REVEALING AND EXPLORING THE LITERATURE'S KNOWN UNKNOWNS: IGNORANCE AND HOW IT DRIVES SCIENCE

by

MAYLA RACHEL BOGUSLAV

B.A., Columbia University, 2016

B.A, The Jewish Theological Seminary, 2016

A thesis submitted to the

Faculty of the Graduate School of the University of Colorado in partial

fulfillment of the requirements for the degree of

Doctor of Philosophy

Computational Bioscience Program

2023

This thesis for the Doctor of Philosophy degree by

Mayla Rachel Boguslav

has been approved for the Computational Bioscience Program

by

James Costello, Chair

Lawrence E. Hunter, Advisor

Sonia Leach, Advisor

Teri Hernandez

Noemie Elhadad

Date: February 13, 2023

Boguslav, Mayla Rachel (Ph.D., Computational Bioscience Program)

Revealing and Exploring the Literature's Known Unknowns: Ignorance and How It Drives Science

Thesis directed by Professor Lawrence Hunter and Associate Professor Sonia Leach

ABSTRACT

Background: Research progresses through accumulating knowledge such that a previously unexplored subject (an unknown unknown) becomes an active research area exploring the questions (known unknowns), until a body of established facts emerges (known knowns). This work aims to help illuminate this process using biomedical natural language processing (BioNLP) to identify, categorize, classify, and explore known unknowns or ignorance statements from the scientific literature. The goal is to help researchers, students, funders, and publishers find the most pertinent research questions or scientific goals for knowledge from the literature based on a topic or experimental results. To do this, researchers must have foundations in both knowledge and questions. However, staying up-to-date on both of them is difficult because of the exponential growth of the scientific literature. Many tools exist to help with knowledge including information extraction systems and knowledge-bases that can be explored by topic or help contextualize experimental results. For the questions, there are some information extraction systems focused on hedging, uncertainty, speculation, factuality, epistemics, and meta-knowledge and a search engine focused on directions and challenges based on a topic (it cannot support queries by experimental results). This prior work mainly focused on the phenomenon in relation to knowledge (*i.e.*, how hedged, certain, speculative, factual, or meta the knowledge is), with only the search engine explicitly focused on new knowledge. With the importance of finding pertinent questions or goals for scientific knowledge, there is a need to go a step further and categorize these statements based on their entailed goal for scientific knowledge (*i.e.*, actionable next steps). Further, no knowledge-bases of such statements, *i.e.*, ignorance-bases, exist that provide summaries and visualizations of such statements based on a topic or experimental results. Thus, we aim to rectify this to help students, researchers, funders, and publishers better understand the state of our collective scientific ignorance (known unknowns or knowledge goals) at scale and across

disciplines, hopefully resulting in an accelerated research process. To determine the feasibility of these general computational ideas, we apply them to the prenatal nutrition field to help find pertinent questions that could affect mothers and offspring globally.

Methods: To capture goals for scientific knowledge, we created a novel BioNLP task to identify, characterize, and classify statements of ignorance based on their entailed knowledge goal. Through manual annotation, we created a taxonomy of ignorance, annotation guidelines, a corpus, and classification models. We also identified the biomedical concepts (ontology concepts) in the ignorance statements to understand their biomedical subjects. We systematically characterized the factors that contributed to the accuracy and efficiency of several approaches to biomedical concept recognition, while aiming to improve performance. Together, the ignorance and biomedical concept classification formed the first ignorance-base on the prenatal nutrition literature. To demonstrate its power, we present two methods of exploration: (1) exploration by topic (*e.g.*, vitamin D) to show that ignorance statements can provide new ideas for future research, and (2) exploration by experimental results (*e.g.*, vitamin D and spontaneous preterm birth gene list) to help a researcher contextualize their results in the ignorance landscape providing questions that their results may bear on potentially from other disciplines.

Results: We show that it is possible to characterize known unknowns as knowledge goals (ignorance taxonomy), that humans can identify statements of ignorance in the literature (annotation task to create a corpus), and that they can be automatically identified (ignorance classification). For the biomedical concepts, we present an automatic biomedical concept recognition system that performed comparably with state-of-the-art systems with some substantial efficiencies in the time and computational resources required for tuning and training. Combining these, we created the first ignorance-base on the prenatal nutrition literature. Exploring it by the topic vitamin D showed that it is possible to find other areas of research (immune system, respiratory system, and brain development) with lots of questions (ignorance statements) that are ripe for future research. Exploring it by a gene list, provided a novel research area (the brain) that implied a different discipline (neuroscience) of which could be explored for answers. Overall, the

ignorance-base provided a question foundation rooted in knowledge goals and ideas for future research.

Conclusion: Our goal is to help students, researchers, funders, and publishers better understand the state of our collective scientific ignorance in order to help accelerate translational research through illuminating the known unknowns and their respective goals for scientific knowledge.

The form and content of this abstract are approved. We recommend its publication.

Approved: Lawrence Hunter and Sonia Leach

Dedicated to all researchers, especially students, searching for a question. The perfect one may be in the place you least expect.

ACKNOWLEDGEMENTS

This work was funded by two National Library of Medicine grants: T15 LM009451 awarded to Mayla R. Boguslav and R01 LM013400 awarded to Dr. Lawrence E. Hunter.

I would like to thank my advisors, Drs. Larry Hunter and Sonia Leach for allowing me to pursue my own ideas and interests. Thank you for always providing helpful feedback as both breadcrumbs to the next big idea (Larry) or sitting with me to go line by line through code or a paper (Sonia). Thank you for also seeing me as a whole person, not just as a researcher, when I faced unexpected challenges in my personal life. Thank you for shaping me into the confident and capable researcher I am today by fostering my curiosity and helping me through any struggles. I would also like to thank my committee for their patience, support, and encouragement throughout this process. To Dr. Jim Costello for guiding me and believing in me as a researcher starting from my rotation project. To Dr. Teri Hernandez for her expertise and her warmth, encouragement, and always having my best interest in mind. To Drs. Michael Paul, Zhiyong Lu, and Noemie Elhadad for serving as outside committee members and bringing new perspectives to my project and committee. This dissertation would not have been possible without this mentorship and guidance.

To the current (and former) members of the Hunter and Leach labs, Dr. Kevin Bretonnel Cohen, Dr. Anis Karimpour-Fard, Dr. Mike Bada, Dr. William Baumgartner Jr., Dr. Elizabeth White, Dr. Ignacio Tripodi, Dr. Tiffany Callahan, Dr. Laura Stevens, Harrison Pielke-Lombardo, Dr. Katherine Sullivan, Lucas Gillenwater, Brook Santangelo, Nourah Salem, Sagi Shaier, and Dr. Adrien Bibal for supporting my research. I am very thankful to Anis, Mike, Bill, and Kevin for their support, wisdom, and friendship. I am also thankful to the other students and postdocs in the computational bioscience program, especially Dr. Brian Ross, Dr. Cody Glickman, Dr. Tiffany Callahan, Rutendo Sigauke, Emily Mastej, and Brook Santangelo for their encouragement and willingness to listen and help with my problems regardless of their personal deadlines. To Dr. Kristen Wade, Dr. Iain Konigsberg, and Hamish Pike from the genetics program for their friendship and support. I am incredibly thankful to Dr. Janet Siebert and Dr. Kristen Wade for being my partners in graduate school through everything.

I am incredibly grateful to all of the collaborators I have had the honor of working with throughout my PhD, including many already mentioned. Their specific contributions are described at the end of each forthcoming Chapter. I especially would like to thank all of my annotators, Dr. Elizabeth White, Dr. Katherine Sullivan, Stephanie Araki, and Emily Dunn for their countless hours of very detailed oriented work that not only helped me shape my task but also provided the fundamental data for this work.

Finally, I owe a huge amount of gratitude to my family and friends for supporting me throughout my program. My parents, Linda and Bruce, for teaching me to problem solve and fostering my curiosity from a young age. To my siblings, Arielle and Louis, for the many fun and philosophical conversations about life, questions, and knowledge. To my adopted Colorado dance family, especially Laurie and Louis Morris, Joan and Ken Saliman, Sharon Swartz, Bev Michaels and Scott Fisher, Lauren Gallagher, Bennett Robinson, the Bogan Family and the Boulder dance group, for their support, friendship, and great dancing. To my adopted Chicago dance family, especially Gabrielle Gordon and Iris Pinto for their friendship and emotional support. To Rabbi Mendel and Estee Popack and their kids for welcoming me into their Jewish home. To Laurie and Louis Morris, Marilyn Hirsch and Gary Gordon, Arielle Boguslav, Camp Ramah in the Rockies, and Kristen Johnson for providing me with family and internet during difficult times. To Erica Shear, Evan Goldstein, Mimi Kaplan, Jessie Leider, Dr. Oded Stein, Jenn Gutterman, and Katie Palazzolo for being there for me from the beginning, providing support, encouragement, and laughs along the way. Lastly, to Gabrielle Gordon and Joshua Mendel for being my people through all the adventures, providing countless hours of emotional support and pancakes.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

CHAPTER I

INTRODUCTION

Motivation

Research begins with a question. It progresses through accumulating knowledge such that a previously unexplored subject (an unknown unknown) becomes an active research area exploring the questions (known unknowns), until a body of established facts emerges (known knowns) [1–3]. We aim to help illuminate this process using biomedical natural language processing (BioNLP) to identify, categorize, classify, and explore known unknowns highlighting their entailed goals for scientific knowledge (*i.e.*, actionable next steps). These known unknowns are discussed in the scientific literature as statements about knowledge that does not exist yet, including goals for desired knowledge, statements about uncertainties in interpretation of results, discussions of controversies, and many others; collectively we term them **statements of ignorance**, borrowing the term from Firestein [1]. For example, “these inconsistent observations point to the complicated role of vitamin D in the immune modulation and disease process" (PMC4889866) is a statement of ignorance. The biomedical concepts within such statements are also important to understand the biomedical subjects of these known unknowns (**biomedical concept recognition**). These include “vitamin D" and “immune" from the example. We aim to reveal both the statements of ignorance and the entailed biomedical concepts to help students, researchers, funders, and publishers better understand the state of our collective scientific knowledge and **ignorance** (known unknowns). This work provides all the necessary methods and tools to create an **ignorance-base**, a knowledge-base containing representations of known unknowns for students, researchers, funders, and publishers to explore the landscape of ignorance at scale and across disciplines, hopefully resulting in an accelerated research process.

A knowledge foundation underlies known unknowns. Many **knowledge-bases** exist to capture the known knowns from domain experts, the scientific literature, and other data sources such as experimental results [4]. To capture the biomedical concepts, many of them utilized biomedical ontologies (*i.e.*, community controlled biomedical vocabularies) both to ameliorate

ambiguity and variability, and for interoperability. These knowledge-bases are important based on their variety of applications [4], including finding and interpreting information based on a single input topic, such as a concept, or a set of input topics that may be related, such as those from experimental results. For example, a graduate student or researcher interested in learning about the field of prenatal nutrition might consult a database of dietary supplements [5]. Or a researcher might perform a functional enrichment analysis to characterize a list of genes associated with vitamin D and preterm birth by finding relevant known biomedical concepts [6].

Both of these examples explore knowledge-bases of known facts, but equally important are the questions or known unknowns surrounding a topic or a set of experimental results. But how do researchers find the most pertinent questions? Researchers gain these skills in graduate school, where the goal is to identify and provide at least some solutions for a question that is unanswered. There are many books [7–11] and articles [12–16] discussing how to choose the most pertinent question or topic, and yet there is only one automated system, the COVID-19 challenges and directions search engine [17]. Lahav *et al.* [17] created a search engine mainly focused on COVID-19 to help researchers discover scientific challenges and directions (their two known unknown categories) by providing relevant articles and sentences for each input biomedical MeSH (Medical Subject Headings - a controlled vocabulary that is part of the Unified Medical Language System from NLM used for indexing, cataloguing, and searching for biomedical information and documents [18]) term. However, their two categories of known unknowns were quite broad, when most other works included more nuanced categorizations (*e.g.*,[19–27]). Their input concepts were not grounded in ontologies, limiting the ability to connect their work to other ontology-based efforts (*i.e.*, many knowledge-bases). Also, they cannot support queries by experimental results as standard methods for contextualizing experimental results (functional enrichment analysis) use knowledge-bases and ontologies [28–32]. Further, they did not go beyond the identification of sentences to provide summaries or visualizations of the immense amount of data outputted. Tables alone do not suffice to truly explore all the outputted data (thousands of sentences can be found per input concept). They “hope to build more tools to explore and visualize challenges and

directions across science" in future work [17]. Thus, no prior work has created a knowledge-base to capture the nuance of the known unknowns drawing from prior work, connected them to ontologies for integration with other knowledge-bases, and provided summaries and visualizations of the outputs to help researchers find the most pertinent questions.

Such an automated system could be useful to a wide variety of scientists ranging from graduate students looking for thesis projects (*e.g.*, [12]) to funding agencies tracking emerging research areas (*e.g.*, [33]). It could help facilitate interdisciplinary interactions amongst researchers by finding questions from another field that bear on a topic or a set of experimental results (*e.g.*, [34]). It could also help track the evolution of research questions over time as a longitudinal analysis (*e.g.*, [35]). Further, automatically identifying questions would allow us to query existing databases for information (*e.g.*, [36]). Thus, there is a need for such an automated system to capture questions or known unknowns.

Many efforts exist to capture the known unknowns [4], but none have been used to create an ignorance-base. Known unknowns were represented through understanding the phenomenon [20, 21, 23–25, 27, 37–67], creating taxonomies where a hierarchy of terms was linked by specified relationships [43, 46, 68–73] and ontologies specifying relationships among controlled vocabularies [74–77], annotating literature to create corpora [40, 41, 50, 55, 63, 78–85], and automating the identification of known unknowns through classification tasks [20, 42–45, 47–49, 51–54, 56, 58–62, 64–67, 86–94]. Some efforts have also sought to capture known unknowns completely by creating theoretical frameworks, determining if the task was feasible for humans to perform, and automating it [17, 19–27, 46, 95, 96]. However, none of these works, including Lahav *et al.*, focused on the entailed knowledge goal of the known unknowns nor have provided summaries and visualizations of the results, especially for an input topic or a set of experimental results. As such, no existing system can support exploring the questions surrounding a topic (a set of ontology terms) or set of experimental results.

For experimental results specifically, current methods for contextualizing them (standard functional enrichment analyses) use knowledge-bases and ontologies [28–32], and natural

language processing (NLP) tools over the biomedical literature [34, 97–106]. Some of this prior work not only aimed to characterize genes but also to help define new research areas (*e.g.*, [34] as one of a few goals), generate new hypotheses (*e.g.*, [107]), and find information about genes of unknown function and fill gaps in knowledge (*e.g.*, a preprint [105] employs manual curation). Thinking beyond a gene list, if we consider pathway models as experimental results, tools exist to associate pathway models to the literature (*e.g.*, [108]) and some took uncertainty into account (*e.g.*, [44, 109]). However, these works focused on confidence and relevance, respectively, rather than explicitly representing statements of known unknowns. Thus, to the best of our knowledge, there is no prior work that can facilitate the widespread search for questions to find new avenues for exploration (knowledge goals) from a set of experimental results.

This work aims to change this by: (1) creating methods and tools to identify, categorize, and classify known unknowns based on their entailed knowledge goals, (2) improving upon existing methods to classify the biomedical concepts within the statements of ignorance to ontologies, and (3) creating a knowledge-base containing representations of known unknowns, formally defined as an **ignorance-base** (based on [1]), to **explore by topic** and **explore by experimental results**, as done in the knowledge-bases. Identifying questions or knowledge goals that need answers, allows us to then look to other fields and knowledge-bases to help answer them. Our goal is to provide students, researchers, funders, and publishers with a mechanism to better understand the state of our collective scientific ignorance (known unknowns) in order to help accelerate translational research through the continued illumination of and focus on the known unknowns and their respective goals for scientific knowledge.

While these ideas and methods are generally applicable across biomedical research, we chose to apply this efficacy work to the field of prenatal nutrition as a starting point. Due to ethical and legal considerations and complexities in studying pregnant mothers and fetuses, prenatal nutrition is understudied and serves to benefit from the identification of questions that are well studied in other fields [110–113]. Fetal development is a critical period and exposure to nutrition has a lifelong impact [114]. For example, the micronutrient vitamin D is very important for

maternal and fetal health, affecting the immune and musculoskeletal systems, neurodevelopment, and hormones [115–119] (see Figure 1.1). Abnormal vitamin D levels are associated with gestational diabetes mellitus, preterm delivery, frequent miscarriages, adipogenesis, pre-eclampsia, obstructed labor, Cesarean sections, reduced weight at birth, respiratory issues, postpartum depression, and autism [115]. If we can identify the known unknowns in prenatal nutrition, even just with regard to the role of vitamin D, then we can look to other fields to help us answer the questions raised, potentially avoiding the additional risks faced when studying pregnant women and their offspring. Also the prenatal nutrition field is a good case study for these ideas because it contains a diverse literature with all types of studies from all over the world, meaning there is a higher potential for generalizability to other fields. Thus, this work has the potential to generalize beyond prenatal nutrition, and more specifically to facilitate new interdisciplinary interactions that could advance the study of an underserved population and potentially help accelerate translational research for mothers globally.

Ignorance Task Description And Results

Our goal is to create an ignorance-base as a formal representation of **statements of ignorance** using the example of the prenatal nutrition literature. We define statements of ignorance as statements of incomplete or missing knowledge categorized based on the entailed **knowledge goals** (*i.e.*, the next actionable step based on the given unknown). A knowledge goal focus allows us to provide actionable next steps to the students, researchers, funders, or publishers. Following the example above, “these inconsistent observations point to the complicated role of VITAMIN D in the IMMUNE modulation and disease process" (PMC4889866) is a statement of ignorance with the entailed knowledge goal to determine the correct role of vitamin D in the immune modulation and disease process by creating novel methods or conducting new experiments to study the complicated role. The ignorance statement, with its entailed knowledge goal, was identified based on the underlined words or **lexical cues** that communicate knowledge is missing. All lexical cues map to a categorization of knowledge goals, an **ignorance taxonomy**. For the example, the cue inconsistent is an *indication of alternative research options or*

controversy of research, observations and point are *indication(s) of proposed or incompletely understood research topic or assertion*, complicated is an *indication of difficult research task*, and role is an *indication of indefinite relationship among research variables*. Using these definitions, we show that it is possible to characterize known unknowns as knowledge goals (ignorance taxonomy), that humans can identify statements of ignorance in the literature (annotation task to create a corpus), and that they can be automatically identified (ignorance classification task). We found that our annotation guidelines were robust to multiple annotation tasks and yielded a high quality ignorance corpus that was rich in statements of ignorance. Statements of ignorance also seemed to employ a rich vocabulary using many different types of lexical cues. With all the data, we trained and evaluated high performing classifiers for statements of ignorance and lexical cues. We created methods and tools to identify, categorize, and classify known unknowns or ignorance statements based on their entailed knowledge goals.

However, identifying ignorance statements and lexical cues alone is not enough to fully understand the known unknowns. The biomedical subjects (or concepts) of ignorance statements are necessary to both understand them fully and be able to connect them to other knowledge-bases. We identified biomedical concepts from the open biomedical ontologies (OBOs) (*e.g.*, the all caps words in the example sentence) using **concept recognition** methods, *i.e.*, automated recognition of references to specific ontology concepts from mentions in text [120, 121]. Concept recognition is a critical task in BioNLP that underlies information extraction from the literature [120–122]. The ultimate goal is to run automatic concept recognition systems over the entirety of the vast biomedical literature (with more than one million new articles per year indexed in PubMed), which means that the efficiency of the such systems is important, as well as the accuracy [123]. At the same time, this task is made difficult by the ambiguity and variability of language with the non-standard usage of abbreviations, synonyms, homonyms, and phrases to describe biomedical concepts [120, 124]. For example, if researchers are interested in the "Golgi-associated PDZ and coiled-coil motif-containing protein" (PR:000008147) and want to automatically extract sentences from the literature containing it, they will have difficulties due to a

synonym: “FIG" is a synonym for this protein and also shorthand for a figure in a paper. To help with these difficulties, many recent open shared tasks including BioCreAtIve [125], the BioNLP Open Shared Tasks (BioNLP-OST) [126], and the recent Covid-19 Open Research Dataset Challenge [127], made concept recognition a focus, creating a community wide effort. All of these shared tasks provided data, evaluation details, and a community of researchers, making them very useful frameworks for further development of such tasks. To continue these efforts, we focused on the Concept Annotation Task of the CRAFT Shared Tasks at BioNLP-OST 2019 (CRAFT-ST): to systematically characterize the factors that contribute to the accuracy and efficiency of several approaches to concept recognition, while aiming to improve state-of-the-art performance. We found that our best-performing systems performed comparably with the state of the art on the task [128] (occasionally extending it by a modest degree), and some offered substantial efficiencies in the time and computational resources required for tuning and training. Our best performing system was Conditional Random Fields (CRFs) [129] or Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT) [130] with Open-Source Toolkit for Neural Machine Translation OpenNMT [131]. Furthermore, we analyzed the strengths and weaknesses of all systems to suggest promising avenues for future improvements as well as design choices that can increase computational efficiency at a small cost in performance. The best performing biomedical concept recognizers were used to identify the biomedical concepts (or subjects) in ignorance statements.

The ignorance-base then is a combination of the identification and classification of ignorance statements and biomedical concepts over the prenatal nutrition literature (1,643 articles in PMCOA). We explored it by topic and by experimental results. We show it is possible to automatically find statements of ignorance related to a topic motivated by researcher’s search for a pertinent question or topic in vitamin D. The goal was to find areas of research (biomedical concepts) with lots of questions (ignorance statements) that are ripe for future research. Analogous to exploring by topic using knowledge-bases, we used the ignorance-base to find other keywords and knowledge goals for the researcher by determining if other concepts were enriched

in vitamin D ignorance statements compared to all vitamin D statements (**ignorance enrichment**). We found the concepts IMMUNE SYSTEM, BRAIN DEVELOPMENT, and RESPIRATORY SYSTEM to be enriched. All of these biomedical concepts with their corresponding ignorance statements provided lots of interesting questions and ideas. In a similar vein, to help the researcher understand the general landscape of questions surrounding a topic and narrow in on a type of question to ask (ignorance category), we present a summary of the ignorance categories and map out how they change over time. We identified the ignorance categories that were over-represented in a subset of ignorance statements as compared to all ignorance statements (**ignorance-category enrichment**). By narrowing the researcher's search to specific ignorance categories, we argue that this provided the researcher with a filtered set of knowledge goals to explore as potential questions for future work. For example, the researcher could choose a topic that is a complete unknown (*indication of unknown or novel research topic or assertion*) or a topic where there are alternate existing hypotheses to explore (*indication of alternative research options or controversy of research*). Further, we argue how this can help track emerging research areas or perform a longitudinal analysis of the evolution of research questions over time for funding agencies and publishers [33, 35, 132–139]. Exploration by topic can be used by any researcher to better understand the question landscape surrounding a topic or set of OBO concepts.

We also show that it is possible to contextualize experimental results in terms of statements of ignorance to understand what questions may bear on them, potentially from another field. Similar to exploration by topic, the goal was to find areas of research (biomedical concepts), which ideally imply another field, with lots of questions (ignorance statements) that are ripe for future research. We used a gene list connecting vitamin D and spontaneous preterm birth (sPTB) from the literature [6] as our example. We mapped the genes to the OBOs to conduct the same analyses as exploration by topic and some added analyses based on the connections between the genes, including gene list coverage, functional enrichment analyses, and comparing our findings to the paper. If vitamin D played a role in preventing sPTB, it would affect mothers everywhere. We found ignorance enrichment of the concepts IMMUNE SYSTEM and BRAIN DEVELOPMENT.

Yadama *et al.* also found the immune system through their functional enrichment analysis and suggested it modulates the effects of maternal vitamin D intake and sPTB [6]. They suggested an increase in maternal vitamin D intake and that future work needs to be done to fully determine the relationship between vitamin D, the immune system, and sPTB. Our exploration provided immune system ignorance statements that can be explored in future work. At the same time, they did not find any concepts related to brain development or mention anything related to the brain in their paper [6]. Thus, we found a novel concept that relates to questions based on their gene list and provide all the ignorance statements for future exploration. Also, the concept brain development implies the field of neuroscience, providing a field for the researchers to look to for answers. The ignorance-base provided a novel putative research area and specific ignorance statements or knowledge goals to pursue in future work based on a gene list. For all findings from the ignorance-base we provide the ignorance statements of interest as a starting point to explore the new found connections as well as discussions of possible conclusions and future questions to ask based on reviewing them with a prenatal nutrition expert (see Figure 1.1 for our results shown in orange). Any researcher with results that can connect to the OBOs can explore the ignorance-base for insights into their results.

The purpose of this work was to highlight the power and importance of focusing on ignorance statements as driven by knowledge goals. We began by improving upon the task of concept recognition as it underlies most information extraction tasks [120–122], including our ignorance task. For the ignorance task, we created a taxonomy of ignorance to characterize statements of ignorance based on their entailed knowledge goals. We showed that humans can identify such statements in the literature using novel annotation guidelines to create an ignorance corpus of 91 articles. We automated classifying ignorance using the corpus drawing heavily from the biomedical concept recognition work. Lastly, using both the ignorance and biomedical concept classifiers, we created the an ignorance-base to explore by topic and experimental results, motivated by a researcher looking for pertinent questions in vitamin D and a vitamin D and spontaneous preterm birth gene list, respectively. Even with all this work, we only scratched the

surface of the power and possibilities of focusing on ignorance statements. We suggest future directions in each Chapter and overall in the Discussion section. (See Figure 1.2 for an overview of this work.) (A list of the formal terms we have introduced here and their definitions are shown in Table 1.1.)

The rest of this Chapter provides an introduction to the relevant BioNLP literature for readers less familiar with the field, a personal motivation for the project, and a summary of the work with contributions to the field of BioNLP. The rest of the dissertation is as follows:

- Chapter 2: Improving Biomedical Concept Recognition: Concept Recognition as Machine Translation
- Chapter 3: Known Unknowns: Capturing and Classifying Ignorance
- Chapter 4: Creating an Ignorance-Base: Exploring Known Unknowns in the Scientific Literature
- Chapter 5: Discussion

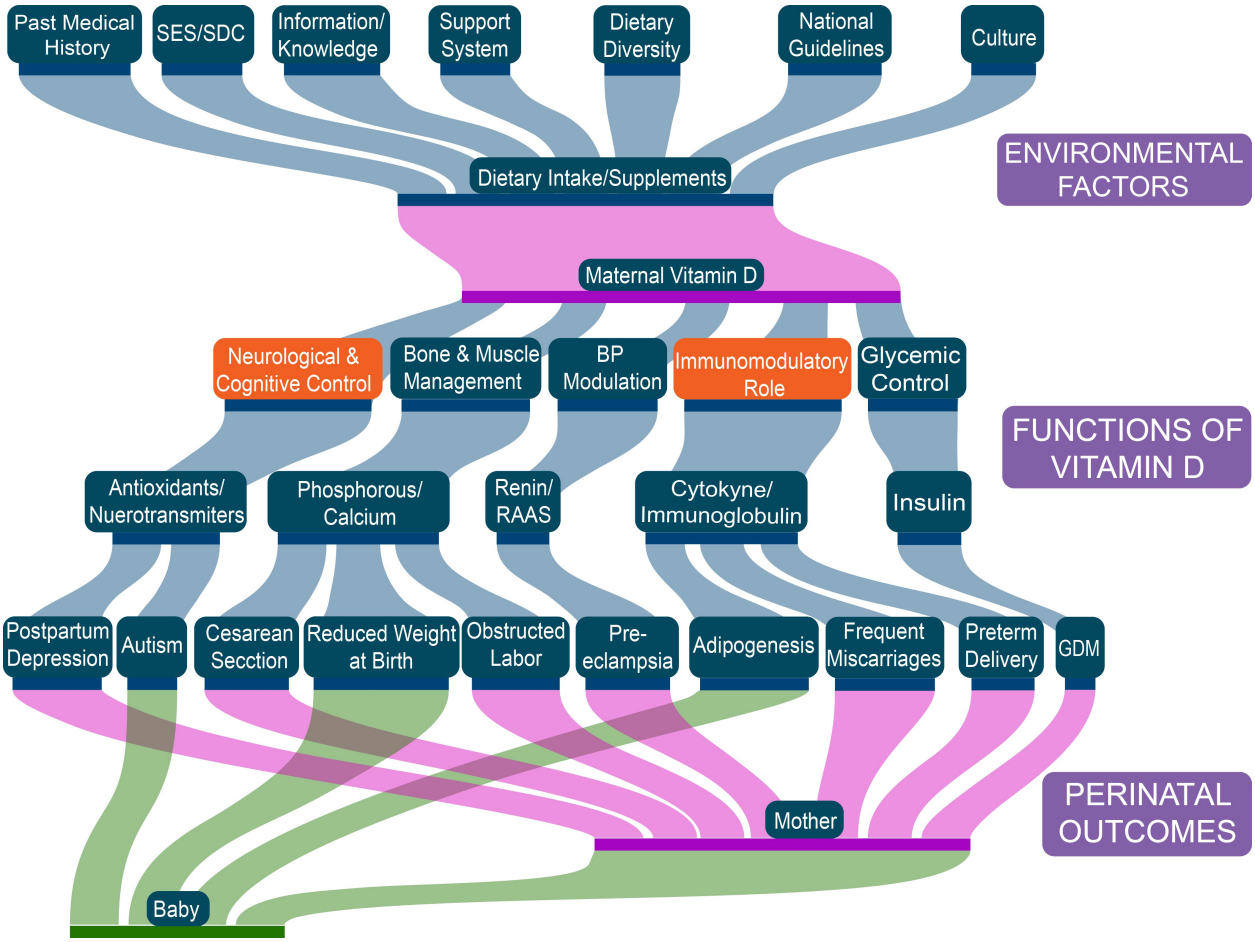


Figure 1.1: Relationship between society, maternal nutrition (vitamin D), and the effects on mother and offspring: a Sankey diagram created based on Figure 3 from [115]. The orange color represents the findings from the exploration methods that the concepts related to brain development and immune system were enriched in ignorance statements and possible novel avenues to explore. SES/SDC = socioeconomic status/sociodemographic characteristics; BP = blood pressure; GDM = gestational diabetes mellitus.

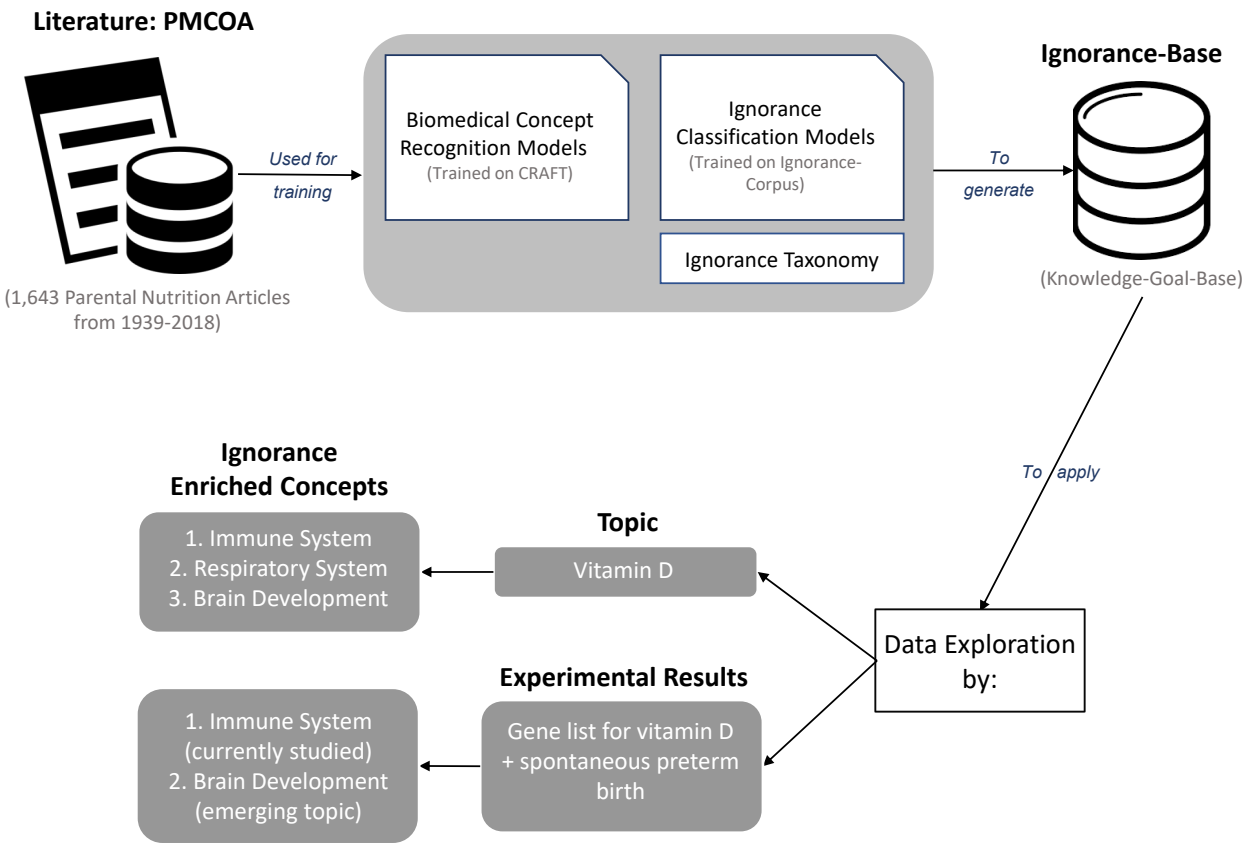


Figure 1.2: Graphical Abstract of the entire dissertation work.

Table 1.1: Term definitions.

Term	Definition
Ignorance	community/collective/scientific known unknowns
Knowledge-base	a database of known information
Statements of ignorance	statements of incomplete or missing knowledge categorized based on the entailed knowledge goal
Knowledge goal	the next actionable step based on the given unknown
Ignorance-base	a knowledge-base, created from the literature, with additional annotations for the sentences that are ignorance statements
Biomedical concept classification/recognition	automatically identifying and mapping biomedical entities to ontologies
Ontologies	controlled vocabularies with specified relationships
Open biomedical ontologies (OBOs)	an effort to create standardized ontologies for use across biological and medical domains
Lexical cue	words or phrases that signify a statement of ignorance
Taxonomy of ignorance	a categorization of ignorance statements based on the entailed knowledge goal
Exploration by topic	automatically find statements of ignorance related to a topic from the ignorance-base
Exploration by experimental results	contextualize experimental results in terms of statements of ignorance from the ignorance-base to understand what questions may bear on them
Ignorance enrichment	a method to identify biomedical concepts that are over-represented in a set of ignorance statements as compared to all sentences, and thus may be a new promising avenue to explore in relation to the input topic
Ignorance-category enrichment	a method to identify ignorance categories that are over-represented in a subset of ignorance statements as compared to all ignorance statements in order to illuminate the types of knowledge goals to pursue and to map out how they change over time

Relevant Biomedical Natural Language Processing (BioNLP) Introduction

This work both helped improve a canonical BioNLP task as well as created a new task using many BioNLP techniques. Thus, we provide a summary of the field of BioNLP, focusing on tasks relevant to this work for those more unfamiliar with the field. NLP is a subfield of computer science and linguistics that aims to create computer applications with inputs of natural or human language. BioNLP is NLP applied to the biomedicine field. The two fundamental difficulties of human language are ambiguity and variability. The context of a word can be critical to

understanding its meaning (*e.g.*, table can be a piece of furniture or a statistical table). There are multiple words that mean the same or very similar things (*e.g.*, street and road). Both of these issues along with the massive number of words in language make analyzing text difficult in general and is the reason there are still many open problems in BioNLP.

The Importance of Concept Recognition

BioNLP, as focused on the scientific literature, was originally motivated by the need to stay up-to-date on the biomedical research as its vastness increased exponentially and continues to [122, 140, 141]. Efforts in BioNLP continued through the observation that it can help find new facts and generate hypotheses through piecing together facts scattered throughout the literature, termed **literature-based discovery** (LBD) [122, 142, 143]. LBD aims to infer new relationships between biomedical entities and processes. Our work has a similar goal to LBD, but focuses on finding sentences with the entities instead of links between them. Both LBD and our ignorance work rely on identifying and extracting the biomedical entities.

Thus, **named-entity recognition** or **span detection** (the task of delimiting a particular textual region that refers to some specific semantic class or classes in text, *i.e.*, a text mention), became a major focus of BioNLP [144–146]. One of the first tasks was to identify text mentions of genes for different model organisms. BioNLP grew alongside efforts to create genomic databases like GeneBank (the earliest one in 1982) as genetic assays increased in size warranting the synthesis of more information. Molecular biologists wanted to understand their experimental results in the context of other biomedical entities, including other genes, drugs, and proteins, as well as the relevant biomedical literature (early tools included MedMiner [147], iHOP [148], and Textepresso [149]). Our work presents an analysis of some of the most currently widely used and high-performing algorithms to help improve performance for span detection, the first step in concept recognition. The best performing algorithms were used to help understand the biomedical subjects of the ignorance statements. Further, just as previous work aimed to contextualize experimental results in knowledge, one aim of our work is to contextualize experimental results in the context of knowledge goals or ignorance. The goal is to help researchers find pertinent

questions by staying up-to-date on the collective scientific ignorance, similar to the original goal of BioNLP.

The second step of concept recognition is **concept normalization**. To fully understand biomedical text mentions, mapping a text mention to a specific entry in a database or other model of the world (*i.e.*, concept normalization) is helpful [122]. The first main task focused on gene normalization to both map found genes to a reference database such as Entrez Gene [150] and to find information about it in the literature. The need for normalization has expanded as databases have expanded. Concept normalization is not a simple task, since span detection is necessary for it. We must identify what concepts to map and any errors in the span detection perpetuate to concept normalization (*i.e.*, if a text mention of a concept is not identified then it cannot be normalized). For both tasks, the fundamental difficulties of language, the sheer massiveness of the biomedical vocabulary, and the need for domain knowledge also cause problems [122]. This dissertation presents a new way of thinking about concept normalization to improve performance, drawing on another BioNLP task (machine translation - automating translation from one language to another).

To help with the difficulties of language and in an effort to formalize the biomedical vocabulary, biomedical **ontologies** were created and much prior work normalized to them. The goal of an ontology is to capture what is, *i.e.*, what exists in the world. In practice an ontology is a set of concepts in a controlled vocabulary, including definitions, unique identifiers, synonyms, and relationships (is-a and has-part) between concepts [122, 151]. The definition constitutes the concept, not the term; so if a term changes, there is no need to change the concept, but if the definition changes then the concept identifier is deprecated and a new concept created. Thus, if identified text mentions from span detection are mapped to ontology concepts through concept normalization, the problems of ambiguity, variability, and the need for domain knowledge (definitions in the ontology) are solved. We use the biomedical ontologies for the concept recognition task in this dissertation. Further, we create a taxonomy of ignorance, similar to an ontology but without the relationships. The goal of the ignorance taxonomy is to capture the epistemics of what is not known, instead of what is known. This presents its own challenges.

One of the first BioNLP ontologies was the Gene Ontology (GO) [152–154], invented to help researchers determine if a gene from one organism had the same function as a gene from another. As each model organism database had its own naming scheme, scientists from three model organisms databases joined together to create a gene function ontology [122]. Today, the goal of the GO is to develop consistent species-independent descriptions of gene products in different model organism databases with three different branches: biological processes, cellular components, and molecular function [154]. Many other biomedical ontologies also exist for other types of biomedical entities such as, genomic sequences (the sequence types and features ontology), proteins (the protein ontology), and chemicals (chemical entities of biological interest). Since there are interactions between all of these entities, efforts to increase integration and interoperability created the **Open Biomedical Ontology (OBO)** Foundry [155], which relies on the basic formal ontology standards from philosophy [151]. We focus on ten different OBOs for this work and create 13 categories of ignorance under our ignorance taxonomy.

The Scientific Literature

The main source of text for the biomedical scientific literature is **PubMed** [156], which is used in this work. It is a repository of biomedical literature containing only abstracts with links out to the full-text if available. Due to publishing requirements, the full-text is not always available. With this limitation, many previous BioNLP systems focused only on abstracts and not the full-text. In the last decade, there has been a move towards the full-text and the acknowledgement that tools may and do perform differently on abstracts compared to full-text [157].

> "Article bodies are typically separated into sections, with a common set of sections being an introduction, a materials and methods section, a results section, and a discussion section. Materials and methods sections are notorious sources of false positives for many types of applications, and they are often omitted from processing in system evaluations, but the ability to handle them is crucial for extracting the information on methods and on biological context that biologists find crucial in interpreting experiments and the output of text mining applications." [122]

The length and the varieties of formatting also make full-text more difficult than abstracts. Further, added features such as figures, tables, and captions are only in the full-text. Even with these challenges, research has shown the need to use full-text articles over just abstracts because of their richness and ability to help with search [158]. Full-text articles are available in **PubMed Central Open Access (PMCOA)** [159], which is used in this work for all the same reasons.

With the continued growth of PubMed, the need for methods to help researchers keep up with their field and possibly others is apparent. Thus, the goal is to find relevant documents to a search, the tasks of **information retrieval** and **document classification** [122, 160]. Researchers can no longer keep up with their field let alone other fields their work may touch on. Interdisciplinary boundaries broke down as genomic science grew and continues to today.

> "For example, the gene relaxin has recently been found to play a role in human responses to drug treatment for heart failure [161]. This has required cardiologists to research relaxin. However, it turns out that relaxin has been studied for years by obstetricians, since it is also the hormone ripening protein, responsible for preparing the uterus for labor. Thus, cardiologists find that they must not only stay current with their own literature, but also must become familiar with a body of literature from the obstetrics community." [122]

Information retrieval is central to our two novel ignorance-base exploration methods, with the goal to help researcher sweep across disciplinary boundaries, like relaxin. Evaluations of information retrieval systems often rely on experts in a field to either create a test set that identifies the relevant literature including a relevance judgment or to evaluate the results of the systems [162]. We rely on experts to judge the output of our exploration methods as to both relevance and ability to help researchers stay up-to-date.

The ultimate goal of BioNLP is automation. In order to do so, many researchers have turned to manually creating **corpora** for their specific task, a set of documents labelled for a specific phenomenon using annotation guidelines. To ensure a high quality gold standard corpus, the annotation task usually includes at least two annotators that independently mark up the text.

From these two marked-up texts, the **inter-annotator agreement (IAA)** is calculated to measure how well the annotations agree to get a value between 0.0 (no agreement) and 1.0 (full agreement) [163]. The higher the IAA, the more the annotators (annotations) agree with each other. Any disagreements amongst annotators are adjudicated and the corpus is updated accordingly. For the biomedical space, not only is linguistic expertise necessary, but also expertise in the biomedical domain of interest. For example, GENIA [164, 165] is one of the most influential corpora for genomic NLP with 1,999 PubMed abstracts manually marked up for both semantic/biological information and aspects of linguistic structure, including sentence boundaries and part of speech. For the biological information, the GENIA project built their own ontology, which has not been widely adopted and is captured in the other OBOs. Another widely used corpus is **The Colorado Richly Annotated Full Text corpus (CRAFT)** [166–170]. It represents the first attempt to do extensive linguistic and semantic annotation of a large collection of full-text journal articles and has been updated regularly to add new semantic and linguistics annotations. It contains 97 articles from PMCOA randomly selected from the mouse genome literature and contains annotations to ten OBOs:

1. Chemical Entities of Biological Interest (ChEBI): compositionally defined chemical entities (atoms, chemical substances, molecular entities, and chemical groups), subatomic particles, and role-defined chemical entities (i.e.,, defined in terms of their use by humans, or by their biological and/or chemical behavior)

2. Cell Ontology (CL): cells (excluding types of cell line cells)

3. Gene Ontology Biological Processes (GO_BP): biological processes, including genetic, biochemical/molecular-biological, cellular and subcellular, organ- and organ-system level, organismal, and multiorganismal processes

4. Gene Ontology Cellular Components (GO_CC): cellular and extracellular components and regions; species-nonspecific macromolecular complexes

5. Gene Ontology Molecular Function (GO_MF): molecular functionalities possessed by genes or gene products, as well as the molecular bearers of these functionalities (though note that only five of the proper classes of the ontology were used for annotation, while the corresponding extension classes were instead used for the large majority of the classes of this ontology)
6. Molecular Process Ontology (MOP): chemical reactions and other molecular processes
7. NCBI Taxonomy (NCBITaxon): biological taxa and their corresponding organisms; taxon levels
8. Protein Ontology (PR): proteins, which are also used to annotate corresponding genes and transcripts
9. Sequence Ontology (SO): biomacromolecular entities, sequence features, and their associated attributes and processes
10. Uberon (UBERON): anatomical entities; multicellular organisms defined in terms of developmental and sexual characteristics

For each OBO, there is also a corresponding extension OBO, **OBO_EXT**. These extension classes were created by the CRAFT semantic annotation lead for the concept annotations but are based on proper OBO classes; they were created for various reasons, particularly for the purposes of semantic unification of similar classes from different ontologies, unification of multiple classes that were difficult to consistently differentiate for annotation, and creation of similar but corresponding classes that were either easier to use or better captured the ambiguity of the annotated text mentions. The CRAFT corpus is our gold standard data for improving performance on biomedical concept recognition. Further, the best models are used to identify biomedical concepts from these ten OBOs in the ignorance statements.

Automation and Evaluation

For automation, there are three main types of algorithms: (1) rule-based, (2) machine-learning based, and (3) a hybrid of the two methods. Rule-based methods rely on a set of rules to determine the outcome of the task. These rules tend to be both linguistic and biological in nature. The goal is to identify patterns and create rules. For concept recognition, dictionary-based methods, a type of rule-based, aim to find exact matches of concepts from an ontology in text. Thus, the rule is to find exact matches between the text and a concept. Adding more rules or changing the existing ones updates the method. In general, rule-based methods rely on the examination of many examples of the phenomenon, expert information, and pattern recognition.

Machine-learning based methods rely on computers to “learn" the patterns and distinguishing factors for a specific problem. The two main types of machine learning are supervised - the correct answers are provided (*e.g.*, corpora) so the machine can check its work (classification), and unsupervised - the correct answers are not provided and so the machine has to find hidden patterns in the data (clustering). With the many complications of language, supervised machine learning is used more often then unsupervised, but with the very time-consuming trade-off of needing to create manually annotated corpora for each task (CRAFT and ignorance corpora for this work). The supervised machine learning algorithm aims to infer a “function" that captures the relationship between the input data (*e.g.*, text) and the output results (*e.g.*, a concept label), while minimizing the error of doing so. Many classical algorithms exist including support-vector machines, regression, k-nearest neighbor, and naive Bayes. Newer methods include neural networks that aim to mimic the neural connections in animal brains, using a weight function to determine if information should be transmitted between neurons or nodes similar to the function of a synapse. The hybrid method combines both the previous two methods in any order. Some systems do pre-processing rule-based steps and others post-process. All of these methods, except for unsupervised machine learning, are used in this work.

For the rule-based and supervised machine learning algorithms where the answers are known, the most common evaluation metrics are **precision**, **recall**, and **F1 measure**. Precision is

a measure of how often the right answers appear in the chosen right answers (false positives). Recall is the measure of how often the right answers appear compared to all right answers (false negatives). F1 measure is the harmonic mean between the two, aiming to balance false positives and false negatives.

$$\begin{cases} Precision = \frac{true\ positives}{true\ positives + false\ positives} \\ Recall = \frac{true\ positives}{true\ positives + false\ negatives} \\ F1\ measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \end{cases} . \quad \text{(I.1)}$$

These are the evaluation metrics used in this work.

There are many other types of tasks in BioNLP [122]. **Relation extraction** aims to extract information about a very focused type of relationship from free text. **Summarization** aims to reduce a document to a shorter form usually by extracting segments of text to assemble into a summary. **Question answering** is the task of generating answers to questions in natural language. Many more tasks exist in BioNLP, but are not used in this work. However, they could be applicable to future work.

Lastly, many tasks, corpora, classification models, and evaluation frameworks exist because of **shared tasks**. A shared task is a community competition or challenge with a stated task and evaluation metric. In general, shared tasks have created new test collections, fostered new algorithm development and improved results, formed and strengthened research communities, pushed research into real-world applications, and provided education and outreach [171, 172]. For example, the BioCreAtIvE community (Critical Assessment of Information Extraction systems in Biology) has conducted many shared tasks since their inception in 2004 [125], with explicit goals to attract researchers, address problems of biomedical importance, create legacy training and test data suites, and allow for the assessment of the state of the art on real biological tasks. We used the framework from another open shared task community, the BioNLP Open Shared Tasks (BioNLP-OST) [126], to improve biomedical concept recognition for the Concept Annotation Task of the CRAFT Shared Tasks (CRAFT-ST) [170]. Further, we create a task, a corpus, and classification algorithms for our ignorance work that could be used in a future shared task. The

ultimate goal is to harness community wide effort and focus on identifying, categorizing, classifying, and exploring ignorance to help researchers find the most pertinent questions to research.

Personal Motivation

My motivation for this project stemmed from my need to find a pertinent question to study and not knowing where to start. Coming into a computational biology (compbio) graduate school program I did not know the questions in compbio let alone in mathematics, the field I studied as an undergraduate. How would I find a question to answer for my dissertation given this situation? This then became my focus and question: How do we help researchers, especially students, understand the questions in a field so that they themselves can decide on a question to answer for their work? Everyone needs a knowledge foundation and researchers especially also need a curiosity and question foundation. This was the motivation for the ignorance-base and exploration methods. I hope this work not only helps students and researchers find and understand questions, but also sparks continued efforts that focus on the questions to provide better curiosity and question foundations for research.

Summary/Contributions

Questions drive science. This work shows that harnessing the questions in the scientific literature can lead to new and exciting avenues for exploration by researchers interested in a specific topic or provide questions that may bear on a specific set of experimental results. I provide an initial characterization of the ignorance task through an annotation task with a taxonomy of ignorance and a classification task to show automation is possible [173]. To identify the biomedical subjects of the ignorance statements, we improve upon biomedical concept recognition creating state-of-the-art classifiers [174]. Combining the ignorance work and the biomedical concept recognition work [174], I create a novel ignorance-base, along with two new exploration methods (preprint [175]). As all of this work was done in the prenatal nutrition field, I also made contributions to the field through the results and discussions of the results.

Separately, I have also made contributions to the understanding of the inter-annotator agreement (IAA) and whether it is the upper limit of machine performance [176]; to speeding up calculations for simulating heterogeneous populations, such as the development of cancer, using Boolean models [177]; to providing pre- and post-processing techniques to improve precision in concept normalization based on an error analysis [178]; to improving performance on relation extraction for chemicals and proteins (or genes) by conducting an error analysis [179]; and to the understanding of how the focus on a given disease in research papers has changed over time through discussions of the ideas [136].

CHAPTER II

IMPROVING BIOMEDICAL CONCEPT RECOGNITION: CONCEPT RECOGNITION AS MACHINE TRANSLATION

Background

Automatic recognition of references to specific biomedical concepts from mentions in text (hereafter **concept recognition**) is a critical task in biomedical natural language processing (BioNLP) and underlies many other BioNLP tasks [122, 180–183], including our ignorance task. Biomedical concept recognition is necessary to understand the biomedical subjects of the ignorance statements. The identification of such concepts ameliorates ambiguity and provides the biomedical subjects of text, but its automation poses many difficult computational challenges. For example, if researchers were interested in the "Golgi-associated PDZ and coiled-coil motif-containing protein" (PR:000008147) and wanted to automatically extract sentences from the literature containing it, they would have difficulties due to a synonym: "FIG" is a synonym for this protein and also shorthand for a figure in a paper. The ability to distinguish between the protein and a figure is necessary for automatic extraction of the information and any other downstream analyses of it, such as automatically summarizing the other biomedical concepts within the protein sentences. The identifier in parentheses next to the protein helps with this. It is an ontology ID from the Protein Ontology [184], an ID in a controlled biomedical vocabulary focused on proteins that require semantically coherent definitions, regardless of the variability in textual expression of the represented concept [155] (*i.e.*, there is a canonical definition provided by the ontology for each term even if there are multiple ways to express it in text).

The target of most biomedical concept recognition efforts have been the Open Biomedical Ontologies (OBOs), including the Protein Ontology [184] and others such as the Gene Ontology [153] and the Human Phenotype Ontology [185], which provide canonical definitions. Each of these ontologies contain more than 10,000 specific classes. Thus, treating concept recognition as a classification task requires tens of thousands of classes, rendering machine learning approaches for this multi-classification problem impractical. The current state of the art (*e.g.*, [128]) was a

hybrid system which involved the use of a rule-based dictionary lookup system, neural networks, and language models. The rule-based dictionary lookup system was specifically to combat the problem that the models will only ever predict concepts seen in training data. The very large number of ontology classes to be recognized and the limited amount of gold-standard training data have impeded the creation of end-to-end systems based entirely on machine learning that can also predict concepts not seen in the training data.

Thus, our goal was to create an end-to-end system based entirely on machine learning that can also predict concepts not seen in the training data. Recently, Hailu *et al.* [186] recast the concept recognition problem as a type of machine translation and demonstrated that sequence-to-sequence machine learning models have the potential to outperform multi-class classification approaches. They and others in the NLP literature, often divided concept recognition into two tasks that were performed separately and differently (see Figure 2.1): Span detection (also referred to as named entity recognition or mention detection), which delimits a particular textual region that refers to some ontological concept (*i.e.*, a text mention) [180–182]; and concept normalization (also referred to as named entity normalization or entity linking), which identifies the specific ontological class to which the textual region or mention refers (an ontology class ID) [183]. It was possible to approach these problems jointly with a single system, but evaluation approaches generally scored them separately and then combined the scores for the full system. Further, different tasks may require only one. Thus, approaching the problem separately allowed these methods to be useful for either task or the combination. The ultimate goal is to be able to run such systems over the entirety of the vast biomedical literature (with more than one million new articles per year indexed in PubMed), meaning that the efficiency of such systems is important, as well as their accuracy [123]. Thus, we systematically characterized the factors that contributed to the accuracy and efficiency of several approaches to sequence-to-sequence machine learning.

Span detection [180–182] has long been conceived of as a sequence-to-sequence analysis task. The outputs are defined as sequences of tags identifying the beginning words or characters of mentions (labelled “B"), words or characters inside mentions (labelled “I"), and words or

characters outside of mentions (labelled “O"), collectively referred to as BIO tags [187]. We evaluate the most widely used and high-performing sequence-to-sequence algorithms for span detection to determine the best-performing algorithm as well as options for low-resource settings. This included conditional random fields (CRFs) [129] (which may be utilized with limited resources), bidirectional long short-term memory networks (BiLSTMs) [188, 189] (which required substantial resources, spurring us to evaluate whether hyperparameter settings tuned on a simple model can be reused in more complex models to save some resources), and language models (in particular, ELMo [190] and BioBERT [130], which were the current state of the art in NLP). In our evaluation, these systems showed widely varying performances and resource requirements, and the performance was not strongly correlated with resource requirements. Furthermore, though they were relatively rare, we attempted to identify discontinuous concept mentions, *i.e.*, mentions that span two or more discontinuous strings of text, a textual phenomenon that only a few researchers have specifically focused on recently [191, 192] and that others have chosen to ignore [128, 186]. We propose a simple extension to BIO tags that captured at least some discontinuous spans for almost all evaluated ontologies.

For concept normalization, we recast the classic multi-class classification problem as a machine translation problem. Most of the previous rule- or machine-learning based methods fell short on predicting text mentions not seen in the training data because of the very large number of classes and the very limited number of examples [183]. To combat this, Hailu *et al.* [186] proposed the idea to use sequence-to-sequence methods for concept normalization, expressing both the text mentions and the concept IDs as character sequences thinking of the task as machine translation (*i.e.*, mapping each letter or character of the example protein above to the corresponding character of the ontology ID: G o l g i - a s s o c i a t e d P D Z a n d c o i l e d - c o i l m o t i f - c o n t a i n i n g p r o t e i n to the ontology ID P R : 0 0 0 0 0 8 1 4 7). They investigated several alternative approaches used in machine translation to apply here and found that the popular Open Neural Machine Translation system (OpenNMT) [131] was the best-performing, although many other neural machine translation systems had very similar performance [193]. To the best of

our knowledge, this was the first attempt to use machine translation for concept normalization and we continue this endeavor. We focus solely on OpenNMT for this exploration because the good performance of sequence-to-sequence approaches to normalization with the ontology class identifiers (*e.g.*, PR:000008147) as the targets was surprising, as the identifiers were intended to be arbitrary and devoid of semantic content [155, 194]. We explored the reasons underlying this surprisingly good performance by evaluating a variety of alternative identifier schemes, suggesting that there are, in fact, semantic signatures in the class identifiers.

We used the Concept Annotation Task of the CRAFT Shared Tasks at BioNLP Open Shared Tasks 2019 (CRAFT-ST) as the framework for this work [170]. Concept recognition has been the subject of many open shared tasks, including BioCreAtIve [125], the BioNLP Open Shared Tasks (BioNLP-OST) [126], and the recent Covid-19 Open Research Dataset Challenge [127]. Shared tasks are valuable because they provide data, evaluation details, and a community of researchers, making them very useful frameworks for further development of such tasks. We chose CRAFT-ST not only because it was one of the most recent shared tasks involving concept recognition, but also because of the richness of data in the CRAFT (Colorado Richly Annotated Full-Text) corpus [166, 168, 169] and the ability to compare our results to the current state of the art [128]; however, all methods described here can be applied to other corpora as well. The CRAFT corpus contains 97 full-text articles from PubMed Central Open Access with extensive linguistic and semantic annotations to multiple OBOs. Further, it has been used for training and/or evaluation in many concept recognition tasks [166] including the CRAFT Shared Task [170] and this work. We found that our best-performing sequence-to-sequence systems performed comparably with the top performing system on the shared task (occasionally extending it by a modest degree), and some offered substantial efficiencies in the time and computational resources required for tuning and training. Furthermore, our analysis of the strengths and weaknesses of such systems suggested promising avenues for future improvements as well as design choices that could increase computational efficiency at a small cost in performance. Thus, we report on extensive studies of alternative methods and hyperparameter selections in an exhaustive

evaluation framework that has been previously developed for open shared tasks [126]. These results not only identified the best-performing systems and parameters across a wide variety of ontologies but were illuminating with regard to the widely varying resource requirements and hyperparameter robustness of alternative approaches. Further, we aimed to be transparent about our classification process in order to facilitate usability and reproducibility [195]. Thus, we present the full end-to-end system for concept recognition along with its components of span detection and concept normalization. Although all of this work focused on identifying biomedical concepts, it can be applied to other types of concepts and potentially many other types of linguistic phenomena.

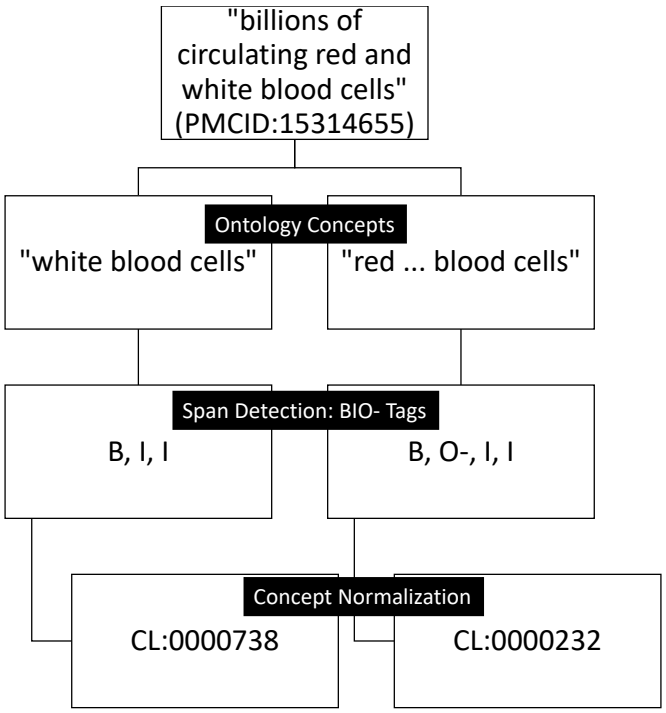


Figure 2.1: Example of the full translation pipeline. Each step was seen as a translation problem. The input was text and the final output was the ontology class identifiers for each detected text mention.

Related Work

Dictionary-based and rule-based methods mainly dominated concept recognition approaches for the Open Biomedical Ontologies due to the enormous number of concepts to identify. Funk *et al.* [196] performed a systematic evaluation of some of these dictionary-lookup systems, finding that ConceptMapper [197, 198] generally performed the best. They not only

identified the highest-performing systems but also the best parameter settings (finding they were not the default settings), for each of the ontologies from CRAFT version 1.0 [168, 169]. They achieved F1 scores between 0.14 and 0.83. The most recent dictionary-based approach was NOBLE Coder [199], which implemented a “general algorithm for matching terms to concepts from an arbitrary vocabulary set". Tseylin *et al.* [199] compared the performance and speed between NOBLE Coder and ConceptMapper, and found that their system outperformed ConceptMapper, but ConceptMapper was the fastest. (It was unclear if the parameters found in Funk *et al.* [196] were used for this work.) This highlights the importance and sometimes trade-off between speed and accuracy. These dictionary-based methods have remained popular because of their interpretability and configurability with no necessary training [200].

Going beyond rule-based systems, recent advances in machine learning have resulted in many hybrid systems that apply machine-learning-based post-processing to dictionary-based systems. Many of these works utilized the CRAFT corpus [166, 168, 169] for training and evaluation. For example, Campos *et al.* [201] proposed a hybrid system employing dictionary matching and a machine learning system for biomedical concept recognition. Nunes *et al.* [202] used this system to create a a biomedical concept recognition annotation and visualization service named BeCAS. Groza *et al.* [203] approached the task as an information retrieval problem and explored case sensitivity and information gain. Basaldella *et al.* [204] proposed a hybrid system named OntoGene’s Entity Recognizer (OGER), which focused first on high recall through a dictionary-based entity recognizer, followed by a high-precision machine learning classifier (see [205] for an updated version of this system). Furthermore, the group who developed this system had the highest-performing method in the 2019 CRAFT Shared Task [128] (UZH@CRAFT-ST), combining an updated version of OGER with two neural approaches, thereby tackling concept recognition as a single task instead of two. As we tackled the same task through the same framework, we used their results as a baseline for the full concept recognition system. Full machine-based systems have only recently been developed with the advent of deep learning (*e.g.*, [206, 207]). Even these deep learning methods seemed to fail when there were not enough

examples in the training data. Thus, our work continues the use of deep learning models to test efficiency, accuracy, and generalizability over the CRAFT corpus [166, 168, 169], especially for concepts not in the training data.

As mentioned earlier, much prior work in concept recognition split the task into two sub-tasks: span detection (named entity recognition or mention detection) [180–182] and concept normalization (named entity normalization or entity linking) [183]. Concept recognition was spilt to apply different approaches for different applications. Our work focuses on sequence-to-sequence algorithms for both tasks and many recent publications have used these approaches for span detection but not concept normalization. For example, Huang *et al.* [208] proposed a model based on a BiLSTM combined with a CRF for BIO tagging and achieved better tagging accuracy for part-of-speech tagging, chunking, and span detection than with a CRF alone. Throwing in character and word embeddings, Lample *et al.* [209] used the same neural architecture as Huang *et al.* for span detection. Adding a CNN to the mix, Ma *et al.* [210] proposed an end-to-end sequence-tagging model based on a BiLSTM-CNN-CRF approach. Thinking beyond any specific language, Gillick *et al.* [211] created an LSTM-based model that reads text as bytes and outputs the span annotations. With their focus on bytes, their representations and models generalized across many languages, creating the first multilingual named entity recognition system.

In the biomedical domain, Habibi *et al.* [212] applied the BiLSTM-CRF proposed by Lample *et al.* [209] for span detection on a wide range of biomedical datasets and found that their model outperformed the state-of-the-art methods. The same architecture was used by Gridach [213] to identify spans of genes and proteins. Zhao *et al.* [214] proposed a multiple-label strategy to replace the CRF layer of a deep neural network for detecting spans of disease mentions. To identify spans of chemicals, Korvigo *et al.* [215] applied a CNN-RNN network, Luo *et al.* [216] proposed an attention-based BiLSTM-CRF, and Corbett *et al.*[217] explored a BiLSTM and CRF separately as well as combined to create ChemListem and further added transfer learning. Similar to Luo *et al.*, Unanue *et al.* [218] used a BiLSTM-CRF to identify spans of drug names and

clinical concepts, while Lyu *et al.* [189] proposed a BiLSTM-RNN model to detect spans of a variety of biomedical concepts, including DNA, genes, proteins, cell lines, and cell types. Wang *et al.* [219] applied multitask learning with cross-sharing structure using a BiLSTM-CNN-CRF model, which included a BiLSTM that learns shared features between ten datasets with gene, protein, and disease categories, and a private BiLSTM specific for each task, borrowing their base model from Ma *et al.* [210]. Sequence-to-sequence algorithms were quite popular for span detection.

More recent advances in deep learning for a variety of different NLP tasks, including span detection, have been achieved with language models. BERT [220], the science-specific language model SciBERT [221], and the biomedicine-specific language model BioBERT [130], all required little fine-tuning to perform well for named entity recognition. The other main language model, ELMo [222], and its biomedical equivalent [190], also performed well on named entity recognition. Peng *et al.* [223] evaluated both BERT and ELMo on the Biomedical Language Understanding Evaluation (BLUE) benchmark and found that BERT, with extra biomedicine-specific documents, outperformed ELMo. The success of these methods showed the value of deep learning and language models for span detection. However, the performance of machine learning methods in NLP depends crucially on hyperparameter selection (as does the performance of many dictionary- and rule-based methods, *e.g.*, [196]). Large computational resources and time requirements, particularly for repeated training and testing under different hyperparameterizations, can limit the extent of hyperparameter searches to find optimal values. None of these works conducted extensive studies of alternative approaches based on efficiency and resources as we do here.

We also applied a sequence-to-sequence method (machine translation) by character for concept normalization for generalizability. To the best of our knowledge this has not been attempted before. Due to the very large number of ontology classes to be recognized and the limited amount of gold-standard training data, the first approaches to concept normalization were rule-based dictionary lookup systems that mapped text mentions looked up in a dictionary to the

corresponding ontology identifier [183, 197, 198, 224]. For example, the Gene Ontology Cellular Component class GO:0005886 might be textually referenced by “plasma membrane", “cell membrane", “cytoplasmic membrane", or “plasmalemma", among others, *e.g.*, abbreviations such as “PM". All concept normalization methods must cope with the challenges related to the variability and ambiguity of human language. Not only are there many lexical variants that refer to the same ontological concept, but each of those words have morphological variants (*e.g.*, nucleus, nuclei, nuclear, nuclearly). Furthermore, many individual words are ambiguous, depending on the surrounding context for the proper mapping to an ontological class; for example, the word “nucleus" can refer to an atomic nucleus (ChEBI:33252), a cell nucleus (GO:0005634), or an anatomical nucleus (UBERON:0000125), among other senses, depending on the surrounding context. Thus, these dictionary-based methods have remained popular because of their interpretability and configurability with no necessary training [200].

At the same time, research has moved on to machine-learning based methods due to the superior performance, applying everything from linear methods to deep learning methods [183]. For example, Li *et al.* [225] generated concept normalization candidates using a rule-based system and then ranked them using a convolutional neural network that harnessed semantic information. Liu *et al.* [226] used an LSTM to represent and normalize disease names. Similarly, Tutubalina *et al.* [227] used recurrent neural networks, including LSTMs, to normalize medical concepts in social media posts. However, most of these methods fall short on predicting text mentions not seen in the training data because of the very large number of classes and the very limited number of examples. Thus, we evaluated the use and surprisingly good results of machine translation on the character level to help with generalizability.

To the best of our knowledge, this was the first attempt to use machine translation for concept normalization, and so there was no prior literature for this application. Machine translation usually aims to translate from one language to another (see, *e.g.*, [193, 228]). Here we were translating text mentions to ontology identifiers (OBO IDs). Deep learning has been applied to machine translation, and recently, neural machine translation had proven superior to previous

methods. Successful approaches belonged to the family of encoder-decoders, which encode source text into fixed-length vectors from which a decoder generates translations [193]. This was a sequence-to-sequence method that mapped sequences of characters or tokens in one language into sequences of characters or tokens in another language. Bahdanau *et al.* [229] introduced a key innovation by adding an attention mechanism to the decoder to relieve the encoder of needing to encode all information from the source text to fixed-length vectors. With this approach, the information can be spread throughout the sequence of text, and the decoder can select the most useful parts to predict the next character or token. It was this approach that Hailu *et al.* exploited [186] and the approach we further explore here. Since this has never been attempted, we compared our results to ConceptMapper [196, 198, 230], as a baseline model.

Lastly, concept recognition, span detection, and concept normalization used a variety of text representations ranging from bags of words to embeddings of words and characters [231, 232]. However, how to represent complex mentions, including overlapping and discontinuous mentions (see Figure 2.1 for an example) was unclear. Previous research had mostly ignored these complex mentions due to their rarity and difficulty, and only recently have some researchers tried specifically to tackle them by extending existing sequence tagging frameworks (as we did here), or by looking at a given sentence as a whole to determine the relationships between concept mentions (see [191, 192] for an overview of previous work). To represent them explicitly, we evaluated alternative text representations using a modified sequence-to-sequence BIO tag representation [187]. The method proposed here is simpler than previous methods and still identified some of these complex mentions.

Methods

Our goal was to improve and evaluate the performance of different methods on concept recognition. Specifically, to explore the performance, efficiency, and underlying reasons for the surprisingly good performance of the proposed machine learning approach toward concept recognition, using the 2019 CRAFT Shared Task framework from the BioNLP-OST task [170]. This framework provided all data and an evaluation pipeline, facilitating direct comparison to the

best-performing system in that evaluation [128]. Per the setup of the CRAFT Shared Task, 67 full-text documents were provided as training with 30 unseen documents held out as an external evaluation set (CRAFT version 3.1.3). Every Open Biomedical Ontology in CRAFT (*e.g.*, the ChEBI chemical ontology, the Uberon anatomical ontology) presented different challenges and benefited from different methods, hyperparameters, and/or training regimes. Thus each OBO was trained, tuned, and evaluated independently for both span detection and concept normalization. Furthermore, training all ontologies separately solved both the problems of overlapping ontology annotations from different ontologies and the multi-classification problem between ontologies, and it simplified the process of adding new ontologies. For all ontologies, the goal was to optimize F1 score, the harmonic mean between precision and recall, which was the metric for comparison of all models. We also considered the resources needed for all models.

Materials and Evaluation Platform

The Colorado Richly Annotated Full-Text (CRAFT) corpus [166, 168, 169] was used to train, tune, and evaluate our models. CRAFT has had four major releases. The CRAFT Shared Task used version 3.1.3, which was released in July 2019 [170]; we thus used this release for all tasks and we refer to it as CRAFT without the version number going forward. Each release contained the ontology files corresponding to the annotations at that time (most ontologies do not have version numbers). CRAFT version 3.1.3 included 67 full-text documents with 30 documents held back, whereas more recent versions included the previously held-back 30 documents [233]. The total corpus was thus a collection of 97 full-text articles. These articles were focused (though not exclusively) on the laboratory mouse and have been extensively marked up with both gold-standard syntactic and semantic annotations. Among the syntactic annotations, segmented sentences, tokens, and part-of-speech tags were used to extract features for all algorithms tested. CRAFT versions 3.0 and above were a significant improvement from version 2.0 with regard to the concept annotations. Among other changes, relative to version 2.0, the concept annotations were updated using newer versions of the ontologies at the time, annotations were created based on the classes of two additional ontologies (the Molecular Process Ontology and the Uberon

anatomical ontology), and extension classes of the OBOs were created and used to annotate the articles, resulting in a substantial increase in annotation counts. (The extension classes were created by the CRAFT semantic annotation lead for the concept annotations but were based on proper OBO classes; they were created for various reasons, particularly for the purposes of semantic unification of similar classes from different ontologies, unification of multiple classes that were difficult to consistently differentiate for annotation, and creation of similar but corresponding classes that were either easier to use or better captured the ambiguity of the annotated text mentions.) CRAFT contained the semantic annotations to single or multi-word concepts from ten OBOs (all from 2019):

1. Chemical Entities of Biological Interest (ChEBI): compositionally defined chemical entities (atoms, chemical substances, molecular entities, and chemical groups), subatomic particles, and role-defined chemical entities (*i.e.*, defined in terms of their use by humans, or by their biological and/or chemical behavior)

2. Cell Ontology (CL): cells (excluding types of cell line cells)

3. Gene Ontology Biological Processes (GO_BP): biological processes, including genetic, biochemical/molecular-biological, cellular and subcellular, organ- and organ-system level, organismal, and multiorganismal processes

4. Gene Ontology Cellular Components (GO_CC): cellular and extracellular components and regions; species-nonspecific macromolecular complexes

5. Gene Ontology Molecular Function (GO_MF): molecular functionalities possessed by genes or gene products, as well as the molecular bearers of these functionalities (though note that only five of the proper classes of the ontology were used for annotation, while the corresponding extension classes were instead used for the large majority of the classes of this ontology)

6. Molecular Process Ontology (MOP): chemical reactions and other molecular processes (not in version 2.0)
7. NCBI Taxonomy (NCBITaxon): biological taxa and their corresponding organisms; taxon levels
8. Protein Ontology (PR): proteins, which are also used to annotate corresponding genes and transcripts
9. Sequence Ontology (SO): biomacromolecular entities, sequence features, and their associated attributes and processes
10. Uberon (UBERON): anatomical entities; multicellular organisms defined in terms of developmental and sexual characteristics (not in version 2.0)

For each of the ten ontologies used in the CRAFT corpus, there were two annotation sets: a core set and a core+extensions set, each with their own challenges. The core set consisted solely of annotations made with proper classes of the given OBO, and the core+extensions set consisted of annotations with proper OBO classes as well as classes created as extensions of the ontologies. The unique identifier for an OBO class was a class identifier (class ID) that consisted of the ontology namespace, a colon, and a unique number identifier; for example, NCBITaxon:10088 was the unique identifier for *Mus*, the taxonomic genus of mice. The lengths of the class IDs were not consistent within and between ontologies: CL, GO, MOP, SO, and UBERON classes had seven-digit IDs (*e.g.*, CL:0000014 (germ line stem cell), GO:0016020 (membrane), MOP:0000590 (dehydrogenation), SO:0000040 (genomic clone), UBERON:0001004 (respiratory system)), while PR classes had nine-digit IDs (*e.g.*, PR:000000035 (BMP receptor-type 1A)). ChEBI class IDs ranged from one (*e.g.*, ChEBI:7 ((+)-car-3-ene)) to six (*e.g.*, ChEBI:139358 (isotopically modified compound)) digits, and NCBITaxon entry IDs (other than those representing taxonomic ranks) ranged between one (*e.g.*, NCBITaxon:2 (Bacteria)) and seven (*e.g.*, NCBITaxon:1000032 (*Enterobacter* sp. P19-19)) digits. An extremely small subset of

classes of the OBOs used for CRAFT concept annotation instead have textual IDs, *e.g.*, NCBITaxon:species, the NCBI Taxonomy class representing the taxonomic rank of species.

For the extension classes, which were identifiable by their namespace prefixes always ending in "_EXT", concept normalization was even more difficult because the class IDs were even more varied. For four of the OBOs used, one or more parallel hierarchies of extension classes were programmatically created for either the entire OBO or for one or more subhierarchies of the OBO so as to create corresponding classes that were more abstract and/or more straightforward to use for concept annotation. The class ID for such an extension class was the same as the original OBO class on which it was based with the exception of "_EXT" appended to the namespace, *e.g.*, GO_EXT:0004872 (bearer of signaling receptor activity), based on the original GO:0004872 (signaling receptor activity). In addition to these programmatically created extension classes, there were many manually created extension classes, which had entirely textual IDs of human-readable labels with the namespaces of the ontologies of which they were extensions appended with "_EXT", *e.g.*, ChEBI_EXT:calcium. In the case of a manually created extension class that was an extension of more than one ontology, the namespace was an underscore-delimited concatenation of the namespaces of the extended ontologies, *e.g.*, GO_MOP_EXT:glycosylation. Of relevance to our concept normalization work, many of the manually created extension classes had long textual IDs, *e.g.*, GO_UBERON_EXT:innervation_entity_or_process. This variance contributed to poor performance on concept normalization generally.

CRAFT generally contained many examples for each ontology providing enough training and evaluation data for our models, except for GO_MF, MOP, and MOP_EXT. Statistics regarding the annotations and annotation classes of the training and evaluation data sets can be seen in Tables 2.1 and 2.2, respectively. For GO_MF, only a very small number of proper classes were used for annotation, while the corresponding extension classes (GO_MF_EXT) were instead used for the large majority of classes. Both MOP and MOP_EXT did not have many examples in general. Even so, CRAFT contained 68,027 core training annotations with 28,516 core evaluation annotations and 112,415 core+extensions training annotations with 46,820 core+extensions

evaluation annotation over all the ontologies, which was a lot of data. More information about the CRAFT Corpus can be found at [233].

Table 2.1: Statistics for the concept annotations in the training (67-document) and evaluation (30-document) data sets for all ontologies in CRAFT. Note: avg stands for average.

Ontology	# training set annotations	avg/median # training set annotations per article	# evaluation set annotations	avg/median # evaluation set annotations per article
ChEBI	4548	68/45	2200	73/45
ChEBI_EXT	11915	178/142	5248	175/142
CL	4043	60/32	1749	58/32
CL_EXT	6276	94/64	2872	96/64
GO_BP	9280	139/108	3681	123/108
GO_BP_EXT	13954	208/158	5847	195/158
GO_CC	4075	61/33	1184	39/33
GO_CC_EXT	8495	127/91	3217	107/91
GO_MF	375	6/2	94	3/2
GO_MF_EXT	4070	61/34	1822	61/34
MOP	240	4/2	101	3/2
MOP_EXT	386	6/2	111	4/2
NCBITaxon	7362	110/90	3101	103/90
NCBITaxon_EXT	7592	113/97	3219	107/97
PR	17038	254/198	6409	214/198
PR_EXT	19862	296/246	7932	264/246
SO	8797	131/118	3446	115/118
SO_EXT	24955	372/341	9136	305/341
UBERON	12269	183/130	6551	218/130
UBERON_EXT	14910	223/165	7416	247/165

With all the data, performance of all systems was measured using the CRAFT Shared Task evaluation platform [170], which gave partial credit for predictions on either task. The benefit of a shared task framework was the standardization of the evaluation framework. Briefly, for the concept annotation task, the evaluation platform made use of the method proposed by Bossy *et al.* [234], which incorporated flexibility in matching both the boundaries (*i.e.*, the start and end character positions) of a predicted concept mention to the reference (span detection), and in the predicted ontology class identifiers (concept normalization). This flexibility allowed the scoring metric to assign partial credit to inexact matches in two different ways. Partial credit was assigned for overlapping boundaries using a Jaccard index scheme over the characters in the matches, and

partial credit for inexact ontology class ID matches was computed using a semantic similarity metric that made use of the hierarchical structure of the ontologies. The final score for a given predicted concept was based on a hybrid of both the boundary and ontology class ID match scores, and included precision, recall, F1 score, and an aggregate score called the slot error rate (see Bossy *et al.* [234] for the exact equations). The evaluation platform was made available as a versioned Docker container, in order to facilitate reproducibility and comparison of future systems to those that participated in the 2019 CRAFT Shared Task. Version 4.0.1_0.1.2 was used to evaluate the systems described here. All evaluation information can be found at the CRAFT GitHub site (specifically at `https://github.com/UCDenver-ccp/craft-shared-tasks/wiki/Concept-Annotation-Task-Evaluation`).

Table 2.2: Statistics for the concept annotation classes (or unique concept annotations) used in the training (67-document) and evaluation (30-document) data sets and for those added as additional training data for concept normalization for all ontologies in CRAFT. Note: avg stands for average.

Ontology	# training set annotation classes	avg/median # training set annotation classes per article	# classes added to training set	# evaluation set annotation classes	avg/median # evaluation set annotation classes per article
ChEBI	1463	22/18	58214	627	21/20
ChEBI_EXT	2852	43/38	58439	1167	39/39
CL	581	9/7	2163	253	8/9
CL_EXT	651	10/8	2168	286	10/10
GO_BP	1586	24/21	29213	682	23/23
GO_BP_EXT	2511	37/33	29301	1090	36/37
GO_CC	677	10/9	4052	212	7/6
GO_CC_EXT	896	13/12	4086	296	10/9
GO_MF	49	1/1	10951	19	1/1
GO_MF_EXT	738	11/11	10031	377	13/12
MOP	85	1/1	3574	32	1/1
MOP_EXT	108	2/1	3578	40	1/1
NCBITaxon	690	10/9	1175661	315	11/9
NCBITaxon_EXT	757	11/10	1175682	346	12/10
PR	1278	19/18	213371	466	16/16
PR_EXT	1534	23/22	213531	588	20/19
SO	1216	18/18	2256	544	18/19
SO_EXT	3172	47/47	2405	1409	47/48
UBERON	2048	31/24	14057	1040	35/31
UBERON_EXT	2409	36/29	14113	1217	41/38

We report on the performance, as well as the necessary resources and efficiencies of the algorithms. Some computations were performed on a contemporary laptop (MacBook Pro Mid 2015), but many required the use of an NIH-funded shared supercomputing resource [235] that included:

- 55 standard compute nodes with 64 hyperthreaded cores and 512GB of RAM
- 3 high-memory compute nodes with 48 cores and 1TB of RAM
- GPU nodes with Nvidia Tesla k40, Tesla k20, and Titan GPUs
- A high-speed Ethernet interconnect between 10 and 40 Gb/s

We used both the CPUs and all GPUs. All computation was written in Python 3 with associated packages. All code and models can be found at: `https://github.com/UCDenver-ccp/Concept-Recognition-as-Translation`. For our concept normalization evaluation, the ConceptMapper baseline model for comparison can be found at: `https://github.com/UCDenver-ccp/Concept-Recognition-as-Translation-ConceptMapper-Baseline`.

Span Detection

Span detection was the first task of our two-part approach to concept recognition. Even though there was quite a lot of previous work on this task, we aimed not only to explore the state-of-the-art methods (using language models), but also to find low-resource methods (using a CRF) and explore whether we can exploit hyperparameter tuning of simpler methods for more complex ones that build on the simple methods (using BiLSTMs). Thus, the goal was to explore the performance and resources of six canonical span detection algorithms [130, 222, 236] using the CRAFT shared task framework [170]. The underlying target representation for all algorithms were BIO tags [187], which were used to label each word in the sequence as beginning (B), inside (I), or outside (O) an ontological concept mention; for example, the BIO tagging for the text mentions “red blood cells" and “white blood cells" in the phrase “red and white blood cells" can be seen in Table 2.3.

Table 2.3: BIO(-) labeling for the discontinuous and overlapping ontology class mentions in the phrase “red and white blood cells" (from PMCID:15314655). (The O- would simply be O in the canonical BIO labeling.)

	red	and	white	blood	cells
labels for the annotation of *red blood cells*	B	O-	O-	I	I
labels for the annotation of *white blood cells*	O	O	B	I	I
final labeling	**B**	**O-**	**B**	**I**	**I**

We also aimed to determine if we could capture complex mentions including overlapping and discontinuous concepts using a modest extension of BIO tags. Discontinuous annotations and overlapping annotations [192] made BIO labeling challenging to define formally. Discontinuous annotations were annotations composed of two or more non-contiguous text spans. The difficulty in translating these text mentions to a sequence of BIO tags was how to label the intervening text between the two discontinuous spans, *i.e.*, as I (inside) or O (outside)? Since discontinuous annotations were rare (no more than 7% of the total words in all concept mentions (see Table 2.4), many previous systems ignored them (*e.g.*, [128, 186]). Here we introduced and evaluated a novel approach simpler than previous work (*e.g.*, [191, 192]) to represent such annotations. We observed that discontinuous annotations contained fewer words within their component text spans compared to the words between the spans, except for PR (see Table 2.4). Thus, to exploit more data, we created a new label, O-, which signified the text between the discontinuous spans of these annotations (see Table 2.3 for an example of “O-"). We refer to this expanded set of BIO tags as BIO(-) tags. We report the results on these discontinuous spans.

Overlapping spans created a different problem: the possibility of multiple labels for a word in the sequence, only one of which can persist for training purposes. For example “red", “and", and “white" all had two conflicting labels depending on the concept mention labeled (see Table 2.3). To address this problem, we prioritized the beginnings of concept mentions (B) to capture as many concepts as possible, even if some words in the multi-word concepts will not formally be captured. In our example then, the annotations based on the BIO(-) tags would be “red..." and “white blood cell" separately. Other approaches (*e.g.*, [237]) rewrote texts with conjunctions to unwind them, but that would not allow us to use the shared task framework, which depends on the

Table 2.4: Quantification of discontinuous and overlapping words in all concept mentions. All numbers are based on the number of words, not concepts.

Ontology	# words in all concept mentions	% words in discontinuous mentions	% words between text spans of discontinuous mentions	% words overlapping multiple mentions
ChEBI	5985	0.3%	0.6%	0.1%
CL	6576	4.3%	4.3%	2.6%
GO_BP	12956	5.2%	7.0%	1.6%
GO_CC	5864	1.5%	2.1%	0.5%
GO_MF	376	0%	0%	0%
MOP	257	0%	0%	0%
NCBITaxon	7696	0.03%	0.03%	0.03%
PR	23261	0.5%	0.2%	0.9%
SO	10348	1.2%	1.8%	0.5%
UBERON	15681	2.0%	2.3%	0.8%

unaltered text for evaluation. Our approach received partial credit for the overlapping concepts in the total score.

With CRAFT preprocessed into word-tokenized BIO(-) tag format, the preprocessed data was used to train, tune, and evaluate all span detection algorithms. The input to each algorithm was a sentence as a sequence of words represented as word features, word embeddings, character embeddings, or language-model-contextualized word embeddings that were then mapped to a sequence of BIO(-) tags as the output. Most tuning was parallelized among the ten different ontologies for time and memory efficiency. We report both the macro- and micro-F1 scores for tuning because the “O" (outside) category for the BIO(-) tags appeared significantly more often than any other tag. The micro-F1 scores take this into account and thus were significantly higher, whereas the macro-F1 scores only look at the raw scores, weighting each tag equally. Thus, the macro-F1 scores were low due to the infrequent “O-" tag. Each algorithm and its corresponding input representation is described in more detail below, focusing on the resources needed, the parameters to tune, and the final training framework used.

Conditional Random Fields (CRFs)

CRFs represented a low resource setting option for span detection because they use the least computational resources of any of the sequence-to-sequence approaches and can generally be

easily trained on contemporary laptops (MacBook Pro Mid 2015). Thus, if the performance is the best or close to the best, then it is a good low resources option. A CRF is a discriminative algorithm that utilizes a combination of arbitrary, overlapping and agglomerative observation features from both the past and future to predict the output sequence [129]. The words in a sentence were the input sequence along with features for each word including the case of the word, the three words before the word for added context, its part-of-speech tag, and the part-of-speech tags for the two words ahead of the word of interest. The output sequence was a BIO(-) tag for each word. The output sequences were directly connected to the inputs via the sentence- and word-level features. To optimize F1 score for each ontology, we tuned the CRF by conducting a randomized search of the hyperparameter space of both the L1 and L2 regularization penalties, with 3-fold cross-validation, which guaranteed that the global optimum was found [129]. With these optimal parameters, we evaluated the model using 5-fold cross-validation. All of this can be done on a contemporary laptop (MacBook Pro Mid 2015).

Bidirectional Long Short Term Memory (BiLSTM)

BiLSTMs represented a resource-intense option that much prior work has used and built upon. All of these works required significant efforts and resources put into tuning. To create efficiencies, we aimed to test whether the parameters for the simplest model could generalize to more complex models built on the simple one, instead of tuning more complex systems each time. An LSTM is a special form of a recurrent neural network that by default remembers information for long periods of time, allowing for more distant context to be used by the algorithm [188]. It is a chain of memory cells stitched together to allow long-term and short-term memory. Each memory cell contains four neural networks, including a forget gate (information to drop), a new information gate (information to add), and an output gate to the next cell (information to propagate). The LSTM architecture lends itself to sequence-to-sequence tasks such as ours, in which the input was a sequence of words and the output was the corresponding sequence of BIO(-) tags. However, the inputs must be vectors, so we used an embedding layer that mapped each word in the training data to a fixed-length vector to create the semantic vector space of the

training data. Due to varying lengths of sentences, we padded all input sequences and output sequences to the maximum number of words among all sentences (approximately 400). Additionally, since the context for a word can be before or after the word itself, we used Bidirectional LSTMs (BiLSTMs), which first run through the sequence forwards and then again backwards, thereby allowing the usage of the context on either side of a word [189].

Tuning a BiLSTM is resource-intensive, so we aimed to test whether the parameters tuned on a simple BiLSTM could translate across more complex BiLSTM models. The hyperparameter search attempted to optimize F1 score; however, unlike the CRF setting, there were no guarantees of finding optimal parameters. The large hyperparameter space and the high cost of each evaluation were barriers to identifying optimal values. We followed established heuristics for tuning a BiLSTM [238], using GPUs to speed up evaluation. The four main hyperparameters to tune were the optimizer, the batch size, the number of epochs, and the number of neurons or hidden units [238]. The classic optimizer is the stochastic gradient descent (SGD), but newer approaches have proven less variant and faster [239], including RMSProp [240] and Adam [241]. For named entity recognition tasks, many have used SGD [208–210, 212–214, 216, 218], and a smaller number have used RMSProp (Root Mean Square Propagation) [217, 219] and Adam [215]. We chose to focus on RMSProp, which updates the learning rate for each parameter separately and automatically using the exponential average of the square of the gradient in order to weight the more recent gradient updates over the less recent ones [240]. Thus, the learning rate required very little tuning, and so we chose not to change the default learning rate of 0.0001. Additionally, it implicitly performed simulated annealing in that it automatically decreases the size of the gradient step if it was too large, so as not to overshoot the minima. Note that that there was no guarantee that a global minima will be reached with any of these optimizers and testing each one was very time consuming.

More time and resources were necessary to tune the other hyperparameters. Since, batch sizes, epochs, and neurons are interrelated, multiple time-consuming experiments of different combinations of the three were needed to find the best parameters. The batch size determined the

number of examples to run together to help speed up the time-consuming training process. The larger the batch, the faster the runs, but the less nuanced the results. Typical batches range from 1 to 64, and thus we tested 18, 36, 53, and 106 (all of which cleanly divided a test set of 10% of the training data). The epochs were the number of repeat experiments to run, as the LSTM can result in very different results based on each random initial condition. The larger the number of epochs, the more time the LSTM takes to run. Due to our limited memory and time, we tested a small (10) and larger (100) number of epochs. For all runs, 10% of the data was used for validation for each epoch, as well as overall. Lastly, the number of neurons or the hidden states (N_h), can be determined with the formula [242]:

$N_h = \frac{N_s}{(\alpha * (N_i + N_0))}$

N_i = number of input neurons

N_0 = number of output neurons

N_s = number of samples in training data set

α = an arbitrary scaling factor, usually between 2 and 10

We found that varying α between 2 and 10 yields between 3 and 12 neurons. Testing every number between 3 and 12 would be very time consuming, and so we took the two extremes of 3 and 12 to test.

Overall, tuning a simple BiLSTM is very time consuming. We conducted 16 experiments (one optimizer, four batch sizes, two epoch sizes, and two neuron sizes) per ontology to find the optimal hyperparameters for the BiLSTM for each of the ten ontologies (160 experiments total). In terms of evaluation, F1 score cannot be optimized directly with an LSTM; instead, the models aim to minimize errors. We chose the categorical cross-entropy as the loss function, as it is the default loss function for multi-class classification problems such as ours, in which four categories were used for BIO(-) tagging. After all this work, training and tuning a more complex BiLSTM would be even more time-consuming. Thus, we tested the hyperparameters tuned here for all other more complex BiLSTM approaches (specifically, BiLSTM-CRF, char-embeddings, and BiLSTM-ELMo), in hopes that we can save resources going forward.

BiLSTM combined with CRF (BiLSTM-CRF)

A BiLSTM-CRF is the architecture of a regular BiLSTM with a CRF as the last layer [208]. The BiLSTM provides the feature weights for the CRF, which provides sequence-level features. The tuning processes would be exactly the same as for the previously discussed BiLSTM; however, we used the same tuned hyperparameters for each ontology found for the BiLSTM and then added the CRF layer on top. This determined if the simple BiLSTM parameters could be used in a more complex model, a BiLSTM-CRF. Thus, since the tuning for the BiLSTM was already done, the added CRF layer was trained and tuned on CPUs, saving significant time and resources.

BiLSTM with character embeddings (Char-Embeddings)

The BiLSTM and BiLSTM-CRF approaches used word embeddings for all of the words in the training data; thus, any word not in the training set cannot be translated to a BIO(-) tag and will be unknown. Our goal was also generalizability to words not seen in the training data and so to combat this, we tried a different underlying sequence representation based on the characters, creating character embeddings for each word. As each word was a sequence of characters, this approach can create a representation for any unknown word using these character embeddings. Training the character embeddings added a significant amount of CPU time and memory. Thus, we again utilized the same parameters as the tuned BiLSTM to save time and resources.

BiLSTM and Embeddings from a Language Model (BiLSTM-ELMo)

To provide even better word representations that included context and generalizability, we tested our simple BiLSTM parameters applied to BiLSTM-ELMo. It was a BiLSTM with a new underlying sequence representation from the language model ELMo [222]. The original ELMo was a language model trained on the 1 Billion Word Benchmark set that included approximately 800 million tokens of news crawl data from the general-domain WMT 2011. ELMo representations were contextual (with modeling of word polysemy), deep (as the BiLSTM was pretrained on a large text corpus), and character-based (thus allowing for representations of words unseen in training). In testing the BiLSTM-ELMo with our simple BiLSTM parameters, our resources did not suffice and we ran out of memory quickly. Thus, we experimented with the same

aforementioned batch sizes and found that a batch size of 18 could run for all ontologies. Thus, we used the optimal hyperparameters with a batch size of 18 for each ontology, saving us from needing more resources.

Bidirectional Encoder Representations from Transformers for Biomedical Text Mining (BioBERT)

Lastly, we tested and determined the resources for another state-of-the-art language model, BioBERT [130], that has been shown to perform quite well for span detection with limited tuning. BioBERT was a biomedical-specific language model pre-trained on biomedical documents from both PubMed abstracts (PubMed) and PubMed Central full-text articles (PMC) based on the original BERT architecture [130]. Briefly, BERT was a contextualized word representation model pre-trained using bidirectional transformers. It then used a masked language model to predict randomly masked words in a sequence from the full context on either sides of the word (instead of scanning one direction at a time), creating bidirectional representations. BioBERT added the additional layer of biomedical-specific training data, as it is known that general-domain-trained algorithms do not usually perform well in the biomedical domain [243, 244]. For our task, we chose to use the BioBERT+PubMed+PMC model because it was the most similar to CRAFT, the articles of which all appear in PMC. Due to the generalizability of BERT and BioBERT, they require minimal fine-tuning to utilize for other tasks, especially for this task since the documents in CRAFT were most likely included in the PMC training data. Thus, we utilized the default fine-tuning parameters for named entity recognition, including a learning rate of 1×10^{-5}, a training batch size of 32, evaluation and prediction batch sizes of 8 each, and 10 training epochs [130]. Similar to the LSTM models, validation was done within each epoch. Training and tuning BioBERT required minimal resources, aside from a GPU for speed.

Concept Normalization

With text mentions identified as biomedical concepts, the final step in the concept recognition task was concept normalization, *i.e.*, the normalization of the detected spans or text mentions of concepts to their respective unique ontology class identifiers (class IDs). For example, the text mention “white blood cell" would be normalized to the class ID CL:0000738

and the text mention “red ... blood cell" to the class ID CL:0000232 (see Figure 2.1). Our goal was to create a machine-learning based model that can generalize beyond the training data. To do so, we extended the idea from Hailu *et al.* [186] to determine if machine translation works for concept normalization, “translating" from text mentions to class IDs. Surprisingly it worked.

Aiming to automate concept normalization beyond rule-based systems, we reframed and explored this task as a translation problem by translating the characters of all the text mentions to the characters of the ontology class identifier. To the best of our knowledge, this was the first attempt at such reframing. Usually, for translation from one language to another, there is an assumed underlying structure and semantics of both languages that is captured at least in part by any algorithm that aims to automate the translation process. For this task, the input in the form of English text mentions contained the structure and semantics of the English language [245], as well as the rich history of its development [246, 247]. On the other hand, the unique numeric class IDs to which these text mentions were annotated supposedly contained no structure or semantics [155, 194]. Thus, from the outset, one would not expect translation from text mentions to class IDs to work, yet our results suggest that it does. We designed a series of experiments to understand what signals in the ontology class IDs were important to the performance. We used ConceptMapper [196] run on CRAFT (v3.1.3) as a baseline to compare and evaluate performance.

To implement this idea, we used the popular Open Neural Machine Translation system (OpenNMT) [131] for machine translation. It implemented stacked BiLSTMs with attention models and learned condensed vector representations of characters from the training data, processing one character at a time. One layer of the sequence-to-sequence LSTM model included four main components: an encoder, a decoder, an attention mechanism, and a softmax layer. There were stacks of multiple layers of encoders, attention mechanisms, and decoders before the softmax layer at the top. The input to OpenNMT was the sequence of characters for the text mentions (*e.g.*, “w h i t e b l o o d c e l l", with each character separated by a space to show it was a sequence of characters and not words). The size of this input sequence for the encoder was the length of the longest text mention in the training data by character count, which could be from 1 to

1000 characters depending on the ontology (*e.g.*, “white blood cell" has 16 characters including the spaces between words). Any input that was shorter than the maximum character length was padded at the end with null characters. Then for each text mention, the output was the class ID, similarly in the form of characters (*e.g.*, for the input “w h i t e b l o o d c e l l", the output was “C L : 0 0 0 0 7 3 8" including the ontology namespace and the colon). Analogously, the output size was the maximum number of characters among the ontology class IDs, with added null characters if the sequence was shorter (*e.g.*, “CL:0000738" has 10 characters). The maximum number of characters in the class IDs ranged among the ontologies from 7 to 20 characters in the core annotation sets and 10 to 83 in the core+extensions sets. This discrepancy arose from the naming convention of extension classes with both the additional “_EXT" in the namespace along with the textual class IDs, as detailed in the materials for CRAFT. We used the default parameter settings to begin to explore this idea as this approach was resource-intensive, requiring large amounts of memory and CPUs.

Training OpenNMT required pairs of text mentions and class IDs (concept pairs). Since generalizability was our goal, in addition to the linked text mentions and class IDs of the CRAFT concept annotations for training, we also added the primary labels and synonyms (extracted from the .obo files distributed as part of the corpus) of classes not used for annotation in the corpus simply due to the fact that they were not mentioned in the articles of the corpus. Dictionary-based systems include the entire ontologies with synonyms in their dictionaries and so we too included all data. For each of these classes, we used the name of the class and its synonyms as the text mentions that map to it; for example, CL:0000019, the Cell Ontology class representing sperm cells, did not occur in the CRAFT training data, so we added to the training data its primary label “sperm" and its exact synonyms “sperm cell", “spermatozoon", and “spermatozoid" as quasi-mentions that map to this ontology class. (See Table 2.2 for counts of ontology classes whose primary labels and synonyms were added to the training data.) Note that CRAFT annotators curated lists of unused classes that either were too difficult to reliably annotate with and/or for which extension classes were alternatively created or used; the labels and synonyms of

these classes were not added as training data. By adding the metadata of these classes from the .obo files, we not only added a significant amount of training data (amounting to thousands of more classes per ontology), but also ensured that all current ontology classes were captured by OpenNMT (with the exception of those purposefully not used by the CRAFT curators). We discuss if adding more data helped or hindered generalizability.

The added ontology concept pairs only occurred once in the training data, but in the CRAFT annotations, there exist multiple pairs of the same string and class ID (the same text occurs in different places). We assessed how the frequency of a concept pair affects the results. With all of these concepts, training sets can be assembled at the type level (for which there was one mapping of a given text mention to a class ID regardless of frequency) or the token level (for which all mappings of text mentions to class IDs were included, even though some were the same string and class ID, only occurring in different places in the text). Training and testing with tokens took into account the frequency of occurrence in the corpus (token-ids), while using types ignored frequency of occurrence in training and evaluation (type-ids). The token-ids were used for the full end-to-end system as it captured all the data including frequency. However, we compared token-ids to type-ids as well to determine if there was a performance difference, and type-ids were used for some experiments to better understand how concept normalization as machine translation worked. We further explored the performance with and without the extension classes. The extension classes greatly increased the size of the training data, and had somewhat different performance characteristics. Tuning was over a 90-10 data split for training to validation over all tokens, and default training parameter settings were used. These experiments with differing amounts of data, helped determine the optimal amount of data for generalizability to the evaluation set.

As framing concept normalization as a machine translation problem was unconventional, we aimed to explore how this approach might be exploiting general semantic information in the ontology class IDs used as output (the core set only) by transforming them in various ways to see how performance changed. To start, the type-ids were compared to the token-ids, and going

forward all further experiments used the type-ids, as they were a smaller set and thus faster to run. If the frequency of the text mention and class ID mattered, then we would see a drop in performance from token-ids to type-ids. The next approach was to use the same IDs but scramble the relationship between text mention and class ID (“shuffled-ids"). Another was to replace the class IDs with random numbers of the same length, drawn without replacement as to have no repeats (“random-ids"). If there was information in the specific class IDs, we would expect to see a drop in performance for shuffled-ids relative to type-ids. If there were information in the distribution of class IDs but not in specific ones, then we would expect a further drop in performance for random-ids relative to shuffled-ids. We further tested to see if we could add information to the class IDs by alphabetizing them by the text mention and assigning consecutive IDs (“alphabetical-ids"). Text mentions that had similar prefixes (*e.g.*, proteins “BRCA1" and “BRCA1 C-terminus-associated protein") had consecutive alphabetical IDs (PR:089212 and PR:089213, respectively), potentially giving the sequence-to-sequence learner additional information. For all experiments, we maintained the ontology prefixes before the unique number identifiers and only changed the numbers (as seen in the example above for alphabetical-ids). However, not all text mentions mapped to class IDs solely with numbers. For example, the class ID for the text mention “phylum" was NCBITaxon:phylum. These types of text mentions and class IDs were rare in ontologies in the core set and thus were not changed in any experiments. In the evaluation of alternative output targets, we calculated the percentages of both exact matches between ontology class identifiers overall and on a per-character basis, as OpenNMT translates per character. Furthermore, we conducted error analyses of these runs to better understand what underlies the translation and to suggest future improvements.

Results

Overall we achieved near- or above-state-of-the-art performance on the concept annotation task of the CRAFT Shared Tasks using only machine learning, BioBERT for span detection and OpenNMT for concept normalization, with direct comparison to Furrer *et al.* [128] on the full end-to-end system using the corresponding evaluation platform as described above (see Tables 2.5

and 2.6). Not surprisingly, BioBERT outperformed almost all other span detection algorithms, except for the CRF (for GO_BP, NCBITaxon, GO_MF_EXT, and NCBITaxon_EXT). Note that due to resource limitations, we tested all span detection algorithms on the core set and applied only the top two algorithms (CRF and BioBERT) to the core+extensions set. The BiLSTMs overall did not perform the best, but it did seem that reusing the simplest model parameters in the more complex models of BiLSTM-CRF and a BiLSTM with character embeddings (Char-Embeddings) either maintained the same performance or sometimes increased performance, especially for the Char-Embeddings model. However, these simple model parameters could not be reused for the BiLSTM-ELMo model. OpenNMT performed quite well for concept normalization. Overall, in comparison to UZH@CRAFT-ST, among the core sets, our best models modestly outperformed the best system for ChEBI, CL, GO_CC, MOP, SO, and UBERON, whereas among the core+extensions sets it just barely outperformed the best system for CL_EXT, GO_CC_EXT, and MOP_EXT. Even for the ontologies whose results were lower than the best system, the results were usually in close proximity (within 0.10 F1 score), except for PR and PR_EXT with our results significantly lower than UZH@CRAFT-ST. Note that UZH@CRAFT-ST yielded the best performance for GO_MF among all ontologies, most likely due to the very few annotation classes included in the ontology (see Tables 2.1 and 2.2). Also, for some ontology annotation sets (specifically, ChEBI, CL, GO_CC, UBERON, CL_EXT, GO_MF_EXT, and UBERON_EXT), all systems (including UZH@CRAFT-ST) performed less competently, with F1 scores below 0.80. Even with this state-of-the-art performance, there is still room for improvement for all ontologies in the core and core+extensions sets.

This work not only provided a state-of-the-art machine learning concept recognition pipeline, but also the factors and resources that contributed to it for usability and reproducibility [195]. Since partial credit was awarded both for span detection and concept normalization, we evaluated each separately to understand what drove the performance in the full end-to-end system (see Tables 2.8- 2.14 for span detection and Tables 2.15- 2.22 for concept normalization). For span detection, most F1 scores were around or above 0.9. The small drop in performance from

span detection to the full system seems to mainly be from discontinuous spans and concept normalization. All span detection algorithms were able to capture some discontinuous spans using our modified BIO(-) tag representation (see Tables 2.10 and 2.11), but the performance was still quite low at around 0.10. For concept normalization, it does appear that machine translation was a salient avenue to explore, which at least for the core set was comparable to the state of the art on the task at hand (see Tables 2.15- 2.18). However, generalizability to concepts not seen in the training data was still quite poor overall with only 15% of the 6,940 unseen core concepts and 18% of the 8,151 unseen core+extensions concepts mapped correctly. At the same time, the character-level match was much higher and there seems to be structure in the OBO IDs (see Table 2.21 and 2.22). It seems plausible that we can add more structure to the IDs to increase performance potentially (*e.g.*, alphabetical-ids). There is still room for improvement for both span detection and concept normalization.

In terms of resources (see Table 2.7), the CRF was a good low-resource option compared to BioBERT for span detection. Our results suggest that we can save lots of resources by reusing the simplest model parameters (BiLSTM) in the more complex models as a starting point (see Tables 2.8 and 2.9). OpenNMT required a lot of resources for concept normalization. To conserve resources, we provide the best hyperparameters for all algorithms and ontologies.

Overall, it seems that translation was a salient avenue to explore for a machine-learning based approach to concept recognition. The rest of this section provides more details on the overall training resources, the span detection results, and the concept normalization results.

Training Resources

Training these algorithms required significant computational resources due to the amount of data, tuning, and optimizing necessary. It is important to consider access to hardware, memory, and time when deciding on which algorithms to use for a task. Thus, here we report on those factors to aid other users in making these decisions (see Table 2.7). We report on these resources for the core annotation set only, as those for the core+extensions annotation set were similar.

Table 2.5: Full end-to-end system evaluation on the core set comparing F1 score. For all results shown here, the span detection algorithm is listed, and the concept normalization algorithm was OpenNMT. UZH@CRAFT-ST was the best performing system from Furrer *et al.* [128] in the CRAFT-ST, shown as a comparison to our methods. The best-performing algorithm is bolded with an asterisk*.

Ontology	CRF	BiLSTM	BiLSTM-CRF	Char-Embeddings	BiLSTM-ELMo	BioBERT	UZH@CRAFT-ST
ChEBI	0.7882	0.6394	0.5027	0.5942	0.0550	**0.7885***	0.7700
CL	0.6779	0.5134	0.3859	0.5611	0.0526	**0.6994***	0.6657
GO_BP	0.7505	0.5137	0.3642	0.6182	0.0720	0.7405	**0.8037***
GO_CC	0.7225	0.1689	0.3049	0.3244	0.0506	**0.7762***	0.7645
GO_MF	0.9778	0.9770	0.9778	0.8906	0.3704	0.9783	**0.9838***
MOP	0.8129	0.7721	0.7158	0.5985	0.0930	**0.8742***	0.8705
NCBITaxon	0.9026	0.7736	0.8391	0.8518	0.0948	0.8910	**0.9694***
PR	0.4040	0.3136	0.2827	0.2732	0.0516	0.5295	**0.8026***
SO	0.8987	0.4106	0.4096	0.7815	0.0813	**0.9054***	0.9027
UBERON	0.7474	0.6812	0.5029	0.6901	0.0793	**0.7670***	0.7488

Table 2.6: Full end-to-end system evaluation on the core+extensions set comparing F1 score for the top two algorithms found in the core set. For all results shown here, the span detection algorithm is listed, and the concept normalization algorithm was OpenNMT. UZH@CRAFT-ST was the best performing system from Furrer *et al.* [128] in the CRAFT-ST, shown as a comparison to our methods. The best-performing algorithm is bolded with an asterisk*.

Ontology	CRF	BioBERT	UZH@CRAFT-ST
ChEBI_EXT	0.7891	0.8039	**0.8209***
CL_EXT	0.7381	**0.7491***	0.7484
GO_BP_EXT	0.7279	0.7353	**0.8138***
GO_CC_EXT	0.8738	**0.8983***	0.8936
GO_MF_EXT	0.6413	0.6255	**0.7438***
MOP_EXT	0.8000	**0.8651***	0.8437
NCBITaxon_EXT	0.8710	0.8624	**0.9722***
PR_EXT	0.4397	0.5188	**0.8011***
SO_EXT	0.7682	0.7829	**0.9187***
UBERON_EXT	0.7558	0.7711	**0.7714***

The availability of hardware including normal laptop CPUs, hyperthreaded CPUs, and GPUs, both changed the time and the possible experiments that could be tested. For most algorithms, CPUs were sufficient. ConceptMapper, used as the baseline for concept normalization, ran on a normal laptop using CPUs (MacBook Pro Mid 2015). The CRF, BiLSTM-CRF, and Char-Embeddings models ran in reasonable amounts of time on the hyperthreaded CPUs for all ontology annotation sets. However, as the time to train increased, we

switched to GPUs because they often speed up the processes significantly [248]. For the BiLSTM, we tuned it on GPUs and continued the runs on GPUs once we had the optimized parameters. BioBERT also performed significantly faster on GPUs than on CPUs. Some algorithms (such as that for the BiLSTM-ELMo model, for which the portion involving ELMo was the rate-limiting step) needed a GPU for practical running times due to the amount of tasks performed in it [248]. Access to GPUs was helpful in speeding up training time as were hyperthreaded CPUs, since advanced hardware was quite important for some algorithms.

The hardware also dictated the available memory, dictating which algorithms and how many ontologies could be run. Each algorithm was trained, tuned, and tested on each of the ten ontologies. For the BiLSTM alone, that was 160 experiments total. Thankfully, most algorithms required a small amount of memory, except for OpenNMT, which required more than 600GB to train all ontologies. To combat this memory requirement, we parallelized it among the ten ontologies, averaging around 60-80 GBs per ontology. We also parallelized BiLSTMs and BiLSTM-ELMo due to the sheer number of experiments. For the BiLSTM-ELMo, our resources were not sufficient to truly test if the BiLSTM hyperparameters would generalize to it because we ran out of memory quickly. We had to modify our experiment and the results were poor. Thus, memory was an important factor in choosing algorithms.

Time was also an important factor. Choosing the right hardware with the right memory also hinged on the training time of the algorithm. This is particularly relevant for a shared task, where time is very important for meeting deadlines. Time is also important for responsiveness and scalability [200]. We found that the fastest algorithm was the CRF, followed by BiLSTM-CRF and BioBERT. The UZH@CRAFT-ST also used BioBERT, and it took significantly longer than our BioBERT due to a differing number of epochs (55 and 10, respectively). The most time-consuming algorithms were BiLSTM-ELMo, with ELMo taking the majority of the time, and OpenNMT. In both cases though, similar to memory issues, we parallelized among the ten different ontologies to speed up the process. Still though, each ontology required 50-100 hours of

supercomputer time. Since time, memory, and hardware were valuable, we provide all models and the best performing hyperparamter information to hopefully speed up future experiments.

Table 2.7: Hardware, memory, and time used for training for all evaluated algorithms. A given training time specifies the total hours if training for all ontology annotation sets were run consecutively, but these can be parallelized by ontology. *Parallelized per ontology due to time constraints. **Runs significantly faster on GPUs. ***Total free RAM available. ConceptMapper runs on CPUs but has no training, as it is a dictionary-based lookup tool, hence the specifications as N/A.

Algorithm	Hardware	Training Memory (GBs)	Training Time (hrs)
CRF	CPUs	2-13	1-4
BiLSTM*	GPUs/CPUs**	17	29
BiLSTM-CRF	CPUs	7	15
Char-Embeddings	CPUs	30	84
BiLSTM-ELMo*	GPUs	42	700-1000
BioBERT	GPUs/CPUs**	5	20
UZH@CRAFT-ST BioBERT* [128]	GPUS	120***	200
OpenNMT*	CPUs	620	515
ConceptMapper [196]	CPUs	N/A	N/A

Span Detection Results

Overall, BioBERT and CRF performed best for span detection on both the core and core+extensions annotation sets (see Tables 2.8 and 2.9, respectively) as evaluated on the set of 30 held-out documents. In low-resource settings, the CRF would suffice. In general, performance was very good, with all ontology F1 scores for both sets above 0.77 and most above 0.90. Note that the best results were seen for GO_MF, MOP, and MOP_EXT, most likely due to the fact that there were relatively few annotations in these sets (see Tables 2.1 and 2.2). Even though the BiLSTM models did not have the best performance, the hyperparameters from our simple model were a good starting point for the more complex models, except for BiLSTM-ELMo. The BiLSTM-ELMo model performed the worst and was most likely the reason the full end-to-end system was so poor for this algorithm. Even with some poor results, all algorithms could detect some discontinuous spans, and some discontinuous spans could be detected by at least one algorithm for all ontologies except for ChEBI, ChEBI_EXT, and GO_MF_EXT, which have among the fewest discontinuous spans (see Tables 2.10 and 2.11 for the core and core+extensions annotation sets, respectively). Due to the rarity of discontinuous spans, these tables suggest that

we can just barely detect them and that more work is needed on these complex mentions specifically. Overall though, sequence-to-sequence algorithms performed quite well for span detection.

Table 2.8: Span detection F1 score results for all algorithms tested against the core evaluation annotation set of the 30 held-out articles. The best-performing algorithm per ontology is bolded with an asterisk*.

Ontology	CRF	BiLSTM	BiLSTM-CRF	Char-Embeddings	BiLSTM-ELMo	BioBERT
ChEBI	0.7234	0.6545	0.5000	0.5280	0.0620	**0.9091***
CL	0.8333	0.5882	0.3774	0.8000	0.0000	**0.9231***
GO_BP	**0.8677***	0.5498	0.3661	0.6346	0.0685	0.8646
GO_CC	0.9412	0.1379	0.2689	0.2581	0.1000	**0.9444***
GO_MF	>**0.9999***	>**0.9999***	>**0.9999***	0.8421	0.0000	>**0.9999***
MOP	>**0.9999***	>**0.9999***	>**0.9999***	>**0.9999***	0.0000	>**0.9999***
NCBITaxon	**0.9959***	0.8551	0.9440	0.9569	0.0711	0.9453
PR	0.4351	0.2979	0.2151	0.0995	0.0339	**0.8199***
SO	**0.9435***	0.4935	0.4897	0.8203	0.1059	0.9081
UBERON	0.7913	0.7206	0.4758	0.7440	0.0854	**0.8826***

Table 2.9: Span detection F1 score results for all algorithms tested against the core+extensions evaluation annotation set of the 30 held-out articles. The best-performing algorithm per ontology is bolded with an asterisk*.

Ontology	CRF	BioBERT
ChEBI_EXT	0.8802	**0.9291***
CL_EXT	0.8000	**0.9677***
GO_BP_EXT	**0.8800***	0.8516
GO_CC_EXT	0.8667	**0.9524***
GO_MF_EXT	0.9211	**0.9231***
MOP_EXT	>**0.9999***	>**0.9999***
NCBITaxon_EXT	**0.9959***	0.9919
PR_EXT	0.5598	**0.7717***
SO_EXT	0.8054	**0.8244***
UBERON_EXT	0.8418	**0.9157***

Our goal was for others to benefit from these extensive studies of alternative methods and hyperparameter selections, and so we present the tuning results for all algorithms since it was the most-time consuming and difficult part of this work (see Tables 2.12- 2.14). Running and evaluating the models was generally quite fast. Note that the micro-F1 scores were higher because

Table 2.10: F1 score results for detection of discontinuous spans for all algorithms tested against the core evaluation annotation set of the 30 held-out articles. (Note that there are no discontinuous spans in the GO_MF, MOP, and NCBITaxon sets.)

Ontology	Support	CRF	BiLSTM	BiLSTM-CRF	Char-Embeddings	BiLSTM-ELMo	BioBERT
ChEBI	14	0	0	0	0	0	0
CL	175	0.1176	0.1158	0.1171	0	0.0107	0.1818
GO_BP	272	0.0952	0	0.014	0	0.007	0.2742
GO_CC	14	0.1053	0	0	0.0526	0	0.3
PR	44	0	0	0	0	0	0.08
SO	45	0.0408	0	0	0	0	0.3
UBERON	118	0.04	0	0	0	0	0.0915

Table 2.11: F1 score results for detection of discontinuous spans for all algorithms tested against the core+extensions evaluation annotation set of the 30 held-out documents. (Note that there were no discontinuous spans in the MOP_EXT and NCBITaxon_EXT sets.)

Ontology	Support	CRF	BioBERT
ChEBI_EXT	19	0	0
CL_EXT	175	0.1164	0.1608
GO_BP_EXT	287	0.085	0.2651
GO_CC_EXT	30	0.1579	0.3721
GO_MF_EXT	20	0	0
PR_EXT	44	0	0.1224
SO_EXT	72	0.1505	0.2979
UBERON_EXT	133	0.0485	0.2128

the “O" (outside) category for the BIO(-) tags appeared significantly more often than any other tag and the macro-F1 scores were lower because of the infrequent “O-" tag.

The CRF required the least amount of resources and was good in a low-resource setting. The optimal L1 and L2 regularization penalties (see Table 2.12) were found in about three hours for each ontology. Of note was that for ChEBI, GO_BP, MOP, and NCBITaxon the model shared the same parameters and yielded very similar F1 scores. This may mean that these ontologies need more tuning with larger parameter spaces or that this signifies a universal parameter for most ontologies. Even still, most macro-F1 scores were above 0.6 except for GO_CC and PR, which were still above 0.5. The lowest score was for PR and the highest for CL, with room for improvement in both. These parameters could be reused for the same model on a similar task or as a starting point for training a new model with minimal resources.

Table 2.12: CRF tuning parameters and resulting tuning F1 scores. The overall memory usage for all tuning was 6 GB.

Ontology	L1	L2	Time (hours)	F1 Score (macro)	F1 Score (micro)
ChEBI	0.0862	0.000186	3	0.61	0.99
CL	0.00477	0.0473	3	0.73	>0.99
GO_BP	0.0862	0.000186	3.17	0.63	0.99
GO_CC	0.269	0.00892	3	0.55	0.99
GO_MF	0.215	0.00392	3	0.66	0.99
MOP	0.0862	0.000186	3	0.65	>0.99
NCBITaxon	0.0862	0.000186	3	0.61	>0.99
PR	0.00477	0.0473	3.25	0.54	0.97
SO	0.315	0.00578	3.22	0.66	>0.99
UBERON	0.221	0.003005	3.3	0.63	0.99

Even though none of the BiLSTM models achieved the best results, it seems beneficial to reuse the hyperparamters from a simple BiLSTM model for more complex models, at least as a starting point (see Table 2.8). Also, tuning the BiLSTM required a significantly longer amount of time to achieve optimal performance on the training data (see Table 2.13) compared to other algorithms. Thus, we provide the BiLSTM hyperparameters as a starting point to help others save resources.

All of the values tested for each hyperparameter were necessary for at least one ontology. On average, training took around 98 hours per ontology, which was due in part to the number of parameters tested. That being said, each tested parameter value for batch size, epochs, and neurons was required by at least one ontology for optimal performance. For batch size, the two most common, 106 (the largest value tested) and 36 (the second lowest value tested), were found to be optimal for four ontologies each, while optimal batch sizes of 53 and 18 were found for ChEBI and UBERON, respectively. As for epochs, an optimal number of 10 was found for most of the ontologies, while an optimal number of 100 was found for GO_CC, NCBITaxon, and SO. For neurons, an optimal number of 12 was found for all ontologies except for GO_MF, for which an optimal number of 3 was found. About the same amount of memory (6.5 GB) was used for tuning all ontologies.

In terms of optimization metrics, we calculated tuning macro- and micro-F1 scores, as for the CRF model. In this case, all micro-F1 scores were above 0.99, while macro-F1 scores were equal to or greater than those for the CRF model, with all scores greater than 0.6. This was in contrast to the results for the final end-to-end system, for which the CRF model outperformed all BiLSTM algorithms. The highest F1 score was that for GO_MF and the lowest for MOP (with GO_MF and MOP having the smallest amount of training data, see Tables 2.1 and 2.2). Overall, we saved a significant amount of resources by using the same hyperparameters for the BiLSTM-CRF and Char-Embeddings models, along with modified values for the BiLSTM-ELMo models, even though none of them performed the best in the full-system evaluation.

Table 2.13: BiLSTM tuning parameters and resulting tuning F1 scores that were used for the BiLSTM-CRF and Char-Embeddings models also.

Ontology	Batch Size	# Epochs	# Neurons	Time (hours)	Memory (GBs)	F1 Score (macro)	F1 Score (micro)
ChEBI	53	10	12	99	6.5	0.67	>0.99
CL	36	10	12	92	6.5	0.74	>0.99
GO_BP	36	10	12	99	6.5	0.68	>0.99
GO_CC	106	100	12	97	6.5	0.66	>0.99
GO_MF	106	10	3	108	8.4	0.99	>0.99
MOP	106	10	12	99	6.4	0.61	>0.99
NCBITaxon	106	100	12	95	6.5	0.96	>0.99
PR	36	10	12	95	6.5	0.71	>0.99
SO	36	100	12	98	6.5	0.66	>0.99
UBERON	18	10	12	97	6.5	0.68	>0.99

Our reuse of hyperparamters to save resources failed with BiLSTM-ELMo due to memory issues. The limit of our resources was reached. Even though this model performed the worst overall, we report the modified hyperparamters used and the training results because the span detection results were on par with the other algorithms (see Table 2.14).

The BiLSTM-ELMo model required the most resources for training. Due to memory issues using the simplest model parameters, we needed to restrict the batch size to 18 and thus took the best results from the tuning of the aforementioned BiLSTM with this batch size. Using a batch size of 18 (the smallest batch size we tested) resulted in some differences in the optimal

numbers of epochs and neurons compared to the BiLSTM results shown in Table 2.13. Most ontologies still required 10 epochs for optimal performance, but a few (GO_CC, GO_MF and MOP) changed from 10 to 100 or vice versa, with the rest remaining the same. For neurons, most ontologies remained at 12, but two switched to 3, including ChEBI and GO_CC, with GO_MF remaining at 3 as well. Comparing F1 scores to the previous algorithms, the tuning micro-F1 scores were exactly the same as those for the BiLSTM, but the macro-F1 scores either decreased slightly (for ChEBI, CL, GO_BP, GO_CC, GO_MF, and SO) or stayed the same (for MOP, NCBITaxon, PR, and UBERON) as those for the BiLSTM. UBERON was the only ontology for which the parameters remained the same from the BiLSTM to BiLSTM-ELMo because it already required a batch size of 18 for optimal performance. It is not clear why all of these shifts occurred, getting at the difficulty in tuning a BiLSTM [238]. Further, the training F1-scores were similar to the CRF and BiLSTM scores, but the evaluation performance was very low. It is also unclear why this is true. More work including time and resources are necessary to understand and tune the BiLSTM-ELMo model.

Table 2.14: BiLSTM-ELMo parameters and resulting tuning F1 scores. Due to limited resources, the batch size was 18 for all ontologies.

Ontology	Batch size	# Epochs	# Neurons	F1 Score (macro)	F1 Score (micro)
ChEBI	18	10	3	0.65	>0.99
CL	18	10	12	0.72	>0.99
GO_BP	18	10	12	0.66	>0.99
GO_CC	18	10	3	0.65	>0.99
GO_MF	18	100	3	0.66	>0.99
MOP	18	100	12	0.61	>0.99
NCBITaxon	18	100	12	0.96	>0.99
PR	18	10	12	0.71	>0.99
SO	18	100	12	0.65	>0.99
UBERON	18	10	12	0.68	>0.99

Concept Normalization Results

Overall, the implementation of concept normalization as a sequence translation task via OpenNMT performed well on almost all ontologies in both the core and core+extensions sets, providing the last step in a fully machine-learning based concept recognition pipeline (see Tables

2.15 and 2.17). OpenNMT significantly outperformed ConceptMapper, except for PR and NCBITaxon_EXT, on the 30 held-out document evaluation set. Further, our results suggest that OpenNMT provided some generalizability to concepts not seen in the training data (see Tables 2.16 and 2.18). However, a weird side-effect of this generalizability was that OpenNMT can predict class IDs that do not exist, *i.e.*, an ID that does not represent an OBO concept (see Tables 2.19 and 2.20). These weird non-existent class IDs seem to exist because it appears that OpenNMT has found some structure in the OBO class IDs (see Tables 2.21 and 2.22). Thus, concept normalization as machine translation is a worthwhile avenue to continue exploring for generalizability and to understand the structure in the OBO IDs.

OpenNMT significantly outperformed ConceptMapper, except for PR and NCBITaxon_EXT, on the 30 held-out document evaluation set, with high character-level matches especially (see Tables 2.15 and 2.17). Evaluation was performed at both the class ID level, in which a correctly predicted class ID in its entirety counts as an exact match, and at the character level, in which each character of the predicted class ID was evaluated and a fractional score was calculated for a partial match. For example, while a text mention of “Brca2" was annotated in the gold standard with the Protein Ontology class PR:000004804, it was predicted to refer to PR:000004803, which differed from the gold standard only in the last digit. (The predicted ontology class ID PR:000004803 represents BRCA1, which also differed from the gold-standard protein BRCA2 in only its last digit.) At the class ID level, this was a mismatch, but at the character level it received a score of $\frac{11}{12} = 0.92$, as 11 of the 12 predicted characters matched. Scores greater than 69% were achieved at the class ID level for all ontologies except for PR and PR_EXT. Comparing the core set to the corresponding core+extensions set, scores for the latter were higher than those of the former for half of the ontologies (ChEBI_EXT, CL_EXT, GO_CC_EXT, PR_EXT, and UBERON_EXT). The class IDs of the extension classes may be easier to predict because there was less variation of text mentions annotated with the extension classes and also possibly due to the use of English words in the extension class IDs (*e.g.*, ChEBI_EXT:calcium, CL_GO_EXT:cell) rather than the numeric IDs of the proper OBO classes.

At the character level, all match scores were above 69% for both sets. Scores for the core set were generally higher than those of the corresponding core+extensions set, except for CL and GO_CC. For all ontologies, in most cases, the character level match scores were near or greater than the class ID scores, indicating that many of the predicted IDs were partly correct (in that they contained correct substrings). Future work can look at these partial matches for improvements.

For OpenNMT all text mentions were normalized to class IDs and so it can capture unseen text mentions, getting at generalizability (see Tables 2.16 and 2.18). Overall, only 15% of the 6,940 unseen core concepts and 18% of the 8,151 unseen core+extensions concepts were normalized correctly. However, more than 50% of the unseen text mentions were normalized correctly for the MOP, SO, MOP_EXT, and SO_EXT sets, and only the NCBITaxon, PR, and PR_EXT sets had none of the unseen mentions normalized. Otherwise around 20-30% were normalized for the rest. On the character-level, the unseen concept results were much higher, suggesting that these models can generalize to some extent beyond the text mentions seen in the training data, most likely due to the character-level translation of OpenNMT. For ConceptMapper, not all text mentions were normalized to ontology class IDs, indicating false negatives. These false negatives were the main reason performance of ConceptMapper was so poor (*e.g.*, no class IDs were predicted for MOP). Since ConceptMapper was a dictionary-based lookup tool, any text mention not in the dictionary will not be captured, but OpenNMT can capture these unseen text mentions. Even though, the text mentions seen in the training data performed far better than the unseen ones on both the class-ID and character levels, our results suggest that OpenNMT was somewhat generalizable and this can be improved upon based on the character-level results.

Even though OpenNMT normalized all text mentions and seemed generalizable, not all predicted class IDs were correct or even real class IDs. These types of mismatches, especially a weird side-effect of our character translation, helped us better understand these results and elicit future research directions. At the class ID level, two scenarios arose: (1) A different real class ID was predicted or (2) a completely non-existent class ID was predicted. The aforementioned Protein Ontology example involving Brca1 and Brca2 was an example of the first scenario. This

Table 2.15: Concept normalization exact match results on the core evaluation annotation set of the 30 held-out documents compared to the baseline ConceptMapper approach. We report both the percent exact match at the class ID level and the character level. We also report the percentage of false negatives (FN) for ConceptMapper (*i.e.*, no class ID prediction for a given text mention). Note that for each ontology the better performance between OpenNMT and ConceptMapper is bolded with an asterisk* for both class ID and character levels.

Ontology	% OpenNMT Class ID	% ConceptMapper Class ID	% ConceptMapper FN Class ID	% OpenNMT Character	% ConceptMapper Character
ChEBI	**82%***	55%	41%	**94%***	58%
CL	**72%***	52%	12%	**92%***	77%
GO_BP	**82%***	29%	59%	**93%***	36%
GO_CC	**81%***	54%	44%	**91%***	55%
GO_MF	**98%***	0%	100%	**99%***	0%
MOP	**95%***	65%	34%	**99%***	66%
NCBITaxon	**87%***	86%	13%	**97%***	87%
PR	10%	**47%***	26%	**76%***	57%
SO	**97%***	75%	21%	**99%***	78%
UBERON	**78%***	64%	34%	**95%***	65%

Table 2.16: Exact match results for the unseen and seen text mentions (relative to the training data) for the core evaluation annotation set of the 30 held-out documents. Reporting the total number of mentions and the number of unique mentions along with the percent exact match on the class ID level and character level for both unseen and seen text mentions.

Ontology	Total/Unique # Unseen Mentions	% Unseen OpenNMT Class ID	% Seen OpenNMT Class ID	% Unseen OpenNMT Character	% Seen OpenNMT Character
ChEBI	345/148	17%	94%	69%	99%
CL	774/208	39%	98%	92%	>99%
GO_BP	727/367	17%	98%	65%	99%
GO_CC	301/85	29%	99%	67%	>99%
GO_MF	3/3	33%	>99%	70%	>99%
MOP	18/7	83%	98%	96%	>99%
NCBITaxon	81/52	0%	89%	72%	98%
PR	2926/388	0%	19%	71%	80%
SO	181/105	62%	99%	89%	>99%
UBERON	1584/514	20%	97%	80%	99%

also occurred with concepts both within an ontology subhierarchy, such as a prediction of ChEBI:24867 (monoatomic ion) for a gold-standard mention of ChEBI:23906 (monoatomic cation), which is a child of the former. One can see the striking resemblance between the two

Table 2.17: Concept normalization exact match results on the core+extensions evaluation annotation set of the 30 held-out documents compared to the baseline ConceptMapper approach. We report both the percent exact match on the class ID level and the character level. We also report the percentage of false negatives (FN) for ConceptMapper (i.e., no class ID prediction for a given text mention). Note that the best performance between OpenNMT and ConceptMapper is bolded with an asterisk* for both class ID and character level.

Ontology	% OpenNMT Class ID	% ConceptMapper Class ID	% ConceptMapper FN Class ID	% OpenNMT Character	% ConceptMapper Character
ChEBI_EXT	**86%***	64%	26%	**84%***	66%
CL_EXT	**82%***	67%	11%	**93%***	84%
GO_BP_EXT	**80%***	34%	44%	**76%***	38%
GO_CC_EXT	**93%***	80%	18%	**94%***	84%
GO_MF_EXT	**69%***	60%	30%	**69%***	64%
MOP_EXT	**92%***	64%	35%	**97%***	44%
NCBITaxon_EXT	83%	**86*%**	13%	**93%***	87%
PR_EXT	**15%***	9%	28%	**72%***	21%
SO_EXT	**92%***	19%	40%	**91%***	22%
UBERON_EXT	**81%***	68%	29%	**92%***	75%

Table 2.18: Exact match results for the unseen and seen text mentions (relative to the training data) for the core+extensions evaluation annotation set of the 30 held-out documents. Reporting the total number of mentions and the number of unique mentions along with the percent exact match on the class ID level and character level for both unseen and seen text mentions.

Ontology	Total/Unique # Unseen Mentions	% Unseen OpenNMT Class ID	% Seen OpenNMT Class ID	% Unseen OpenNMT Character	% Seen OpenNMT Character
ChEBI_EXT	476/188	32%	92%	67%	85%
CL_EXT	775/209	36%	99%	77%	97%
GO_BP_EXT	861/431	26%	89%	57%	78%
GO_CC_EXT	339/113	39%	99%	57%	98%
GO_MF_EXT	515/146	31%	83%	45%	78%
MOP_EXT	21/10	67%	98%	85%	>99%
NCBITaxon_EXT	123/79	1%	86%	68%	94%
PR_EXT	3114/429	0%	25%	66%	75%
SO_EXT	318/183	51%	94%	69%	92%
UBERON_EXT	1609/532	23%	96%	79%	95%

concepts; the only difference is the added “cat", which OpenNMT did not pick up on. However, we did receive partial credit in the final evaluation, as ChEBI:’monatomic cation’ is subsumed by ChEBI:’monoatomic ion’ in the ChEBI ontology. Note that the predicted and the true class IDs were not always very similar such as for Brca1 and Brca2 and were sometimes even very different.

The first scenario seems to be a false negative and we found some syntactic relationships between the error and the truth (more exploration in the Discussion section).

Much less frequent were occurrences of the second scenario (see Tables 2.19 and 2.20). For example, OpenNMT predicted the class ID SO:0002000 and there was no such SO class. The correct class ID was SO:0001179 (U3 snoRNA), which was quite different. It is unclear how to handle these non-existent IDs, aside from assigning it as an automatic error. Interestingly, these errors were evenly distributed between training and validation; there was not much difference between the percentage of predicted non-existent class IDs for the training and validation sets. However, there did seem to be quite a difference between those and the evaluation set, with more predicted non-existent class IDs in the evaluation set, especially for GO_MF (50%) and GO_CC_EXT and NCBITaxon_EXT (more than 20%). Intriguingly, there were no predicted non-existent class IDs in any of the CL data sets nor in the MOP_EXT training and evaluation sets. This could be a proxy for whether these ontologies could stay within the “vocab" of the class IDs only using true class IDs (0% means fully within vocab), and not predicting non-existent ones (>0%). For the second scenario, it is unclear how to interpret these predicted non-existent class IDs and we explore them in the Discussion section.

Table 2.19: Percentage of predicted non-existent class IDs out of the total number of predicted mismatch class IDs for the core set for the training, validation and evaluation sets.

Ontology	% Non-existent Class IDs in Training	% Non-existent Class IDs in Validation	% Non-existent Class IDs in Evaluation
ChEBI	3%	4%	11%
CL	0%	0%	0%
GO_BP	2%	2%	2%
GO_CC	2%	4%	1%
GO_MF	1%	1%	50%
MOP	0%	2%	0%
NCBITaxon	7%	7%	11%
PR	10%	10%	2%
SO	0%	1%	0%
UBERON	0%	1%	2%

Table 2.20: Percentage of predicted non-existent class IDs out of the total number of predicted mismatch class IDs for the core+extensions set for the training, validation and evaluation sets.

Ontology	% Non-existent Class IDs in Training	% Non-existent Class IDs in Validation	% Non-existent Class IDs in Evaluation
ChEBI_EXT	8%	8%	17%
CL_EXT	1%	1%	9%
GO_BP_EXT	2%	2%	3%
GO_CC_EXT	4%	4%	21%
GO_MF_EXT	2%	2%	4%
MOP_EXT	0%	6%	0%
NCBITaxon_EXT	9%	9%	25%
PR_EXT	7%	7%	5%
SO_EXT	1%	1%	2%
UBERON_EXT	1%	1%	1%

Another contributor to both generalizability and the mismatches, aside from the character level translation, seemed to be the amount of data and the apparent structure in the OBO class IDs (see Tables 2.21 and 2.22 at the class ID and character levels, respectively). Focusing only on the core set, concept frequency in the training data had a modest effect on the performance of OpenNMT, except for NCBITaxon and PR. Interestingly, the match results for the token-ids relative to those for the type-ids were the same or very close (CL, GO_MF, and SO) or slightly decreased at both the class ID and character levels, except for MOP at the class ID level and GO_CC at the character level, for which the performance for the type-ids increased. Note that at the class ID level, the match result for PR was already very low for token-ids (under 10%), so any drop was comparatively minuscule. For NCBITaxon, performance completely dropped to zero at the class ID level, indicating that frequency of annotation concepts (which was taken into account in token-ids) was necessary for performance (which is analyzed further in the Discussion section). Whether frequency of concepts helped improve performance depended on the ontology.

At the character level, the drop was nowhere near as drastic as the class ID level, especially for NCBITaxon, due to partial-match scoring. Looking at some of the mismatches for both type-ids and token-ids, different errors were made. For example, for “penicillin" (annotated with ChEBI:173334 in the gold standard), ChEBI:7986 (which is the class ID for “pentoxifylline") was

predicted. These class IDs were nothing alike, but we can see similar structure (specifically, “pen" and “in") in the English concept mention and label of the predicted ID, which was identified by the token-ids method but not by the type-ids method. Our results show that it was better to use the token-ids approach for all ontologies except for MOP, for which the type-ids approach performed better. This suggests that more data was useful.

We also found that the class IDs have some semantic content or structure that may be contributing to the mismatches (see shuffled-ids and random-ids in Tables 2.21 and 2.22). As expected, we did see a decrease in performance from type-ids to shuffled-ids to random-ids at both the class ID and character levels for at least half of the ontologies. The match results for ChEBI, NCBITaxon, PR, and UBERON reduced to 0% for both shuffled-ids and random-ids at the class ID level, whereas those for CL, GO_MF, MOP, and SO generally maintained their level of exact matching between type-ids, shuffled-id, and random-ids, with random-ids performing the best for GO_CC of all experiments (only slightly). Note that NCBITaxon was already at zero for the type-ids. Furthermore, since OpenNMT works at the character level, we could see the structure breaking at the character level (see Table 2.22): for all ontologies there was a large decrease from type-ids to shuffled-ids to random-ids, except for GO_CC, GO_MF, MOP, and SO. This may suggest some accidental structure in the shuffled-ids and/or random-ids generated. The results at the class ID and character levels suggest that OpenNMT identified some structure in the current ontology class IDs.

Adding in more semantic content to the ontology class IDs may be a promising avenue to increase performance. We added in more semantic content by alphabetizing the text mentions and it seemed to perform close to the token-ids results for most ontologies, although it may have added in too much structure (see the alphabetical-ids columns in Tables 2.21 and 2.22). At the class ID level, the token-ids approach performed the best for all ontologies; however, the alphabetical-ids approach recovered the losses from the shuffled-ids and/or random-ids approaches to perform close to the token-ids results for most ontologies, except for CL, GO_MF, MOP, and NCBITaxon, with a similar trend at the character level. This suggests that the alphabetical-ids approach may

have imposed too much structure on the ontology concepts, as incremented alphabetized IDs were given to all text mentions even if they were annotated with the same class ID. For example, the class ID GO:0097617 was used for “annealing", “hybridization", and “hybridizations" mentions, which in the alphabetical-ids approach were mapped to GO:05728, GO:15701, and GO:15702, respectively. Therefore, in the predictions, hybridization was mapped to GO:15702, which was one number different but incorrect nonetheless, whereas annealing was predicted correctly. Thus, performance may be boosted by incorporating stemming and lemmatization into ID mapping, *e.g.*, providing the same alphabetized ID to “hybridization" and “hybridizations" and a different one for “annealing". Future work can explore other mapping methods to both help understand and improve upon the good performance of machine translation for concept normalization. Machine translation is a promising avenue for concept normalization.

Table 2.21: Exact match results for the concept normalization experiments on the core evaluation annotation set of 30 held-out documents. We report the exact match percentage at the class ID level. The highest percentage is bolded and with an asterisk*.

Ontology	Token-ids	Type-ids	Shuffled-ids	Random-ids	Alphabetical-ids
ChEBI	**82%***	65%	0%	0%	78%
CL	**72%***	**72%***	69%	70%	56%
GO_BP	**82%***	79%	64%	27%	52%
GO_CC	81%	80%	81%	**84%***	76%
GO_MF	**98%***	**98%***	**98%***	**98%***	51%
MOP	95%	**97%***	**97%***	92%	80%
NCBITaxon	**87%***	0%	0%	0%	0%
PR	**10%***	3%	0%	0%	8%
SO	**97%***	96%	**97%***	**97%***	96%
UBERON	**78%***	74%	0%	0%	74%

Discussion

By reframing concept recognition as a translation problem, we not only sidestepped the multi-class classification problem, but also achieved above or near state-of-the-art performance (see Tables 2.5 and 2.6) on the concept annotation task of the CRAFT Shared Task with direct comparison to Furrer *et al.* [128] via the corresponding evaluation framework. Overall, on the full system run for the core set, our approach using BioBERT with OpenNMT slightly outperformed

Table 2.22: Exact match results for the concept normalization experiments on the core evaluation annotation set of 30 held-out documents. We report the exact match percentage at the character level. The highest percentage is bolded and with an asterisk*.

Ontology	Token-ids	Type-ids	Shuffled-ids	Random-ids	Alphabetical-ids
ChEBI	**94%***	89%	60%	58%	**94%***
CL	**92%***	**92%***	86%	80%	65%
GO_BP	**93%***	91%	85%	56%	84%
GO_CC	91%	**92%***	**92%***	89%	82%
GO_MF	**99%***	**99%***	**99%***	**99%***	94%
MOP	99%	>**99%***	99%	99%	86%
NCBITaxon	**97%***	74%	73%	68%	74%
PR	**76%***	75%	40%	30%	74%
SO	**99%***	**99%***	**99%***	98%	96%
UBERON	**95%***	93%	69%	54%	88%

the best run from the participants in the CRAFT Shared Task (UZH@CRAFT-ST) [128] for six of the ten ontologies, whereas the latter modestly outperformed our best system for three of the four other ontologies (except for PR which was significantly lower). Furthermore, the BioBERT used here required less resources than the BioBERT for UZH@CRAFT-ST, but the OpenNMT in our system negated that (see Table 2.7). For the core+extensions set, our best-performing run slightly outperformed the best UZH@CRAFT-ST run for CL_EXT, GO_CC_EXT, and MOP_EXT, while the UZH@CRAFT-ST system outperformed our approach for all other ontologies. However, our BioBERT+OpenNMT system attained F1 scores within 0.2 of the UZH@CRAFT-ST scores for six of the seven ontologies, with PR_EXT as the exception. Overall, we achieved above or near state-of-the-art performance.

The errors that led to lower performance in these full runs resulted from issues with span detection, concept normalization, or an interaction between the two. For span detection, there were four classes of errors relative to the gold-standard text mentions: (1) the text mention was not detected at all; (2) the text mention was partially detected; (3) the text mention detected included extra text; and (4) a full extra text mention was detected. Case (1) was a false negative and case (4) was a false positive. For cases (2) and (3) though, partial credit was awarded in the full evaluation pipeline for detecting at least part of the text mention. For concept normalization,

the errors produced were most likely rather opaque and difficult to analyze (such as the non-existent class IDs) compared with *e.g.*, a dictionary-based approach like ConceptMapper (though there may be other reasons this approach would have trouble for these ontologies). Concept normalization as translation needs to be explored beyond the experiments done here, especially since ConceptMapper outperformed OpenNMT for PR and NCBITaxon_EXT. At the same time, some possible reasons for poorer performance may be related to the mixing of two different types of class IDs that include both numbers and English text as well as the varying lengths of the class IDs. The text class IDs tended to be rather long, and OpenNMT performs worse with longer sequence lengths compared to shorter ones (as evidenced by the results on the core+extensions set, where many extension classes had English text class IDs and were longer, and UZH@CRAFT-ST outperformed BioBERT+OpenNMT for all but three ontologies). For the interaction of the two tasks in the full system, case (1) in span detection will lead to no normalization, propagating the error from span detection and resulting in a false negative also for concept normalization. For cases (2) and (3), however, it was possible for the normalization step to still correctly identify the class ID even though the text mention was slightly incorrect. On the other hand, for case (4), the class ID determined will always be wrong because the text mention should not have been detected in the first place, as OpenNMT always outputs a class ID for each inputted text mention. Thus, it is important to understand the implications of a two part system to understand the errors in the final system.

It is also important to focus on each aspect of the full system and each aspect can be used separately for different tasks as well. Overall, span detection algorithms performed very well for all ontologies with all F1 scores for the best algorithms above 0.81 for the core set and 0.77 for the core+extensions set (see Tables 2.8 and 2.9). The lowest, PR_EXT, seems to be suffering from predicting extra text mentions (false positives) as in case (4) of the span detection errors mentioned above. For concept normalization, the exact match percentage on the class ID level was above 72% for the core set (except PR) and above 69% for the core+extensions set (except PR_EXT) (see Tables 2.15 and 2.17). In this case, both PR and PR_EXT were very low at 10%

and 15%, respectively, at the class ID level. At the character level, the exact match percentages were much higher at 76% and 72%, respectively, but these were still lower than most of the other ontologies. Looking at the errors produced for PR specifically, it seems that the English text class IDs error mentioned above was at play for PR_EXT especially, as well as the fact that PR has longer numeric class IDs to begin with and many acronyms and other synonyms compared to other ontologies. As for the interaction of the two tasks, the best F1 scores of the full system were all above 0.69 for the core set and above 0.74 for the core+extensions set, except again for PR and PR_EXT, respectively (see Tables 2.5 and 2.6). It seems that the very poor results of concept normalization for PR and PR_EXT were to blame for these poor full-system results, for which many of the spans were detected correctly but normalized incorrectly. Thus, there is still room to improve concept recognition for all core and extension ontologies, especially for PR and PR_EXT. Future work can directly compare to this work using the CRAFT Shared Task as their framework with the built-in evaluation platform [170].

The goal of this work was not only performance accuracy but also efficient use of resources. If the level of resources needed to train, tune, test, or use these models was outrageously high, then it would not be feasible to extend beyond the ontologies here or use these models for much larger collections of texts such as all of the publications catalogued in PubMed. Thus, some models were less useful than others due to their resource consumption for tuning and training (see Table 2.7), even if they performed better. Recent research quantified the exorbitant financial and environmental cost to deep learning algorithms for natural language processing [249]. They suggested that since researchers have unequal access to computational resources, they should report their training time and hyperparameter sensitivity and perform a cost-benefit analysis (the benefit in terms of accuracy), and they should prioritize computationally efficient hardware and algorithms. They also acknowledge that ideally researchers should have equal access to computational resources. For this work, having access to a GPU greatly sped up tuning and training times, which was in line with other research findings [248]. Compared to the BioBERT runs in UZH@CRAFT-ST [128], the CRF run on a CPU and BioBERT on a GPU were

the most efficient algorithms because the CRF remained on the sentence level and BioBERT fine-tuned a pre-trained language model (not taking into account the pre-training), respectively. If one does not have access to a GPU, or has a large dataset, the CRF would be most efficient for very similar performance, especially for GP_BP, GO_MF, MOP, NCBITaxon, SO, GO_BP_EXT, MOP_EXT, and NCBITaxon_EXT, where the CRF either outperformed BioBERT or remained the same for span detection (see Tables 2.8 and 2.9). In terms of state-of-the-art language models, if one has a small dataset or access to a GPU, then BioBERT was preferred over the other language model, ELMo. ELMo required the most resources and a GPU, and thus may not be practical for this task, as the simple model parameters could not even be tested. This was in line with previous research comparing BioBERT and ELMo [223]. Along with ELMo, the BiLSTM may not be practical for this task, as it required a large amount of resources to tune (see Table 2.13). However a promising avenue for reducing BiLSTM resources was to reuse (or at least start from) the simple model parameters for more complex models to yield sometimes only very slight drops or even increases in performance. For concept normalization, OpenNMT also used a large amount of resources (CPU threads, memory, and time), and other machine translation algorithms should be explored in the future with respect to resources and performance. We also report on all of this to help with reproducibility and future research into this topic [195].

As mentioned above, for span detection, BioBERT performed best, followed closely by the CRF for both the core and core+extensions sets. All algorithms that included a BiLSTM performed worse on this task. However at least one BiLSTM-included model performed well (within 0.2 of the best-performing model) for all ontologies except GO_CC and PR (see Table 2.8). There is always room to tune the BiLSTM parameters more, as BiLSTMs are difficult to effectively tune in general [238]. These results for BiLSTMs were in line with previous methods, which found that BiLSTMs combined with other algorithms performed well on span detection [189, 208–210, 212, 213, 216, 218]. Thus, future work can explore the parameter reusage more, tune the algorithms more, and separately tune the BiLSTM-ELMo component.

All span detection algorithms also detected discontinuous concept mentions for at least one ontology, one of the most difficult types of concept mentions to detect. A recent review on recognizing complex entity mentions, including discontinuous mentions, found that the problem was complex and not yet solved [191, 192]. We offered a new simple approach looking at the words between the discontinuous spans of the mentions, recognizing that they were greater in number compared to the words in the discontinuous spans providing more training data. In the evaluation documents, we were able to detect some discontinuous spans, even if only a few, for all ontologies, except for ChEBI, which contained the fewest discontinuous spans (see Tables 2.10 and 2.11). As reviewed by Dai [192], more system development as well as curation of more examples of these complex mentions is necessary to improve performance.

Regarding concept normalization as sequence translation performed well on all ontologies except for PR and PR_EXT as discussed earlier (see Tables 2.15 and 2.17). There were several advantages to this translation approach. The primary one was that, for the classes of the Open Biomedical Ontologies used in CRAFT and in this work, the output to be predicted was a relatively short string of characters. This task was no longer a massive multi-classification problem with a choice among thousands of different classes. Treating the inputs as sequences of characters, rather than sequences of words, also addressed the problem of unknown or out-of-vocabulary text mentions, as the model can learn sub-word patterns, covering potential text mentions in the evaluation set that were unseen in the training set, whose fragments (character n-grams) may appear in the training process [250] (see Table 2.16 for the number of unseen text mentions). However, poor performance in general most likely stemmed from the added ontology concepts from the OBOs not seen in the CRAFT annotations (see Table 2.2 for column the number of classes added to the training set), which greatly increased the number of concepts (especially for NCBITaxon and PR). It was a bit more complex though, as performance for NCBITaxon was fine, whereas it was not for PR. For NCBITaxon this was the case because CRAFT mainly focused on the laboratory mouse in biomedical studies; thus, the taxon for mice (in addition to that for humans) was overrepresented in this collection, leading to the good results

for the token-ids approach (in which annotation frequency was taken into account) and the very bad results for the type-ids approach (in which there was only one mapping for each unique class). This suggests that for large ontologies the type-ids approach is not feasible, unless there is added training data for the most represented concepts in the corpus to boost the performance. For NCBITaxon specifically, this may mean that its model cannot generalize to articles not focused on laboratory mouse studies. On the other hand, for PR there was no one overrepresented protein as was the case for NCBITaxon, and so the large amount of additional OBO concepts was most likely confusing the model. It seems that the more annotation classes there were, the harder it was to train the model successfully. Note that ChEBI and GO_BP were the next largest ontologies, but their performance with the type-ids approach did not suffer as much as PR and NCBITaxon. Thus adding the labels of their unseen ontology classes seemed to help with generalizability. A further exploration of the performance with different quantities and sets of labels of additional ontology classes not used in CRAFT, such as concepts in other biomedical corpora, may help determine a better training dataset for concept normalization for some ontologies especially.

Combining span detection and concept normalization for the full run, our performance was near or slightly above the state of the art. One explanation is that the full evaluation pipeline gave partial credit to predicted IDs of ontology classes that have some semantic overlap with the correct annotation classes (*e.g.*, the aforementioned example involving ChEBI:'monoatomic ion' and ChEBI:'monoatomic cation'). This may also be due to the duplicate annotations in the training data for the token-ids approach, for which we included all annotations from CRAFT. However, it is interesting to note that the performances of the type-ids experiments were always the same or lower than those for the corresponding token-id experiments, except for MOP, which was higher (see Table 2.21). The duplicates in the token-ids approach may bias the algorithm to recognize more frequent concepts in CRAFT. On the other hand, a number of non-existent class IDs were predicted for almost all ontologies except CL, with more in the core+extensions set compared to the core set (see Tables 2.19 and 2.20). A further exploration of both these predicted

non-existent class IDs and the absence of them for CL may help explain how OpenNMT translates concept mentions to class IDs.

Our results show that the unique ontology class identifiers contained some semantic content or structure (see Tables 2.21 and 2.22), which probably gave rise to the predicted non-existent class IDs in the full run. This was contrary to the OBO Foundry recommendation of ontology class identifiers not having semantic content [155, 194]. There exists obvious semantic content in the IDs of the very small number of proper OBO classes (*e.g.*, NCBITaxon:phylum, representing taxonomic phyla) and the many CRAFT extension classes of the OBOs (*e.g.*, ChEBI_EXT:calcium, representing elemental calcium and calcium ions) that contain English words. Semantic content may also have arisen due to a curation process that produced groups of closely related classes with sequentially adjacent identifiers, (*e.g.*, PR:000004803 and PR:000004804, representing the BRCA1 and BRCA2 proteins, respectively). It appears that OpenNMT created a fuzzy dictionary matching, where generally the mappings were unique with some wiggle room when necessary. By shuffling or randomly assigning class IDs, we saw where this fuzzy dictionary broke the structure at both the class ID and character levels. It appears that OpenNMT was finding patterns between mentions of concepts with consecutively numbered class IDs, such as those for BRCA1 and BRCA2. In fact, many of the mistaken class IDs were only a few characters off from the correct ones (see Tables 2.15 and 2.17 for results at the character level). This therefore provides an opportunity for applying post-processing techniques to the predicted class IDs to fix the mistaken characters, similar to other rule-based error-analysis approaches (*e.g.*, [178]). Future work could include applying pre- and post-processing techniques to the results of OpenNMT concept normalization.

From another perspective, we attempted to boost performance by adding in more structure to the class IDs during the execution of the task, as in the alphabetical-ids experiment. The alphabetical-ids approach did not outperform the token-ids approach, and the performance was most likely dependent on how alphabetical the concepts in each ontology were. For example, the concepts “HaC" (CL:0000855) and “haematopoietic cell" (CL:0000988) have quite different class

IDs but were very close in the alphabetization of the CL classes. Furthermore, stemming and lemmatization may also help boost performance. Thus, a hash or mapping from the current ontology class IDs to the alphabetical IDs with some stemming and lemmatization, along with some post-processing techniques at the character level, may help boost the performance at the class ID level. Both of these avenues warrant future investigation.

The main limitations of this work were due to the data representations and limited amounts of data and resources in general. For span detection the BIO(-) tags proposed here did not properly capture neither overlapping spans nor many discontinuous spans, thus meriting future work. For concept normalization we chose not to change any of the class IDs even though they all had varying lengths within and between ontologies (particularly for the extension classes, whose IDs included much more English text and were longer). Exploring different representations, including different ways to map the text class IDs to number identifiers, may improve performance. In terms of the algorithms themselves, future work should include an exploration of the amount of data needed to train and develop these algorithms for this task. In general it is possible to tune all algorithms more fully, including tuning the learning rate, for which the default or suggested rate was used for all algorithms. In particular, the BiLSTMs were most likely not tuned fully and should be tuned additionally in future work, including using early stopping for determining a more precise number of epochs. However, we provide a good starting point for tuning further. Furthermore, the assumption that the two best-performing algorithms on the core annotations set for span detection would perform well on the core+extensions set may be false. Thus, future work should focus on the core+extensions set more for span detection, starting with the BiLSTMs.

For concept normalization as machine translation, we only explored one machine translation model, OpenNMT, providing preliminary evidence that reframing this problem as translation is a salient avenue to explore in future work on concept recognition. Further, exploration of other algorithms for machine translation may prove fruitful. Also, the unseen text mentions in the evaluation set hint at the generalizability of the concept normalization method, but the generalizability of all systems (the span detection methods, concept normalization methods,

and the full end-to-end system) to other biomedical corpora is unknown. It is unclear how to add synthetic data for span detection, but we did make sure to add all the class identifiers from the OBOs not seen in the training data for the concept normalization task. Future work should test the generalizability of all systems.

This generalizability problem is not foreign to biomedicine [251]. Systems trained and tested on one set of data can get different results on a new set of data. Here we only used CRAFT and we do not know how a different corpus may affect these metrics. However, the CRAFT corpus contained many concept annotations from many different ontologies [166, 168–170]. At the same time, biomedical scientists are interested in their domain of scientific literature, and so domain-specific corpora and ontologies should be looked at - failure of solutions that work on specific domains to work on general data is observed quite frequently in biomedicine [252]. If most biomedical scientists are interested in a specific domain then creating **unique solutions** for each **unique ontology** for each **unique domain** is worthwhile, and then there is no generalizability problem. This is not to suggest that scientists should not care about the entire literature, but instead suggests that focusing on each domain separately could benefit the whole literature. Concepts in each domain are specific to that domain (these specific ontologies already exist [253]) and need to be dealt with in their specific way. This motivated us to treat all ontologies separately, while also providing methods and starting points for any additional ontologies beyond the ones here. All of the methodologies presented can be applied to any ontology with a training and evaluation corpus.

Conclusion

In conclusion, machine translation is a promising avenue for concept recognition that sidesteps the traditional multi-class classification problem. We achieved state-of-the-art results on the concept annotation task of the 2019 CRAFT Shared Task with a direct comparison to previous results. Our span detection algorithms detected some complex mentions and our concept normalization method helped with generalizability. Given the amount of work that goes into shared tasks, shared task resources should be reused if possible. In general, resources need to be

taken into consideration for concept recognition and NLP at large. We provided a low resource setting option as well as reused hyperparameter tuning results from a simple model for more complex models to save resources. Future work should focus on the core+extensions annotations set more for span detection. For concept normalization, further exploration of why machine translation worked and an exploration of other machine translation algorithms may prove fruitful. As the generalizability of this system is unknown, future work should test the generalizability of all proposed systems: span detection, concept normalization, and the full end-to-end system. Any part of this system can be used in the future, as it was all trained, tuned, and evaluated separately.

Summary

This work tested the applicability of translation to concept recognition to create a purely machine-learning based end-to-end system that achieved state-of-the-art performance. We split the task into two subtasks (span detection and concept normalization) and evaluated performance combined and separately. Using a shared-task framework, we systematically characterized the factors that contributed to the accuracy and efficiency of several approaches to sequence-to-sequence machine learning for span detection. The CRF and BioBERT performed the best with the CRF requiring fewer resources. For concept normalization, we recast the classic multi-class classification problem as a machine translation problem and explored the reasons underlying the surprisingly good performance by evaluating a variety of alternative identifier schemes. OpenNMT performed quite well for concept normalization and it seemed to find semantic signatures in the class identifiers. Overall, our best-performing sequence-to-sequence systems performed comparably with the top performing system on the shared task (occasionally extending it by a modest degree), and some offered substantial efficiencies in the time and computational resources required for tuning and training. Furthermore, our analysis of the strengths and weaknesses of such systems suggested promising avenues for future improvements as well as design choices that could increase computational efficiency at a small cost in performance.

Contributions And Acknowledgements

This work was published in BMC Bioinformatics under their special call for Recent Progresses with BioNLP Open Shared Tasks [174]. This work was also based on earlier work focused on improving precision in concept normalization [178].

- **Boguslav, M. R.**, Hailu, N. D., Bada, M., Baumgartner, W. A., Jr, Hunter, L. E. Concept recognition as a machine translation problem. BMC bioinformatics. 2021;22(Suppl 1): 598. Available From: https://doi.org/10.1186/s12859-021-04141-4

- **Boguslav M**, Cohen KB, Baumgartner WA, Hunter LE. Improving precision in concept normalization. Pac Symp Biocomput. 2018;23:566–77. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5730334/

I would like to thank the other authors including William A. Baumgartner Jr., Lawrence E. Hunter, Negacy D. Hailu, Michael Bada, and Kevin B. Cohen.

For the first publication, all authors read and approved the final manuscript. MRB was responsible for the design and implementation of the project, including training and evaluating all models. NDH began this project as his dissertation work determining the main ideas and framework for this project that were adapted by MRB. MB supervised all of the concept annotation work for the CRAFT Corpus, as well as provided guidance and extensive discussions about the use of the corpus in this work. WAB created the CRAFT Shared Task with the evaluation pipeline, as well as provided guidance and extensive discussions on the shared task and computation for this work. LEH supervised the whole project.

I would like to acknowledge the BioFrontiers Computing Core for computing resources and support, especially Jonathon Demasi. I would also like to acknowledge Harrison Pielke-Lombardo for preparing the CRAFT corpus annotations in a convenient format, as well as Asmelash Hadgu who supported Negacy Hailu and his dissertation work that led to this manuscript. I acknowledge Lenz Furrer and Fabio Rinaldi for providing me information on their research. Also thank you to Katherine Sullivan for editing the manuscript.

For the second publication, MB (MRB) ran all experiments, analyzed the data, and wrote the final version of the paper. KBC analyzed the data and wrote the first draft of the paper. WAB performed precursor work and participated in running experiments. LEH conceived of and directed the project. All authors participated in analyzing the data and approved the final version of the paper. The work was aided by discussions with Michael Bada, Negacy Hailu, and Tiffany Callahan.

CHAPTER III

KNOWN UNKNOWNS: CAPTURING AND CLASSIFYING IGNORANCE

Background

One of the main goals of biomedical concept recognition is to help extract information from the literature about specific biomedical concepts [4, 120–122]. Sentences in the literature that contain these biomedical concepts pertain to both facts or known knowns and questions or known unknowns. Distinguishing between the two is fundamental depending on the information needs and both are necessary to conduct research. Information extraction and BioNLP more generally has mainly focused on the literature's role in current knowledge [122, 254] and not on the role it can play in future knowledge or questions, even though posing good questions is as fundamental to scientific progress as analyzing experimental results [1–3]. Thus, this work aims to extract information from the literature based on its role in future knowledge.

Questions are a driving force in the selection of research topics and approaches, as discussed in the philosophy of science literature [1, 2, 69–73, 255–258]. Capturing such questions formally could accelerate research by helping students, researchers, funders, and publishers understand the scientific question landscape or the next goals for scientific knowledge (**knowledge goals**). These questions are discussed in the scientific literature as statements about knowledge that does not exist yet, including goals for desired knowledge, statements about uncertainties in interpretation of results, discussions of controversies, and many others; collectively we termed them **statements of ignorance**, borrowing the term from Firestein [1]. An automated approach to identifying and characterizing these statements of ignorance with their entailed knowledge goals has a variety of significant use cases. One is to facilitate interdisciplinary interactions: a collection of formally characterized statements of ignorance could identify questions from other disciplines that new results might bear on, particularly when considering genome-scale data such as transcriptomics (*e.g.*, [34]). A systematic survey of scientific questions could be useful to a wide variety of scientists, ranging from graduate students looking for thesis projects (*e.g.*, [12]) to funding agencies tracking emerging research areas (*e.g.*, [33]). Another potential application is

longitudinal analysis, for example, tracking the evolution of research questions over time (*e.g.*, [35]). Further, identifying questions would allow us to query existing databases for information (*e.g.*, [36]). However, in previous attempts at such applications, neither statements of ignorance nor the role they play in future knowledge were the focus. (Only recently, after this work was published, was a search engine for challenges and directions for COVID-19 introduced with a focus on future research [17].) By introducing statements of ignorance to these applications there is a potential to develop the scientific ignorance landscape at scale and across disciplines, resulting in an accelerated research process.

No prior work related to ignorance theoretically and computationally including hedging, uncertainty, speculation, factuality, epistemics, and meta-knowledge with many overlaps between them, have focused on the role such statements play in future knowledge, *i.e.*, the entailed knowledge goal or next actionable step based on the statement. (Lahav *et al.* [17] does not focus on knowledge goals either.) For example,"Thus, depending on the cellular environment, the short- and long-term effects of Tax expression can be quite different" was a hedged statement [41], but did not capture a knowledge goal. On the other hand, the sentence, "The exact molecular function of SEPW1 protein is unknown to date" had a knowledge goal (to explore the exact molecular function of SEPW1), but was not hedged. Focusing on the knowledge goals can help accelerate research because it provides an actionable next step for researchers to take. Thus, there is a need to capture statements of ignorance and the entailed knowledge goals both theoretically and computationally.

Our goal is to identify and parse a research article into its statements of ignorance (similar to decomposition of complex questions in the Question Answering literature [259]). Here we describe a novel natural language processing (NLP) task: identification and characterization of statements of ignorance. We present an estimate of the extent of statements of ignorance in a sample of the full text biomedical literature, provide a theoretically-driven taxonomy of such statements, describe a manually annotated corpus of statements of ignorance and their categorization (along with novel annotation guidelines for the task), and demonstrate that

automatically identifying and classifying statements of ignorance is feasible. This work is the first step towards facilitating interdisciplinary interactions based on experimental results, creating a systematic survey of questions for researchers, tracking the evolution of research questions over time, and querying existing databases for potential answers to the questions found.

While these ideas and methods are generally applicable across biomedical research, we focused on one specific area, prenatal nutrition. This body of literature is small enough to be tractable and diverse enough to show feasibility. It is also significant to global health, since women are understudied [110, 111], especially when pregnant [112, 113], due to ethical and legal considerations and complexities. We hope that even this initial pilot demonstration will be significant, as it has the potential to find new research areas to explore and facilitate new interdisciplinary interactions that could advance the study of this underserved population.

Related Work

Previous work in both NLP and more broadly provided the foundation of our approach, but differed in a variety of important ways. There was no canonical name for the phenomenon we called **ignorance**, although prior work has used the related terms **hedging, uncertainty, speculation, factuality, epistemics,** and **meta-knowledge**, each of which has been defined to be somewhat different than our focus here.

Theoretically, our categorization of ignorance differs from previous work in our focus on how these types of statements play a role in future knowledge, *i.e.*, knowledge goals. The original linguistics phenomenon that sparked all these areas of research was hedging. Hedged statements in linguistics were statements that can be true or false to some extent [37]. Recognizing that scientific research articles included hedges, hedging was then defined more specifically within these articles as “any linguistic means used to indicate either a) a lack of complete commitment to the truth value of an accompanying proposition, or b) a desire not to express that commitment categorically" [38]. Hedging highlighted a focus on truth and facts. To help determine the levels of truth, research turned to uncertainty. To make the best determination, one must understand what they do not know. One of the first attempts to understand uncertainty theoretically was for

decisionmakers, especially for law [73]. Scientific uncertainty was defined as the "different kinds of potential error associated with descriptive scientific information" [73]. A taxonomy of six categories of descriptive uncertainty focusing on the errors that can occur within each was created, including conceptual, measurement, sampling, modeling, causal, and epistemic. In the bioscience field specifically, prior work sought to explore speculative language by presenting many examples of the phenomenon and determining that it was feasible for humans to annotate [27]. They focused on expressions of levels of belief including hypotheses, tentative conclusions, hedges, and speculations. Others have recast this phenomenon as factuality, alluding to a continuum that ranges from factual to counter-factual with degrees of uncertainty in between [43]. Lastly, still others [260] coined the term meta-knowledge to encompass different types of interpretive information including confidence levels, hypotheses, negation, and speculation. They determine five categories of meta-knowledge including manner, source, polarity, certainty level, and knowledge-type. All of these works focused on these phenomenon in relation to the current known knowledge (i.e., how certain, speculative, hedged, factual, or meta is the knowledge). We instead aim to focus on how these types of statements can create future knowledge (i.e., knowledge goals). Note that many of the same sentences from prior work may be identified as ignorance statements, but the categorization of the sentences would be different with a focus on the next actionable step. We draw heavily upon this previous work and shift the focus to knowledge goals.

With a focus on the current known knowledge, computationally, biomedical text mining has mainly focused on the identification and classification of ignorance statements in order to discard or de-emphasize it in favor of knowledge only [231], or simply to present it separately from factual information [20, 21, 23–25, 27, 39–67]. Only a few works [17, 260–263] went beyond the current knowledge to focus on future knowledge. For example, using such statements to determine if clinical questions in patient notes were answerable or not [261] or attempting to link an author's findings to statements of intended knowledge gain [260]. Further, Ramona Bongelli *et al.* [262] created uncertainty corpora for both scientific and popular biomedical articles in order to compare the amount of uncertainty and the categories of uncertainty markers

used by writers in each corpus. They lay broader claims about the levels of uncertainty in research ("(1) in all corpora, the percentages of uncertainty are always much lower than that of certainty; (2) uncertainty progressively diminishes over time in biomedical articles..." [262]), and the most common linguistic markers (modal verbs, markers of possibility rather than subjectivity, and third-person subject followed by modal verbs rather than first-person followed by mental verbs). Lastly, Chen *et al.* [263] and Lahav *et al.* [17] were perhaps the closest to our work here. Chen *et al.* [263] overarching goal was to capture the integral role that the epistemic status of scientific propositions play in scientific change, but they did not go beyond identification. Lahav *et al.* [17] "goal is to bolster the ability of researchers and clinicians to keep track of difficulties, limitations, and emerging hypotheses." They focused only on challenges and directions, when there were many other categories with goals for scientific knowledge, such as controversies. Further, they did not explicitly focus on goals for scientific knowledge to provide actionable next steps to researchers. Thus, an explicit and specific characterization of how these types of statements play a role in future knowledge as knowledge goals was necessary.

Introducing a novel NLP task, required defining the task, determining if it was feasible for humans to perform accurately, finding ways to automate it, and testing generalizability. Researchers began by leveraging the linguistic definition of hedged statements [37] to identify **lexical cues**, words or phrases that communicated uncertainty, and **scopes**, the linguistic scope of the hedge cue. Using the notions of cues and scopes, researchers proved that hedged statements could be identified and classified in the scientific literature through theoretical frameworks (including taxonomies and ontologies), the creation of corpora, and the creation of classification models to automate the task. A taxonomy is a "hierarchy consisting of terms denoting types (or universals or classes) linked by subtype relations" [151]. Relevant taxonomies [43, 46, 68–73] spanned all different domains including biomedical and more general domains. Paul Han *et al.* [69] provided an overview of many uncertainty taxonomies from both the biomedical and general domains prior to 2011. In terms of ontologies, four [74–77] described the degree of evidence or confidence underpinning a statement and thus were relevant to our work here. Using these

theoretical frameworks, many corpora were built to show that humans can reliably identify the phenomenon at hand [40, 41, 50, 55, 63, 78–81, 83–85, 264]. With these corpora, researchers aimed to automate the task through all different model types including rule-based, probabilistic models, machine learning, deep learning, and hybrids [20, 42–45, 47–49, 51–54, 56, 58–62, 64–67, 86–92]. Lastly, some works did all three, creating theoretical frameworks, determining it was feasible for humans to perform, and automating it [17, 19–27, 46, 95, 96]. This work also does all three: building an ignorance taxonomy driven by knowledge goals based upon the theoretical frameworks mentioned, creating a corpus with annotation guidelines, and utilizing the same modeling frameworks for a classification task.

Though we focus on English biomedical scientific articles, specifically in the domain of prenatal nutrition, prior work spanned different contexts and languages. The majority of previous studies focused on biomedical research articles, but there was also the clinical domain [19–22, 25, 41, 63, 68, 69, 95, 261], other scientific domains [40, 50, 67, 70–73, 82], and the informal language domain [26, 84]. In terms of languages, most considered English, though others studied Arabic [45, 55, 57], Chinese [22, 54, 83, 91, 96], Swedish [68], French [95], and Spanish [51, 85]. All the articles focused on languages other than English stated the need for more research development in these contexts. Furthermore, some works aimed to think across contexts and domains for their task [39, 46, 51, 56, 81], hinting at generalizability. We demonstrate generalizability in our work by applying our classifiers to a held out test set of prenatal nutrition articles that were unseen by the classifiers.

Methods

To achieve our goal of identifying and characterizing questions stated in the biomedical literature, we first formalized what it means to be a statement of ignorance and the knowledge goal such a statement entails, and then demonstrated that such statements of ignorance exist and can be identified, both manually and automatically, in the literature. To do this, we 1) developed a taxonomy of types of statements of ignorance, 2) created annotation guidelines for recognizing such statements, 3) validated their use through an annotation task to create a manually labelled

corpus, and lastly, 4) used the corpus to benchmark classifiers to automatically recognize statements of ignorance (see Figure 3.1). These resources help ground future research aimed at exploring the state of our collective scientific ignorance.

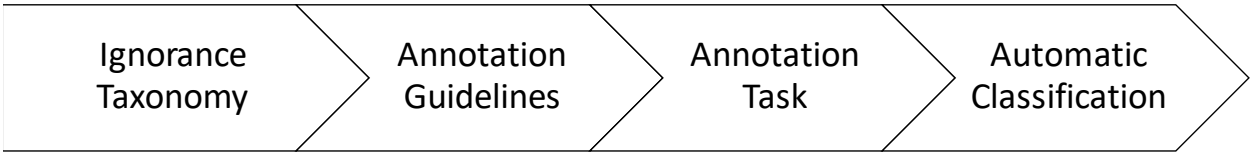


Figure 3.1: Methods Flowchart. A flowchart of the methods.

Materials

Scientific articles from our subject area of prenatal nutrition were taken from The PubMed Central Open Access (PMCOA) subset of PubMed [159], to ensure access to full-text articles and free sharing of data. We queried PMCOA using 54 regular expressions determined in consultation with a prenatal nutrition expert (Teri L. Hernandez), which included keywords such as {prenatal, perinatal, and antenatal} paired with keywords like {nutrition, vitamin, and supplement} (the full query can be found at `https://github.com/UCDenver-ccp/Ignorance-Question-Work`). In total we gathered 1,643 articles, subsets of which were used for each task below. All articles were provided in XML format and converted to text format parsing through the XML using a script in Java. All computation was written in Python 3, with its associated packages. In addition, the annotation task used Knowtator [265, 266] and Protege [267] to annotate the full-text articles; this allowed the ignorance taxonomy to be easily browsable like an ontology, and helped the annotators select the correct level of specificity for each lexical cue. Computation used a contemporary laptop (MacBook Pro Mid 2015) and an NIH-funded shared supercomputing resource [235] that included:

- 55 standard compute nodes with 64 hyperthreaded cores and 512GB of RAM
- 3 high-memory compute nodes with 48 cores and 1TB of RAM
- GPU nodes with Nvidia Tesla k40, Tesla k20, and Titan GPUs

- A high-speed Ethernet interconnect between 10 and 40 Gb/s

We used both CPUs and GPUs to train, evaluate, and predict statements of ignorance. All code and associated materials can be found at the following GitHub pages:

- `https://github.com/UCDenver-ccp/Ignorance-Question-Work`: our preliminary work including the full query, a preliminary ignorance taxonomy, annotation guidelines, an ignorance corpus of 60 documents, and preliminary classification models.
- `https://github.com/UCDenver-ccp/Ignorance-Question-Corpus`: the final ignorance corpus of 91 articles with 8 documents updated and 31 additional documents annotated in order to create a separate held-out test set for evaluation.
- `https://github.com/UCDenver-ccp/Ignorance-Question-Work-Full-Corpus`: the Python scripts to create both the corpus and classification models.

Ignorance Task Description

The goal was to show the feasibility of identifying and classifying statements of ignorance. We produced a gold standard corpus consisting of articles with labeled sentences as **statements of ignorance** along with the **lexical cue(s)** (words or short phrases) that distinctly signify it as such mapped to a categorization of **knowledge goals** (**ignorance taxonomy**). This was done through detecting spans of text either as a whole sentence or as words or short phrases. We also provided classification algorithms that aimed to automate the identification of both the statements of ignorance (sentences) in an article and the specific lexical cues (words or phrases). Taking the example above,"<The exact molecular function of SEPW1 protein is unknown to date>", the goal was for an annotator to identify or an algorithm to classify that this article sentence was a statement of ignorance as shown by the brackets around the sentence. From there, once the sentence was deemed a statement of ignorance, the goal was to identify or classify that unknown (shown underlined in the example) was the lexical cue that signified it as such. Note that one sentence can have multiple lexical cues that signify ignorance. The ignorance taxonomy helped to

distinguish between different lexical cues: the annotator and classifier also needed to map the cue unknown to a specific ignorance category that captures the knowledge goal of the sentence. Following the example, the taxonomy category was an *indication of an unknown or novel research topic or assertion* and the word unknown would be mapped to that by the annotator and the classifier. The knowledge goal was to explore the exact molecular function of SEPW1 protein further to gain any insights, which was implied by the ignorance category.

Ignorance Taxonomy

Indication of unknown or novel research topic or assertion or *unknown/novel* was only one type of ignorance category where the statement indicated something was not known or a lack of information on a topic. The knowledge goal was to explore the unknown further to gain any insights. Characterizing ignorance statements based on their entailed knowledge goals provides researchers with actionable next steps to continue to move research forward. To determine these different types of statements of ignorance and create the ignorance taxonomy, we first manually reviewed a subset of 736 article abstracts among the 1,643 prenatal nutrition articles from PMCOA. We conducted our task, focusing specifically on the lexical cues that signified that knowledge was missing or incomplete as in unknown in the above example. These lexical cues were then grouped together and organized into a taxonomy based on the knowledge goal each cue suggested. For example, the taxonomy category *unknown/novel* included lexical cues such as unknown, uncertain, still unclear, could not find, etc. So each cue mapped to a specific taxonomy category. Lexical cues and categories were also inspired by and added from existing work [1, 2, 41–43, 69, 70, 72] to create an initial **ignorance taxonomy** driven by knowledge goals. The majority of lexical cues corresponded to a single ignorance category, and thus implied a specific category assignment, though some cues, such as challenge, if so, and imply appeared in multiple categories and the correct category assignment then depended on sentence context. For clarity and simplicity, the taxonomy was a hierarchy of both broad (a higher level grouping of ignorance categories) and narrow categories (the ignorance category along with all lexical cues) based on the different types of knowledge goals. Lastly, the taxonomy was dynamic and iteratively updated

further during the annotation tasks as the annotators found lexical cues beyond what was gathered during this manual review, as described below. We present the final taxonomy after all the annotations were completed.

Annotation Guidelines

To fully describe our theoretical framework of ignorance and provide instructions for others to identify statements of ignorance (testing feasibility and reproducibility), we developed annotation guidelines for annotating articles based on the manual review. These guidelines not only helped annotators find statements of ignorance with lexical cues already in the ignorance taxonomy, but also helped them identify new cues to add. In particular, we recognized the importance of lexical cues and example sentences in the manual review to help guide the annotators in their decision to determine if a sentence was a statement of ignorance, the lexical cue, and the ignorance taxonomy category. Thus, the lexical cues gathered from the manual review were used to pre-process unmarked full-text articles, labeling all the cues in the text that were already pre-mapped to the ignorance taxonomy. The task was almost impossible without any pre-processing (unmarked full text articles). The annotators received these pre-processed articles and for each lexical cue marked, decided if it was correct or not: determined whether the sentence containing a marked cue was indeed a statement of ignorance, if yes, they needed to ensure that the ignorance category mapped to by the lexical cue was the correct one or change it, and if not then delete the pre-marked cue. Lastly, the annotators read the whole article and as our lexical cue list was not exhaustive, they identified additional statements of ignorance and their lexical cues that were not yet part of the taxonomy and added in the new mappings.

In order to help annotators reliably make a determination if a sentence and lexical cue signified ignorance, we provided many examples of lexical cues identified in statements of ignorance from the general PMCOA. We gathered and reviewed 150 sentences for each lexical cue to provide both positive and negative examples. For example, "<however, there is contradictory evidence from recent studies regarding the influence of IL-6 on insulin action and glucose metabolism> [4546]" represented a statement of ignorance of an *indication of alternative*

research options or controversy of research (or *alternative/controversy*), where we need to resolve disagreements about the influence of IL-6. At the same time, the sentence: “although yin and yang are contradictory in nature, they depend on each other for existence" was not a statement of ignorance and was a negative example for the contradictory lexical cue. Negative examples helped the annotators avoid the assumption that a lexical cue necessarily entailed a statement of ignorance. The annotation guidelines thus contained the ignorance taxonomy including all definitions and lexical cues for each taxonomy category with both positive and negative examples of statements of ignorance per lexical cue. The annotators referenced these examples throughout the annotation tasks and they will be referenced here to illustrate our methods.

The sentence containing the ignorance lexical cue was also important to capture to understand the biomedical context. We modified the scoping guidelines from BioScope [41] to capture the **scope** of the lexical cues: the biomedical knowledge the lexical cue qualified. Our annotations were different from BioScope, but the scoping principles were similar: to capture the sentence fragment that was the subject of the task at hand. We chose to capture the full sentence that contained a lexical cue as the scope due to difficulties in capturing just fragments. Note that future work can then determine the biomedical context more specifically by identifying the specific biomedical concepts included in these statements of ignorance using existing automatic concept recognition tools (*e.g.*, [121] and Chapter 2). In the example above, the brackets < > signified the scope of the statement of ignorance, *i.e.*, the full sentence. Note that there were no brackets around the negative example as it was not a statement of ignorance. The annotation guidelines described how to add the subject for the true statements of ignorance and how to remove or delete the annotation for the incorrect pre-marked cues. For the full guidelines see: `https://github.com/UCDenver-ccp/Ignorance-Question-Work/tree/main/Ignorance_Guidelines`.

Annotation Task

We conducted two annotation tasks to determine if identifying statements of ignorance was feasible and reproducible, and to create enough data to automate the task. All annotators were

computational biology researchers. In both cases, two independent annotators annotated batches of one to seven articles at a time checking in each article: the pre-marked cues, either deleting or adding the scope for those cues, and adding in any missed cues from the pre-processing. For example, using the positive example above, if the sentence "<however, there is contradictory evidence from recent studies regarding the influence of IL-6 on insulin action and glucose metabolism> [4546]" was in an article, the cue contradictory would be marked already based on the pre-processing and the annotator would note that, check that it should be mapped to *alternative/controversy*, and add the scope as seen with the brackets here. On the other hand, if the negative example above, "although yin and yang are contradictory in nature, they depend on each other for existence", was in an article, contradictory would be automatically marked through the pre-processing, however, the annotators would delete it as it does not signify ignorance. In terms of adding a missing cue, the phrase recent studies from the positive example was not marked and signified an *indication of proposed or incompletely understood research topic or assertion* (or *incompletely understood*), where more evidence was needed. The annotators would recognize that and mark recent studies to the narrow category of *incompletely understood*, thus adding the cue to the taxonomy for the next round of annotation. Note that the scope of recent studies was the same as contradictory in that sentence (*i.e.*, the full sentence) and was captured once for both cues.

To ensure the reliability of the annotations and create final gold-standard documents, we evaluated the quality and agreement of the annotations per batch using inter-annotator agreement (IAA) measures [163, 268]. The IAA measured how well the annotations agreed. Note that previous work showed that IAA did not bound the performance of classification algorithms, but was usually close to the limits [176]. Here, we calculated the F1 score between the two annotations, taking one annotation set as the "reference" and calculating precision and recall with the other one. The F1 score then was the harmonic mean between precision and recall (note that changing the "reference" flipped precision and recall, and thus the F1 score remained the same). The IAA was calculated on the exact text span of lexical cues or scopes chosen, as well as

ignorance category assignments. We also calculated fuzzy IAA when the category assignments matched but not the text span of the cue or the scope, or vice versa.

This fuzzy IAA allowed us to capture partial matches. Spans of the lexical cues can overlap when annotators recognized differing numbers of words for multi-word cues. For example, the annotators might have agreed on the ignorance category but highlighted either need or need to be in the sentence: "Thus doses of D vitamin and calcium supplementation, which may differ from those recommended in normal pregnancy, need to be carefully tailored in thyroidectomised patients." We took the maximum text span as the final span between the two annotators (need to be). This also occurred with the scope annotations, where use of the Knowtator software sometimes resulted in different text spans marked (an autocompletion error). Further, category assignments overlapped when one person annotated to the ignorance category implied by the specific lexical cue while the other annotated to the lexical cue. In the end we chose the narrowest applicable category or lexical cue in the taxonomy, to ensure we were capturing the correct information. For example, if one annotator mapped need to be in the above example to *indication of future research work* (or *future work*) and the other mapped it to the lexical cue need to be which by subsumption implied *future work*, the final annotation would be the lexical cue. All of these fuzzy matches were resolved during the adjudication process to ensure a high quality corpus. We report both the exact IAA as well as the fuzzy IAA. Ideally, both should remain over 80% to trust the annotations and the reliability of the guidelines [163, 268, 269].

The final product of the annotation tasks was a gold standard corpus to be used for automation and exploration. To finalize and maintain quality of the annotations, all disagreements were adjudicated, adjusting the guidelines and taxonomy accordingly with any newly identified cues along with the sentence or scope they were found in as a positive example. These discussions led to updates in all or some of the article annotations, the annotation guidelines, the ignorance taxonomy, and the lexical cue list. This process was repeated each time incorporating new updates.

The first annotation task included Mayla R. Boguslav (M.R.B.) and Elizabeth K. White (E.K.W.) as annotators and followed the process mentioned. Emily Dunn, a prenatal nutrition

researcher, annotated one article along with M.R.B. and E.K.W. and then stopped due to the time commitment. Articles to annotate were first chosen based on length starting with the shortest articles. As the IAA reached 80%, articles were chosen randomly using seeded randomness. M.R.B. and E.K.W. annotated in batches of four to seven articles and adjudicated each batch together, totalling 60 articles.

For the second annotation task, Katherine J. Sullivan (K.J.S.) and Stephanie Araki (S.A.) were the annotators and M.R.B. was a separate adjudicator. K.J.S. and S.A. were trained on eight random articles chosen from the 60 gold standard articles already annotated by M.R.B. and E.K.W. Any changes made to these articles (due to more experience with the task) were marked accordingly. After reaching the required IAA of 70-80%, new articles were chosen randomly using seeded randomness. For the first eight new articles (two batches of four), both annotators annotated the same articles as usual. After reaching IAAs of 80% or higher, we decided to divide the work: each annotator separately annotated 1-2 different articles and then adjudicated all with M.R.B. Since the classic IAA could not be calculated, we calculated an “F1 score" between the original annotation and the adjudicated version to see how reliable the single annotation was compared to the final adjudicated version. We continued annotation when this score stayed above 80%, signifying that the annotator was at least 80% correct in their annotations as discussed with the adjudicator. K.J.S. and S.A. in total annotated 39 articles, eight updated and 31 new ones, totalling a corpus of 91 articles, ready to be used to automate this task.

Automatic Classification

Automating the identification and characterization of statements of ignorance is the first step to help develop the scientific ignorance landscape at scale, ideally resulting in an accelerated research process. To determine if this task was feasible to automate, we tested standard classifiers, using the manually-labelled dataset to evaluate performance (see Figure 3.2). We split the data into a training set of 65 articles (approximately 2/3) and a held out test set of 26 articles (approximately 1/3). Note that the test data included 501 unique lexical cues with no sentence examples in the training data. To avoid batch effects based on the different annotators, we split

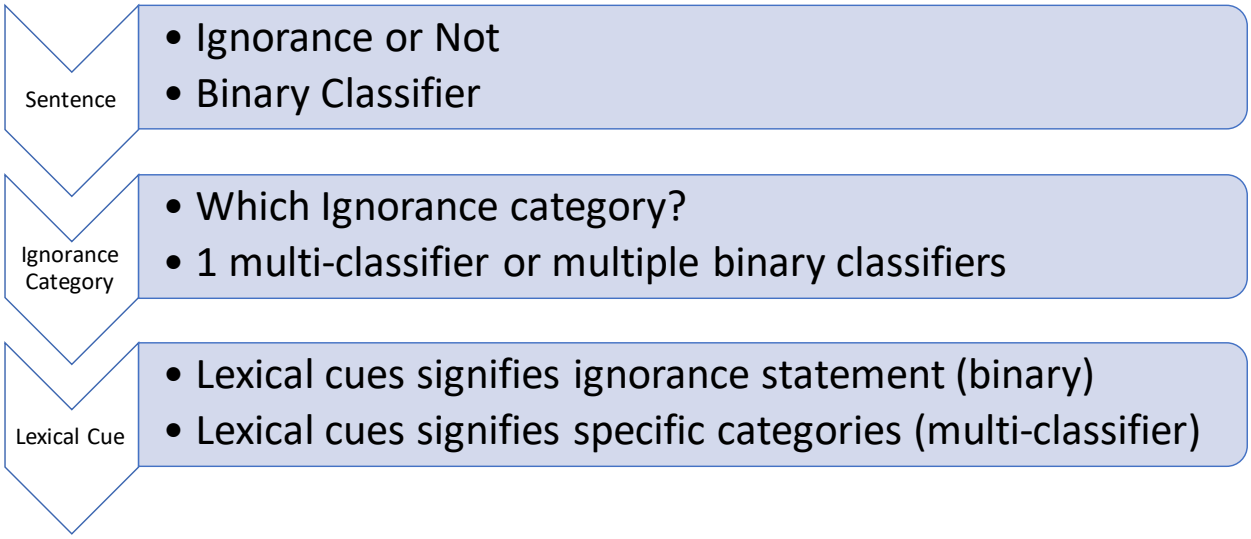


Figure 3.2: Classification Flowchart. A flowchart of the different classification problems.

Table 3.1: Data split for automatic classification in order of annotation tasks. Note that E.K.W. was Elizabeth K. White, M.R.B. was Mayla R. Boguslav, E.D. was Emily Dunn, Gold standard was the previous gold standard up to that point (the first row), K.J.S. was Katherine J. Sullivan, and S.A. was Stephanie Araki. *M.R.B. was an annotator along with the others. **E.D. only annotated one article along with the other annotators and then stopped. ***M.R.B. was the adjudicator in these batches.

Annotation batch	Total Articles	Training Articles	Testing Articles
First Annotation Task: E.K.W., M.R.B.*, (E.D.**)	52	37	15
Training: Gold Standard, K.J.S., S.A.	8	6	2
Second Annotation Task: K.J.S., S.A., (M.R.B.***)	8	6	2
Split Annotation Task: K.J.S., M.R.B.***	11	8	3
Split Annotation Task: S.A., M.R.B.***	12	8	4
Total Articles	91	65	26
Total Sentences	12,055	8,281	3,774
Total Words	416,866	285,439	131,427

each batch separately (see Table 3.1). Any labelled sentences (scopes) or words (lexical cues) in the training set served as the positive samples to classify and any non-labelled sentences or words served as negative samples.

As per the task description, classification can be made at the sentence or word level both as binary and multi-classification problems (see Figure 3.2). At the sentence level, the binary task was to determine whether or not a sentence was a statement of ignorance, labeled in the corpus as scope. Then since each statement of ignorance had at least one lexical cue labeled, the sentence can also be labeled by the ignorance categories of its lexical cues. For example, the positive

example above with lexical cues contradictory and recent studies (added as a new cue), would map to both *alternative/controversy* and *incompletely understood*. This now created a multi-classification problem: to map the sentences to the specific ignorance categories of their lexical cues. Similarly, we can focus on the lexical cues, with the binary task to classify whether a word in an article was in a lexical cue or not as labeled in the corpus (*e.g.*, the words contradictory, recent, and studies would be labeled as lexical cues). For the multi-classification task, the words would be mapped to specific ignorance categories (*e.g.*, contradictory to *alternative/controversy* and {recent, studies} to *incompletely understood*).

For both the sentence- and word-level multi-classification tasks, we created both one true multi-classifier and an ensemble - splitting the task into 13 smaller binary subtasks in which the ignorance category of interest was the positive case and all other sentences/words belonging to a different category were negative cases. Combining all 13 classifiers into an ensemble gave the full categorization for each sentence with no contradictions, since each classifier was predicting different categories. In all cases, each article was tokenized into sentences and then words used in the respective classification tasks. Further, we split the training data 90:10 for training and validation, and then evaluated separately on the held-out test set. For both tasks, we report the F1 scores on the held-out test set of 26 articles for (1) the one binary task (ALL CATEGORIES BINARY), (2) the 13 separate binary tasks that together create an ensemble multi-classifier (each category individually), and (3) the one multi-classifier (ALL CATEGORIES COMBINED - the macro-average).

For the sentence binary classification, we compared a simple artificial neural network (ANN) [270–272], bidirectional encoder representations from transformers (BERT) [220], and Biomedical BERT (BioBERT) [130] in order to compare a basic deep learning model to state-of-the-art language models. Full-text documents from PMCOA were first split into sentences, which were then processed by the CountVectorizer package in Python to tokenize words, remove punctuation and special characters, and produce a count vector of words as input to the ANN. The final tuned ANN consisted of a flattened layer followed by two dense layers, with

early stopping call backs, patience of 5, and dimensions 50 (to be proportionate with the large input shape) and 1. Overfitting was avoided by adding the training truncating functions (epochs ranged between 6 to 12). To allow for faster training and better generalization [273], we chose a batch size of 16 for training. We also applied the stratify function to ensure that the data splitting for the test was equal. Due to an imbalance of our data, we balanced it to the class with fewer samples before training to achieve a stable learning process. We verified this decision by training on the non-balanced data in a separate trial. For BioBERT, we used the vocabulary from BioBERT to train the base BERT model. For both BERT and BioBERT models, we used a batch size of 16 and a learning rate of 1×10^{-5} [130]. The epochs were tuned based on avoiding overfitting using truncating functions. We did not freeze the layers of the pre-trained BERT model and allowed the weights to keep updating during training for better performance. We compared performance of all three algorithms.

The sentence multi-classification problem, classifying each ignorance sentence into a taxonomy category based on its lexical cues, had a similar setup to the sentence-binary task, using a vector of word counts as inputs. We created one multi-classifier. This task though was more complex because some sentences mapped to multiple categories and the categories were not balanced. To combat all of these problems, we decided to also create an ensemble of binary classifiers for each taxonomy category that classified all sentences as either the category of interest or not (meaning the other taxonomy categories). All classifiers, one for each taxonomy category, were run over all the data allowing for one sentence to be classified in multiple categories easily in separate models (*e.g.*, the positive example before would be easily classified as both *alternative/controversy* and *incompletely understood*). Performance of the same three models were compared per ignorance category.

Not only was identifying the sentence as an ignorance statement helpful, but also identifying the specific lexical cues that signified the sentence as such provided a more precise focus on the ignorance occurrence, especially for sentences mapped to multiple ignorance categories. The word-level classification tasks were to determine which words in a sentence

indicated a lexical cue (binary task), and specifically for which taxonomy category (multi-classification). As our taxonomy was very similar to an ontology, we used our prior work in concept recognition (Chapter 2) and applied it to a different type of linguistic phenomenon. (Chapter 2 explored and evaluated some of the canonical algorithms for concept recognition over many different ontologies.) In particular, we made use of the best performing span detection algorithms, namely CRF [129] and BioBERT [130], to determine the words in all lexical cues given an article. CRF models were tuned with L1 and L2 regularization to avoid overfitting. For BioBERT, the NER baseline parameters performed quite well most likely because it was a similar task, and so we did not tune any other parameters [130]. The CRF and BioBERT were trained for each specific iteration of the word-level classification. Thus, we first word-tokenized the articles using WordPunctTokenize in Python. We then assigned each word a BIO- tag based on whether the word was at the beginning of a lexical cue (B), inside of it if it was a multi-word cue (I), outside of it meaning not a word in the lexical cue (O), or if the lexical cue contained a discontinuity (*e.g.*, *no...exist* where the "..." signifies a discontinuity), we labeled the words that exist between the lexical cues (O-). The input to the CRF and BioBERT to train then were the words with their target BIO- tag labels. When predicting, the input was the word-tokenized articles and the output was a BIO- tag for each word. To then determine what the lexical cue was, we re-assembled the BIO- tags by finding the B tags (single word lexical cues), combining the words with B and I in a row (multi-word lexical cues), ignoring the O labels (not lexical cues), and by combining B, I, and O- (discontinuous lexical cues). A more thorough discussion of BIO- tags can be found in Chapter 2.

For the word-level binary task, we took every lexical cue from all ignorance categories, classified the words into BIO- tags, and then re-assembled them into the lexical cues. Note that lexical cue overlaps were possible, but have no effect on performance because the goal was to identify any lexical cue no matter the ignorance category. At the word-level, the binary classifier can find new lexical cues not already in our lexical cue list by finding new combinations of words for example. So, we compared the predictions of the classifier to our set of lexical cues in the taxonomy to determine if the classifier was finding new ones. For the word-level

multi-classification task, we explored two different methods. First, we created an ensemble with binary models (CRF and BioBERT) for each ignorance taxonomy category similar to the sentence multi-classification task. At the same time, however, BIO- tags have the capability to encode which category the word was from (*i.e.*, B-*taxonomy category* or I-*taxonomy category*), and so we created one multi-classifier (CRF or BioBERT) for word-level multi-classification as well. This automation was the first step towards the scalability of the identification and characterization of statements of ignorance in the broader scientific ignorance landscape.

Results

Statements of ignorance employed a rich vocabulary

Researchers discussed ignorance in a myriad of ways in the scientific literature that fall into five broad categories and thirteen narrow ones. The manual review of 736 paper abstracts to develop a taxonomy of ignorance based on implied knowledge goals revealed that abstracts contained on average seven (minimum 0, maximum 24) statements of ignorance, involving 897 lexical cues. Subsequent refinement of 60 full text articles during the first annotation task added 993 lexical cues, and the additional 31 articles added 623 more cues, for a final ignorance taxonomy that included 2,513 lexical cues. Statements of ignorance employed a rich vocabulary.

These cues were distributed across 13 narrow categories combined into five broad categories. These included:

- *Indication of answered research question*
- Epistemic indication of research topic, assertion, or variables: *indication of unknown or novel research topic or assertion*, *indication of explicit research inquiry*, *indication of proposed or incompletely understood research topic or assertion*, *indication of indefinite relationship among research variables*, *indication of largely understood research topic or assertion*
- *Indication of anomalous or curious research finding*

- Indication of barrier to research: *indication of alternative research options or controversy of research*, *indication of difficult research task*, *indication of research problem or complication*
- Indication of future research activity: *indication of future research work*, *indication of future research prediction*, *indication of important consideration for future research work*

Note that the ignorance categories *indication of answered research question* and *indication of anomalous or curious research finding* were both broad and narrow. For the three broad categories that unify multiple narrow categories, “Epistemic indication of research topic, assertion, or variables" contained statements that answer the questions around evidence and how confident we were in it, in increasing order. “Indication of barrier to research" contained statements that indicated that research cannot move forward until the barrier was overcome, including multiple options, obstacles, or complications. “Indication of future research activity" included statements of future needs such as future considerations or work. For annotation purposes, these broad categories helped simplify the 13 categories.

Along with the ignorance categories we provide the full definitions, knowledge goals, and example cues for each narrow category (see Table 4.2). Each category separately also contained a rich vocabulary, as shown by the last column of the table. The increase in lexical cues between each annotation task argues that we have only begun enumerating the myriad of ways to signify a statement of ignorance. Further, our ignorance taxonomy was not only a categorization system of ignorance statements via knowledge goals, but also a depiction of the research life-cycle and how researchers discuss our collective scientific ignorance (see Figure 3.3).

Table 3.2: Ignorance Taxonomy: definitions, knowledge goals, example cues, and total cue count. The categories in bold were only narrow categories. Abbreviations are in italics.

Ignorance Category	**Definition**	**Knowledge Goal**	**Example Cues**	**Total Cues**

indication of *answered* research *question*	A statement of a goal or objective of a study that is attempted or completed during the study.	to find the answer(s) in the article; determine if the question(s) is (are) fully answered in the article	aim, goal, objective, our study, sought, to determine	64
indication of *unknown* or *novel* research topic or assertion	A statement that indicates something is not known (a lack of information), or information is presented for the first time (new or novel) and a significant amount of research is needed; not a statement about the absence of something.	to explore the unknown further to gain any insights	could not find, don't know, elusive, not...established, uncertain, still unclear	155
indication of *explicit* research *inquiry*	An explicit statement of inquiry (with a question mark or question word such as how, where, what, why).	to find answers to the question and/or discover methodologies that will help answer the question	?, what, where, wondered, why	19

indication of proposed or *incompletely understood* research topic or assertion	A positive or negative statement proposing a possible/feasible explanation for a phenomenon on the basis of limited evidence as a starting point for further investigation OR a statement that information is needed to support an assertion or claim, including both positive and negative statements. Either a statement that some evidence already exists, explaining how current findings support previous work, adding confidence to a claim OR a statement that information is limited, more research is needed or is ongoing including limitations – biases or short comings related to the study design and execution.	to gather more evidence to support the claim OR conduct more research to determine the validity of the claim; complete the partial picture; consider the short comings and biases for the next experiment and how it can be addressed.	a good understanding, believe, evidence...limited, has been suggested, hypothesis, no studies, possibly, preliminary stage, remains under investigation, still being discovered, support, trend	797
indication of *indefinite relationship* among research variables	A statement about a connection, link, or association between at least 2 variables; connectedness between entities and/or interactions representing their relatedness or influence.	to confirm the connection, link, or association between variables; determine the full underlying relationship between variables	affect, associated, correlate, factor, influence, interact, link, pattern, tend	198

indication of *largely understood* research topic or assertion	A statement staking a claim to the most likely explanation, relationship, or phenomenon; assumes that there is a good chance this understanding is correct.	to determine if the most likely option is correct or if another option is more feasible	almost all, assumed, concluding, evident, it is clear, most likely, thus	202
indication of *anomalous* or *curious* research finding	A statement of a surprising result, conclusion, observation or situation; the researchers were not expecting the result, conclusion, observation or situation but are intrigued by it.	to explore the surprising result, conclusion, or situation more and determine if the result, conclusion, observation, or situation is repeatable	appeared to be, interestingly, noteworthy, surprisingly	113

indication of *alternative* research options or *controversy* of research	Either an explicit statement of multiple (at least 2) choices, actions, approaches, or methods that need to be experimentally determined, including statements with an implied second option, such as "whether". This includes a statement of disagreement amongst researchers OR a lack of consensus OR at least two possible answers presented as results from different researchers - usually in reference to previous results and stated when results disagree with each other OR contradictions.	to determine the correct option or a better option and if there are disagreements, to determine the truth to break any disagreements	cannot rule out, claims, has been challenged, whether, whilst	221
indication of *difficult* research *task*	A statement of something not easily done, accomplished, comprehended, or solved; or a complicated thing with a multitude of underlying pieces or parts; heterogeneity; excludes medical complications.	to create methods to study the complicated system and to better understand any piece of the complicated system; potentially requires new experiments or better techniques	not feasible, remains...challenge, variability, rarely able to	98

indication of research *problem* or *complication*	A statement of issues, problems, mistakes, or medical complications that are cause for anxiety and/or worry.	to determine the gravity of the concern and determine if it needs to be dealt with before the next experiment or study	issue, error, insufficient, lack of reproducibility, publication bias, underestimated	98
indication of *future* research *work*	A statement of extensions, including next steps, directions, opportunities, approaches, or considerations of the described work that may be implemented at some future time point. This also includes a statement of suggestion or a proposal as to the next best course of action, especially one put forward by an authoritative body; advice telling someone the best action to take.	to determine the next course of action based on this future work proposal	additional research, are needed, continue to explore, further study, more...studies, recommend, warrants, worthy of closer attention	258
indication of *future* research *prediction*	A statement of extrapolation of given data into the future and/or from past observations, without reference to next steps.	to run the simulation or experiment to determine if the prediction is correct; publicize the outcomes of the study to the correct people	allow, expect, if so, serve as a basis, will	27

indication of *important consideration* for future research work	A statement calling for attention including an action needed to be taken immediately or information that needs to be disseminated immediately OR critical: being in or verging on a state of crisis or emergency OR urgently needed OR absolutely necessary.	to take the urgent action ASAP or distribute the knowledge ASAP	call for action, cautious, crucial, emphasis, global problem, high on the agenda, necessary, relevant to note, vital	263

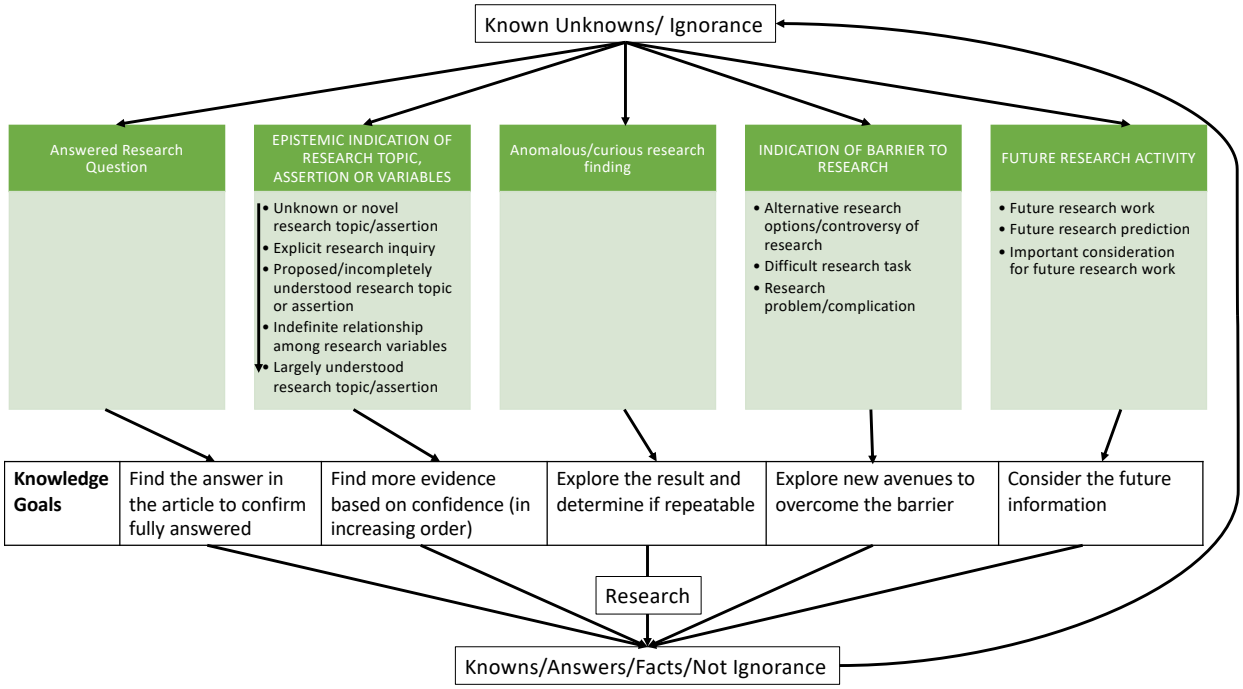


Figure 3.3: Ignorance taxonomy embedded in the research context: Starting from the top, research starts from known unknowns or ignorance. Our ignorance taxonomy is in green (an ignorance statement is an indication of each ignorance category) with knowledge goals underneath. Research is then conducted based on the knowledge goals to get answers; these then filter back to the known unknowns to identify the next research questions.

Table 3.3: Interannotator Agreement (IAA): IAA was calculated as F1 score for all annotation tasks. The IAA for the training was between the two annotators, not including the previous gold standard. *F1 score between annotator and final gold-standard version after adjudication with M.R.B.

Annotation batch	category IAA	scope IAA	fuzzy category IAA	fuzzy scope IAA
First Annotation Task: E.K.W., M.R.B., (E.D.) (60 articles)	78%	87%	79%	90%
Training: K.J.S., S.A. (8 articles)	77%	66%	78%	87%
Second Annotation Task: K.J.S., S.A. (8 articles)	81%	82%	81%	93%
Split Annotation Task: K.J.S., M.R.B.* (12 articles)	88%	92%	89%	95%
Split Annotation Task: S.A., M.R.B.* (12 articles)	89%	92%	90%	96%
All combined	82%	87%	83%	92%

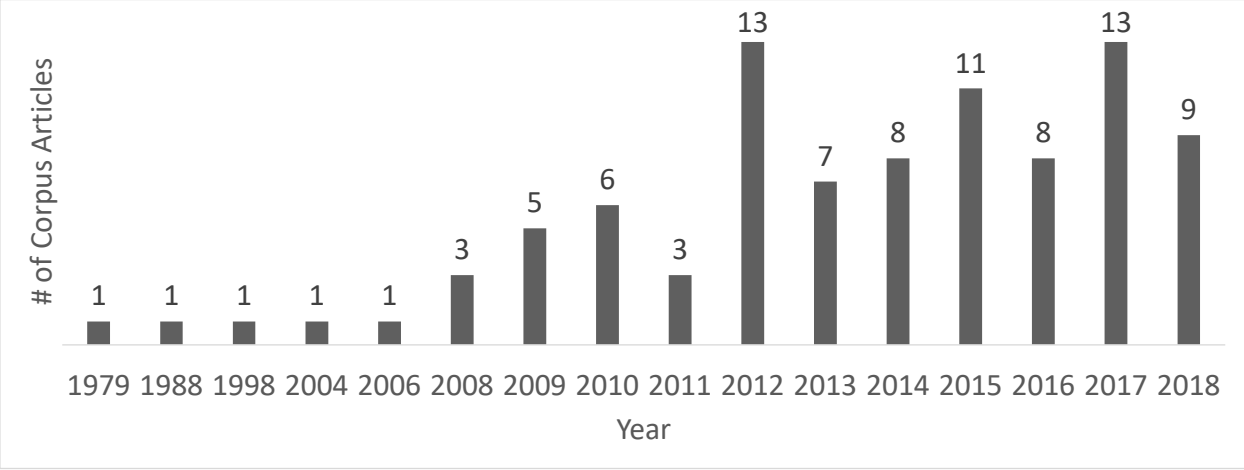


Figure 3.4: Article date distribution for the ignorance corpus (1979-2018).

Robust annotation guidelines yielded a high quality corpus

Our annotations guidelines seem robust, reproducible, and possibly generalizable because they yielded a high quality corpus over two different annotation tasks with five different annotators (see Table 3.3) and articles spanning the years 1979 to 2018 (see Figure 3.4). We can trust the annotations and reliability of the guidelines as the IAA was near or above 80% for the classic annotation tasks [163, 268, 269] and for the split annotations, the annotators were correctly identifying statements of ignorance around 90% of the time. Further, it was easier to train K.J.S. and S.A. for the second annotation task given the gold standard data of the first annotation task as a reference. We believe this annotation work could be extended to more articles and new annotators based on these results.

Within the guidelines themselves, the positive and negative examples for each lexical cue and the flexibility of the guidelines contributed to the feasibility, reproducibility, and generalizability of the annotation task. Surprisingly, almost every cue contained a negative example. Even for seemingly obvious cues such as unknown, there was a negative example. A statement such as “to make inference, the maximum likelihood method is applied to estimate the unknown parameters in the empirical log-odds ratio models given in (3)-(6)" was a negative example in that it was a description of a methodology that helps determine missing parameters. Further, the guidelines allowed for updates with the discovery of new lexical cues by the annotators. This flexibility greatly contributed to the robustness of the data.

Lastly, even with such guidelines not every article could be annotated for this task. K.J.S. and M.R.B. decided that one article (PMC4869271) was too difficult to annotate for this task because:

- The audience seemed to be public health workers whereas this task was more focused on researchers via scientific articles. This may be a true implementation paper.
- The article was full of quotes, making it very difficult to determine what was ignorance. Many of the quotes were also informal, meaning that it was unclear what was a quote or not. This further made the article difficult to follow.
- The article was most likely more of a technical report, public health education, government, and/or an interview rather than a scientific article based on the publication types from the NIH [274]. There did not appear to be much basic research in it.
- It was rather difficult to determine what was ignorance in the study itself versus the implementation of the policies being presented. The majority of the article seemed factual because it was mainly quotes from the public health workers.
- Neither K.J.S. nor M.R.B. were confident in their annotations of the article.

Thus, this article was excluded from the final corpus. (Our task does not apply to all articles, which is addressed more thoroughly in the Discussion.) Even with this exclusion, the high quality of the data gave us confidence in its use for any downstream applications, including exploration and automation.

The scientific literature was rich in statements of ignorance

The scientific literature, exemplified by prenatal nutrition, was rife with statements of ignorance representing a variety of different ignorance categories (see Table 3.4). This helped determine how researchers discuss knowledge goals in the literature both in terms of lexical cues and ignorance categories. More than half (56%) of the total sentences (12,055) in all articles were statements of ignorance, with nearly 20,000 lexical cue annotations and nearly 7,000 scope annotations. Every article had at least one ignorance lexical cue annotation and thus a scope also since the minimum number of annotations for both ALL CATEGORIES and SCOPE was one. At the same time, the minimum number of annotations for each individual ignorance category was zero, meaning that there was at least one article that contained no annotations to each specific category. The most represented categories were *incompletely understood* and *indefinite relationship* with over 6,000 and nearly 4,000 annotations across all articles, respectively. *Explicit inquiry*, *difficult task*, *unknown/novel*, and *future prediction* were the least annotated with under 400 annotations total for each category. Thus, research articles contain many statements of ignorance from a variety of ignorance categories. As all of these statements represent knowledge goals, researchers can take advantage of these statements as ideas for future work in all categories of ignorance.

We also determined the unique lexical annotation information and interestingly, the same small percentage of cues were used quite frequently. Many of the lexical cues were repeats of the same cue even with so many lexical cue annotations (see Table 3.5). In total there were 1,202 unique lexical cue annotations and as expected, all articles had at least one unique ignorance lexical cue annotation since the minimum number of unique annotations for ALL CATEGORIES was one. At the same time there was at least one article that contained no annotations to each

Table 3.4: Annotation Statistics Per Ignorance Category: Total number of lexical cue annotations in all articles and statistics per ignorance category. SCOPE was the number of sentences that contain at least one ignorance lexical cue. Note that all categories except for ALL CATEGORIES and SCOPE (had 1) had zero minimum number of annotations. Note max = maximum.

Category	# total annotations	average # annotations per article	median # annotations per article	max # annotations
answered question	619	6.8	7	23
unknown/novel	328	3.6	2	20
explicit inquiry	177	1.95	1	27
incompletely understood	6471	71.11	67	329
indefinite relationship	3908	42.95	36	165
largely understood	1695	18.63	15	108
anomalous/curious	1012	11.12	9	39
alternative/controversy	1519	16.69	12	94
difficult task	320	3.52	2	25
problem/complication	532	5.85	3	29
future work	948	10.42	6	103
future prediction	366	4.02	2	52
important consideration	1786	19.63	9	199
ALL CATEGORIES	19681	216.27	200	1021
SCOPE	6700	73.63	69	262

specific category as the minimum for each individual category was zero. The ignorance categories followed a similar pattern as above, where the most represented categories continued to be *incompletely understood* and *indefinite relationship* (above or equal to 130) and the least represented were *future prediction* and *explicit inquiry* (under 20). We also calculated the percentage of unique annotations for each category, meaning the number of total unique annotations per category divided by the number of total annotations per category. For all categories these percentages were under 22%, and most were under 10%. Of note, 52% (1301 out of 2513) of the lexical cues did not appear at all in the corpus, meaning there were no sentences with those cues. Thus, a select subset of lexical cues and categories were frequently reused across all annotations.

Further, the distribution of lexical annotations varied by article section, showing that most appeared in the discussion and conclusion sections as expected and surprisingly many also appeared in the methods section (see Table 3.6). There were no ignorance annotations within the

Table 3.5: Unique Annotation Statistics Per Ignorance Category: Total number of unique lexical cue annotations in all articles and statistics per ignorance category. We did not include SCOPE because the number of sentences was the same; we only capture the scope one time no matter how many lexical cue annotations occur within it. Note that all categories except for ALL CATEGORIES (had 1) had zero minimum number of unique annotations. Note max = maximum.

Category	# total unique annotations	% unique annotations	avg # unique annotations per article	median # unique annotations per article	max # unique annotations
answered question	49	8%	4.32	5	14
unknown/novel	69	21%	2.53	2	9
explicit inquiry	16	9%	1.21	1	8
incompletely understood	368	6%	32.66	35	83
indefinite relationship	130	3%	15	15	45
largely understood	106	6%	7.59	7	25
anomalous/curious	63	6%	4.6	5	14
alternative/controversy	100	7%	7.57	7	24
difficult task	40	13%	2.3	2	11
problem/complication	41	8%	2.91	2	12
future work	105	11%	5.34	4	21
future prediction	14	4%	1.84	1	6
important consideration	101	6%	6.73	5	37
ALL CATEGORIES	1202	6%	94.59	104	273

titles of papers. As might be intuitive, the conclusion section contained the most annotations (almost double the next largest section), followed by the discussion and then the method section. The results and abstract sections contained the fewest, with the abstract containing very few in general and on average potentially due to the normally small size of this section. On average, each section had at least 2 annotations and at most 200 annotations. The medians were slightly lower, indicating some outlier articles with many annotations. In fact, the maximum number of annotations in one section was 1,000 in the conclusion section. Further, the conclusion, discussion, and methods sections in all articles contained at least one if not a few more ignorance annotations at the bare minimum. For example, "Owing to the lack of other available data, we used the annual number of live births in Iaşi county reported on 01 July 2009 (n=9,499) to define the size of the reference group [23]" was in a methods section with owing to mapping to *problem/complication*, lack of...data mapping to *unknown/novel*, and reported mapping to *incompletely understood*. The sentence presented a method while also explaining why that

particular one was chosen due to a lack of data. A researcher could choose to focus on all the ignorance statements from the methods section for example to get ideas for gaps in methodologies. Statements of ignorance were rampant throughout the scientific literature in all sections of an article.

Table 3.6: Annotation Counts per Section: Total number of annotations by section in all articles with section delineation and statistics per article. Note that every article contained a title and none of the titles had any ignorance annotations. Note avg = average, min = minimum, and max = maximum.

Section	# total articles	# total annotations	avg # annotations per article	median # annotations per article	min # annotations	max # annotations
abstract	71	123	1.73	1	0	29
introduction	86	2649	30.8	16	0	571
methods	61	2880	47.21	38	1	367
results	54	927	17.17	6.5	0	287
discussion	57	4906	86.07	55	3	330
conclusion	41	8196	199.9	150	2	990

Statements of ignorance and lexical cues can be automatically identified

We can automatically identify statements of ignorance and lexical cues achieving F1 scores around or above 0.8 with many closer to 0.9 (see Table 3.7 for sentence classification and Table 3.11 for word classification). Automatic identification of both statements of ignorance and the specific lexical cues provides researchers with specific knowledge goal statements to explore for potential future research ideas. This automation provides a foundation for future applications to explore the state of scientific ignorance at scale to help accelerate research.

Evaluating on a significant amount of unseen data provided a robust estimate of our classifier performance and hinted at generalizability. The ignorance corpus of 12,055 sentences was split into 8,281 sentences in training (2/3) and 3,774 for evaluating (1/3) which included over 285,000 words with 2053 lexical cues and over 131,000 words with 1382 lexical cues, respectively (see Table 3.1). Further, the evaluation data included 501 unique unseen lexical cues in the training, meaning there were no sentence examples for these cues in the training data. We provide the results for all classifiers recognizing that one classifier did not work for all ignorance categories. (See Figures 3.5 and 3.6 for a comparison of all the models on the test data for

sentence and word classification respectively. For the results for each algorithm including training and testing scores, see Tables 3.8- 3.10 for sentence classification, and Tables 3.12 and 3.13 for word classification.)

The ensemble of 13 different binary classifiers performed the best for both classification tasks and all models were necessary. On the sentence-level, the binary classifier performed quite well at an F1 score of 0.95 on the testing data, while the one multi-classifier failed at 0.12. For the ensemble, all ignorance categories performed near 0.8-0.9 with BERT and BioBERT as the main algorithms and ANN only for *incompletely understood*. The most difficult category appeared to be *alternative/controversy* at an F1 score of 0.79. The ANN performed quite well on training, but scores dropped in testing (see Table 3.8). BERT similarly performed quite well on training while maintaining most of the performance on testing aside from a few ignorance categories (see Table 3.9). Lastly, BioBERT was similar to BERT although the testing scores remained higher than BERT overall (see Table 3.10). Overall, we can automatically identify sentences as statements of ignorance quite well, providing knowledge goal statements to explore in future work.

Table 3.7: Sentence Classification: the best model for sentence classification for each ignorance category and all categories combined.

Ignorance Category	Model	testing F1 score	testing support
ALL CATEGORIES BINARY	BioBERT	0.95	2005
answered question	BERT	0.97	168
explicit inquiry	BioBERT	0.9	92
unknown/novel	BioBERT	0.88	63
incompletely understood	ANN	0.83	225
indefinite relationship	BERT	0.87	1072
largely understood	BERT	0.9	312
anomalous/curious	BERT	0.96	149
alternative/controversy	BioBERT	0.79	441
difficult task	BERT	0.95	93
problem/complication	BioBERT	0.9	202
future work	BioBERT	0.85	195
future prediction	BERT	0.88	55
important consideration	BERT/BioBERT	>0.99	491
ALL CATEGORIES COMBINED	BioBERT	0.12	2005

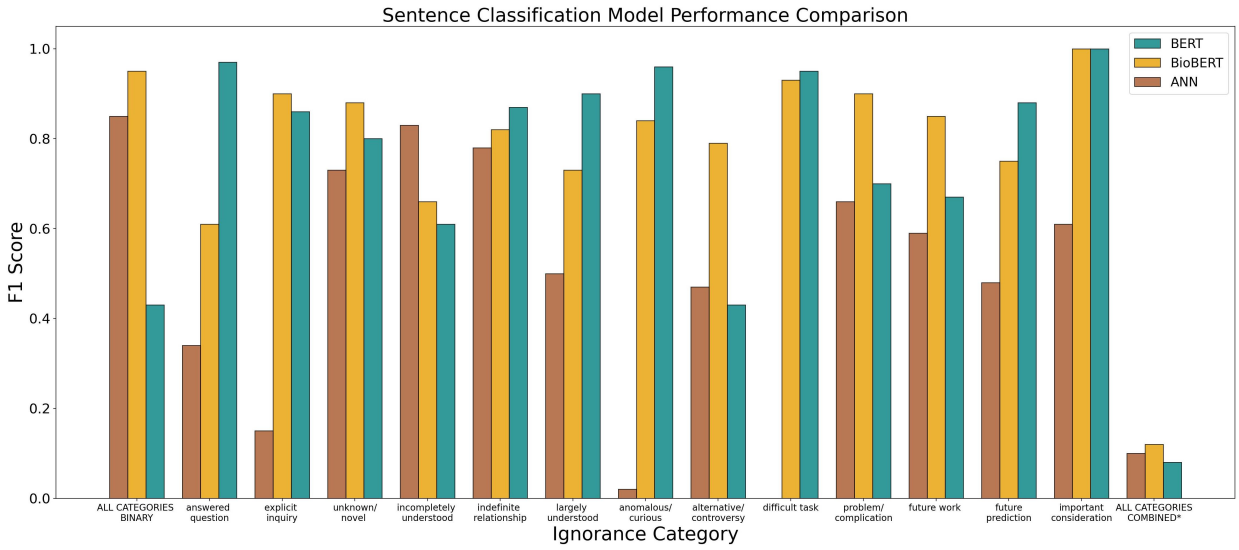


Figure 3.5: Sentence classification summary: A bar plot summary of test F1 scores for sentence classification. *Reporting the macro-average F1 score of all the categories for one multi-classifier.

Table 3.8: ANN sentence classification: Note that one sentence can map to more than one category and so they will not add up to the total binary. *Reporting the macro-average F1 score of all the categories for one multi-classifier.

Ignorance Category	training F1 score	training support	testing F1 score	testing support
ALL CATEGORIES BINARY	0.99	3354	0.85	2005
answered question	0.66	479	0.34	168
explicit inquiry	0.95	524	0.15	92
unknown/novel	0.87	347	0.73	63
incompletely understood	0.9	1582	0.83	225
indefinite relationship	0.86	1965	0.78	1072
largely understood	0.86	523	0.5	312
anomalous/curious	0.92	479	0.02	149
alternative/controversy	0.97	942	0.47	441
difficult task	0.88	342	0.0	93
problem/complication	0.68	324	0.66	202
future work	0.85	869	0.59	195
future prediction	0.93	178	0.48	55
important consideration	0.75	735	0.61	491
ALL CATEGORIES COMBINED*	0.17	3354	0.08	2005

Table 3.9: BERT sentence classification: Note that one sentence can map to more than one category and so they will not add up to the total binary. *Reporting the macro-average F1 score of all the categories for one multi-classifier.

Ignorance Category	training F1 score	training support	testing F1 score	testing support
ALL CATEGORIES BINARY	0.83	3354	0.43	2005
answered question	0.99	479	0.97	168
explicit inquiry	0.87	524	0.86	92
unknown/novel	0.89	347	0.8	63
incompletely understood	0.81	1582	0.61	225
indefinite relationship	0.91	1965	0.87	1072
largely understood	0.96	523	0.9	312
anomalous/curious	0.93	479	0.96	149
alternative/controversy	0.83	942	0.43	441
difficult task	0.92	342	0.95	93
problem/complication	0.92	324	0.7	202
future work	0.83	869	0.67	195
future prediction	0.9	178	0.88	55
important consideration	>0.99	735	>0.99	491
ALL CATEGORIES COMBINED*	0.23	3354	0.10	2005

Table 3.10: BioBERT sentence classification: Note that one sentence can map to more than one category and so they will not add up to the total binary. *Reporting the macro-average F1 score of all the categories for one multi-classifier.

Ignorance Category	training F1 score	training support	testing F1 score	testing support
ALL CATEGORIES BINARY	0.99	3354	0.95	2005
answered question	0.93	479	0.61	168
explicit inquiry	0.99	524	0.9	92
unknown/novel	0.88	347	0.88	63
incompletely understood	0.97	1582	0.66	225
indefinite relationship	0.88	1965	0.82	1072
largely understood	0.86	523	0.73	312
anomalous/curious	0.85	479	0.84	149
alternative/controversy	0.92	942	0.79	441
difficult task	0.91	342	0.93	93
problem/complication	0.82	324	0.9	202
future work	0.83	869	0.85	195
future prediction	0.88	178	0.75	55
important consideration	>0.99	735	>0.99	491
ALL CATEGORIES COMBINED*	0.22	3354	0.12	2005

The lexical cues provided more details as to where the specific knowledge goal was signified. On the word level, it seems to be even easier to automatically classify lexical cues, even ones that were not seen in the training data (see Table 3.11). Both the binary classifier and one true multi-classifier performed quite well using BioBERT mainly, but the ensemble still outperformed both with F1 scores near or above 0.9. Overall, BioBERT performed better than the CRF, but the CRF was generally quite close in performance aside from *future prediction*. Both the CRF and BioBERT maintained their performance from training to testing generally (see Tables 3.12 and 3.13, respectively). The CRF can be used in low resource settings if GPUs are not available for BioBERT prediction, otherwise BioBERT performed best (similar our results in Chapter 2). Further, the best models achieved an F1 score of 0.84 on the 501 unique unseen lexical cues only in the evaluation data with no examples in the training data. These models were quite effective on unseen cues. The few errors seemed to be due to minor variations on the ignorance cues already in the taxonomy, such as plurals or words in different orders. Thus, both our sentence and word classification models performed quite well and can be used in future applications to explore these statements.

Table 3.11: Word Classification: the best model for word classification for each ignorance category and all categories combined. *Reporting the average F1 score of all the categories for one multi-classifier.

Ignorance Category	Model	testing F1 score	testing support
ALL CATEGORIES BINARY	BioBERT	0.89	7601
answered question	BioBERT	0.89	320
unknown/novel	CRF	0.98	155
explicit inquiry	BioBERT	0.97	43
incompletely understood	BioBERT	0.93	2809
indefinite relationship	BioBERT	0.97	1205
largely understood	BioBERT	0.94	618
anomalous/curious	BioBERT	0.96	399
alternative/controversy	BioBERT	0.91	598
difficult	CRF	0.93	128
problem/complication	BioBERT	0.9	238
future work	BioBERT	0.89	391
future prediction	BioBERT	0.94	100
important consideration	BioBERT	0.93	608
ALL CATEGORIES COMBINED*	BioBERT	0.82	6239

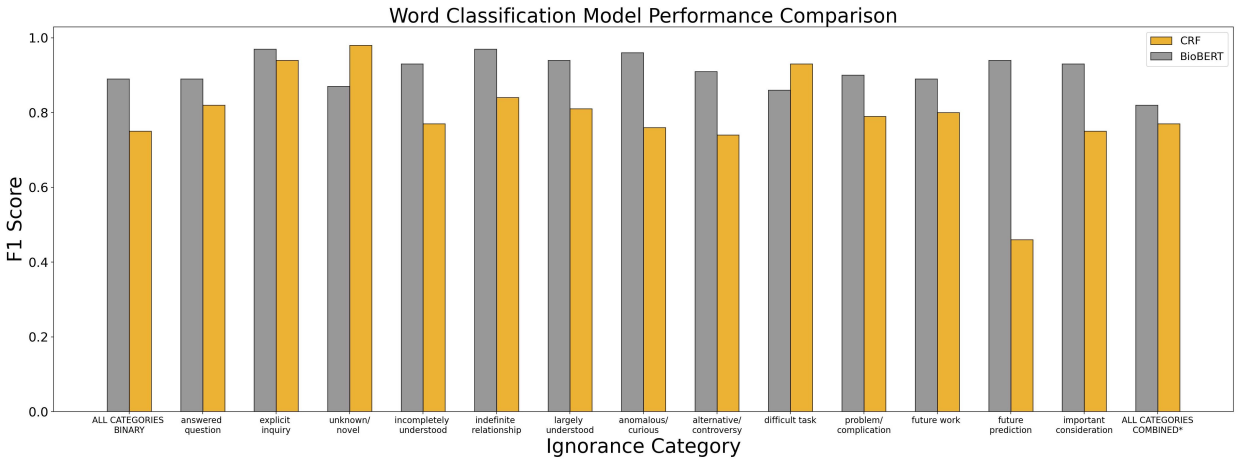


Figure 3.6: Word Classification summary: A bar plot summary of test F1 scores for word classification. *Reporting the macro-average F1 score of all the categories for one multi-classifier.

Table 3.12: CRF word classification. *Reporting the macro-average F1 score of all the categories for one multi-classifier.

Ignorance Category	training F1 score	training support	testing F1 score	testing support
ALL CATEGORIES BINARY	0.76	16748	0.75	7601
answered question	0.8	692	0.82	320
unknown/novel	0.97	401	0.98	155
explicit inquiry	0.94	136	0.94	43
incompletely understood	0.76	5764	0.77	2809
indefinite relationship	0.83	2847	0.84	1205
largely understood	0.79	1313	0.81	618
anomalous/curious	0.73	809	0.76	399
alternative/controversy	0.73	1262	0.74	598
difficult task	0.9	283	0.93	128
problem/complication	0.8	432	0.79	238
future work	0.82	991	0.8	391
future prediction	0.56	341	0.46	100
important consideration	0.75	1484	0.75	608
ALL CATEGORIES COMBINED*	0.77	13828	0.77	6239

Discussion

Capturing and classifying statements of ignorance in terms of their entailed knowledge goal was an important new NLP task that formalized and extended prior work in hedging, uncertainty, speculation, factuality, epistemics, and meta-knowledge into how such statements can serve future knowledge gain. Even though each phenomenon was defined differently, our

Table 3.13: BioBERT word classification. *Reporting the macro-average F1 score of all the categories for one multi-classifier.

Ignorance Category	training F1 score	training support	testing F1 score	testing support
ALL CATEGORIES BINARY	0.9	16748	0.89	7601
answered question	0.92	692	0.89	320
unknown/novel	0.91	401	0.87	155
explicit inquiry	0.95	136	0.97	43
incompletely understood	0.94	5764	0.93	2809
indefinite relationship	0.98	2847	0.97	1205
largely understood	0.94	1313	0.94	618
anomalous/curious	0.91	809	0.96	399
alternative/controversy	0.9	1262	0.91	598
difficult task	0.9	283	0.86	128
problem/complication	0.88	432	0.9	238
future work	0.91	991	0.89	391
future prediction	0.97	341	0.94	100
important consideration	0.96	1484	0.93	608
ALL CATEGORIES COMBINED*	0.82	13828	0.82	6239

ignorance taxonomy relates, subsumes, and nuances much of it. The prior hedging work [37, 38, 41] mainly focused on facts and as stated earlier, a hedged statement was not necessarily a statement of ignorance. Even still, many statements overlapped and contained the same lexical cues including suggests, likely, may, possible, probably, and either...or [41]. For uncertainty, our work related to epistemic uncertainty introduced by Walker [73], where the goal was to think critically about the source of uncertainty and how that impacted decisions. Our work aimed to think critically about the source of unknowns and how that drives research. Uncertainty, speculation, and factuality research [27, 43] also turned to a focus on confidence levels to understand the certainty of the statements. We can map this onto the stages of research under our broad category of “epistemic indication of research topic, assertion, or variables" where research begins with an *unknown/novel* and completes with a *largely understood* idea into knowledge (see Figure 3.3). This continuum included the categories *largely understood* and *incompletely understood*, which mapped to the factuality levels of probable and possible in SemRep [43]. Lastly, the meta-knowledge research [260] aimed to combine many of these ideas as manner,

source, polarity, certainty level, and knowledge type. Certainty level L1 (possible) mapped roughly to *incompletely understood* and level L2 (probable) to *largely understood*. Further, our taxonomy was a more nuanced categorization of their knowledge types of research hypothesis and new knowledge. Also it was more nuanced than recent work focusing on scientific challenges and directions [17]. Future work could include augmenting both the meta-knowledge and recent work to add in our ignorance taxonomy.

Our ignorance taxonomy not only aimed to extend this prior work, but also to be grounded in the field of philosophy of science as a theory of science progression and evolution (see Figure 3.3). Our ignorance taxonomy subsumed a few types of ignorance informally laid out by Firestein [1], including curiosity, possibility, and controversy, which extended to *anomalous/curious*, *incompletely understood*, and *alternative/controversy* in our taxonomy respectively. Kuhn [2] discussed how small discrepancies in the predictions of a theory can accumulate until they cause a crisis necessitating a completely new theory. In our taxonomy, this corresponded to the category of *anomalous/curious*. Pearl *et al.* [70] aimed to forge connections from correlation to causation, the *indefinite relationship* category in our taxonomy, since the literature was filled with associations. Han *et al.* [69] focused on healthcare with categories including probability, ambiguity, and complexity extending to *largely understood*, *anomalous/curious*, and *difficult task* in our taxonomy, respectively. Smithson's [72] taxonomy included incompleteness and probability, which mapped to *incompletely understood* and *largely understood* in our taxonomy, respectively. The philosophical implications of our work can continue to be explored in future work.

All of our work provided a foundation for future tools and methods to explore these statements of ignorance to develop the scientific ignorance landscape at scale and across disciplines, hopefully helping to accelerate research. Using our ignorance taxonomy and lexical cues, we provided robust annotation guidelines that yielded a high quality corpus of 91 prenatal nutrition articles. Human readers reliably identified and categorized statements of ignorance in our annotation task with all IAAs around 80% (see Table 3.3). Much of the prior work also

provided lexical cues to help with their tasks (*e.g., [41, 51, 61, 263]*). We extended these lists to over 2,500 cues, which to the best of our knowledge is the largest lexical cue list. Between our two annotation tasks alone, over 600 lexical cues were added to the taxonomy each time. This also showed that statements of ignorance employed a rich vocabulary. Future work can focus on these cues to better understand how researchers discuss knowledge goals in the literature.

At the same time, lexical cues may not be the only way to capture statements of ignorance and we may have biased the annotators by highlighting the lexical cues before annotation. Sometimes the annotators agreed that a sentence was a statement of ignorance, but either disagreed on the lexical cue or had a hard time deciding on the cue. In these cases, we needed to identify a cue based on our task definition, but there could be other features that capture it better. Lahav *et al.* [17] aimed to not only capture sentences with lexical cues but also augment beyond it. However, they did not discuss what other features helped identify such statements. Future work can explore other linguistic features that may help capture more statements of ignorance. We also did not measure the effect of pre-highlighting lexical cues for the annotators. They found the task infeasible when the lexical cues were not highlighted. Even with the highlights, the annotators continued to find new lexical cues. Future work can explore this effect and determine other means to capture statements of ignorance. We provide our very detailed ignorance taxonomy, extensive annotation guidelines, and our gold standard corpus as in prior work (*e.g.*, [17, 39, 41]), for future work to expand upon.

Based on our corpus, the scientific literature was rich in statements of ignorance both by article and by section (see Tables 3.4-3.6), which are ripe for exploration in future work. This led to the broader question: does ignorance exist in every article? Our corpus suggests that there was an almost even split between ignorance statements mapping to all thirteen categories and not. Every article had at least one ignorance annotation no matter how short the article and just over half of the total sentences in all articles (55%) were statements of ignorance. This was in direct contrast to Bongelli *et al.* in that they claimed that “the percentages of uncertainty are always much lower than that of certainty" [262]. Our corpus begs to differ with respect to ignorance most

likely because our taxonomy was much broader than uncertainty including categories such as *difficult task*, *important consideration*, and *answered question*, which would not be deemed uncertain. More work is necessary to truly make these claims. Further, we also disagree with their claim that “uncertainty progressively diminishes over time in biomedical articles" [262]. Ignorance did not seem to diminish over time in our corpus: all articles spanning from 1979 to 2018 contained ample ignorance. We do however agree with their [262] finding that most ignorance statements were in the discussion and conclusion sections. Surprisingly, the methods section had more statements of ignorance than expected and it seems due to the justifications of techniques with comparisons and explanations in relation to previous work. The introduction section had the next largest, which was the second largest category for Bongelli *et al.* [262]. Future work is needed to explore statements of ignorance by article and across sections. Even so, with our data, it seems that the prenatal nutrition literature at least was filled with many statements of ignorance ripe for exploration.

All of these statements of ignorance mapped to all different types of ignorance spanning all the categories, providing different types of knowledge goals for researchers to explore. Most of them were under the categories of *incompletely understood* and *indefinite relationship* with many articles providing more evidence but not enough yet for claims, and many associations and relationships between entities in the articles, respectively. Further, the most underrepresented category was *explicit inquiry*, showing that most of the statements of ignorance were not explicit, but instead implicit in the 12 other categories, making this task very difficult and nuanced. Our corpus covered a wide range of ignorance categories ripe for exploration.

With all this gold standard data, we showed that we can automate this task to identify statements of ignorance and lexical cues for use in future applications (see Tables 3.7- 3.13 and Figures 3.5 and 3.6). The ensemble of 13 binary classifiers for both the sentence and word classification tasks performed quite well with F1 scores on a held out test set around or above 0.8, with many close to 0.9. The ensemble avoided the difficulties of a multi-classification problem. Also, our word classifiers achieved an F1 score of 0.84 on lexical cues unseen in the training data

(501 cues), hinting at generalizability beyond the cues gathered here. At the same time, we used only the baseline hyperparameters for our classifiers and thus other parameter tunings and other algorithms, such as PubMedBERT [243], may yield improved results. Even still, we demonstrated that it was feasible to automate the identification of statements of ignorance.

The major limitations of this work were the amount of data and the focus on only one field. The goal of this work was to show feasibility of these ideas and methods, and thus it is unclear whether this work generalizes beyond both the data and the field of prenatal nutrition. More data generally boosts performance and helps determine generalizability. Even still, we achieved high F1 scores on our corpus of 91 articles, which seemed to be enough to train, evaluate, and test the classification models. Future work could include more annotations in the prenatal nutrition field to help improve automation. Further, to test generalizability beyond this one field, future work could apply our ignorance classifiers to another field, such as the COVID-19 literature, LitCovid [275], and manually evaluate the annotations. The results from that analysis can be compared to this work to determine generalizability. It remains to be seen if these results generalize beyond prenatal nutrition.

Even with these limitations, this work showed that it was possible to identify ignorance statements in a field that is difficult to study due to ethical and legal considerations [110–113]. This work provided the first step towards finding new research areas to explore and facilitating new interdisciplinary interactions that could advance the study of this underserved population. Discussions of these statements of ignorance can both refine and improve research questions, identify established facts, and facilitate the comparison of approaches for further research. These discussions of scientific questions are of interest to researchers, educators, publishers, and funders because they provide insights and directions for new research and may provide context for existing results. Formalizing and disseminating such statements of ignorance in the scientific literature is an important new NLP task with the potential to greatly impact how we view the literature and scientific progress in general.

Conclusion

The new NLP task of finding scientific questions or statements of ignorance in the scientific literature will not only yield novel text mining tools, but will also help accelerate research through tracing out the evolution of scientific thought in a discipline, pointing out gaps or flaws in existing theories, and providing new avenues for future insights. Here we not only showed that this task was feasible, but created an ignorance taxonomy, a gold standard corpus, and classification models. The ultimate goal is to help enhance literature awareness by creating an ignorance-base (compared to a knowledge-base) of statements of ignorance found in the literature for all to explore, including students, researchers, publishers, and funders.

Summary

This work introduced a novel biomedical natural language processing task that aimed to identify, characterize, and automatically classify statements of ignorance from the scientific literature (statements where scientific knowledge is missing or incomplete). Our interest was on how these statements can play a role in accelerating research (goals for scientific knowledge). We presented a novel ignorance taxonomy driven by these goals for scientific knowledge along with annotation guidelines explaining how to identify such statements. Using this taxonomy and reliable annotation guidelines (inter-annotator agreement above 80%), we created a gold standard ignorance corpus of 91 full-text documents from the prenatal nutrition literature with over 20,000 annotations and used it to train classifiers that achieved over 0.80 F1 scores on a held out test set. By focusing on statements of ignorance there is a potential to develop the scientific ignorance landscape at scale and across disciplines, resulting in an accelerated research process.

Contributions And Acknowledgements

This work was the combination of one publication [173] and one preprint [175].

- **Boguslav, M. R.**, Salem, N. M., White, E. K., Leach, S. M., Hunter, L. E. Identifying and classifying goals for scientific knowledge. Bioinformatics advances. 2021;1:1 vbab012. Available From: https://doi.org/10.1093/bioadv/vbab012

- **Boguslav MR**, Salem NM, White EK, Sullivan KJ, Araki SP, Bada M, *et al.* Creating an Ignorance-Base: Exploring Known Unknowns in the Scientific Literature. 2022;:2022.12.08.519634. Available From: https://doi.org/10.1101/2022.12.08.519634

I would like to thank the other authors on those papers including Nourah M. Salem, Elizabeth K. White, Katherine J. Sullivan, Stephanie P. Araki, Michael Bada, Teri L. Hernandez, Sonia M. Leach, and Lawrence E. Hunter.

For the publication, MRB, SML and LEH conceived of the idea. MRB and EKW annotated all documents for the corpus. MRB and NMS created the classification models. All authors helped write and reviewed the manuscript. I would like to thank Harrison Pielke-Lombardo for his updates to the Knowtator tool; William A. Baumgartner Jr. for his help with the literature and taxonomy work; Teri L. Hernandez for her consultations on prenatal nutrition; Michael M. Bada for his guidance on annotation tasks and the taxonomy; and Katherine J. Sullivan and Tiffany J. Callahan for many discussions about this work.

For the preprint, MRB, SML, and LEH conceived of the ideas. KJS, SPA, EKW, MRB, and MB were a part of the annotation task. MRB and NMS created the classification models. TLH was our prenatal nutrition expert. TLH, SML, and LEH supervised the work. All authors helped write and reviewed the manuscript. I would like to acknowledge the BioFrontiers Computing Core for computing resources and support, especially Jonathon Demasi. I would like to thank Harrison Pielke-Lombardo for his updates to the Knowtator tool; William A. Baumgartner Jr. for his help with the literature and taxonomy work; and Tiffany J. Callahan for many discussions about this work.

CHAPTER IV

CREATING AN IGNORANCE-BASE: EXPLORING KNOWN UNKNOWNS IN THE SCIENTIFIC LITERATURE

Background And Related Work

The ultimate goal of this work is to provide a system for students, researchers, publishers, and funders to explore the scientific ignorance landscape at scale and across disciplines, resulting in an accelerated research process. With both the biomedical concepts (Chapter 2) and ignorance statements (Chapter 3) recognized, it is now possible to create an **ignorance-base**, a knowledge-base focused on the questions or **ignorance statements**. Research begins by accumulating knowledge on an unexplored subject (an unknown unknown). It continues through exploring the questions (known unknowns) until a body of established facts emerges (known knowns) [1–3]. Many **knowledge-bases** exist to capture the known knowns from domain experts, the scientific literature, and other data sources such as experimental results [4]. These knowledge-bases are important based on their variety of applications [4], including finding and interpreting information based on a single input topic, such as a concept, or a set of input topics that may be related, such as those from experimental results. For example, a graduate student or researcher interested in learning about the field of prenatal nutrition might consult a database of dietary supplements [5]. Or a researcher might perform a functional enrichment analysis to characterize a list of genes associated with vitamin D and preterm birth by finding relevant known biomedical concepts [6].

Both of these examples explore the knowledge-bases of known facts, but equally important are the questions surrounding a topic or a set of experimental results. How do researchers find the most pertinent questions? Researchers learn these skills in graduate school where the goal is to identify and provide at least some solutions for a question that is unanswered (a thesis project). The graduate student needs to learn both the questions in the field and how they are asked in order to ask their own. The graduate student is a good example of how researchers find questions and many books [7–11] and articles [12–16] discuss this process, including how to choose a topic or

question. However, to the best of our knowledge there are no automated systems that provide insights, summaries, and visualizations to help researchers find the most pertinent questions based on knowledge goals and the biomedical ontologies to support finding areas of research (biomedical concepts) with lots of questions (ignorance statements) that are ripe for future research. An automated system could be useful to a wide variety of scientists ranging from graduate students looking for thesis projects (*e.g.*, [12]) to funding agencies tracking emerging research areas (*e.g.*, [33]). It could help facilitate interdisciplinary interactions amongst researchers by finding questions from another field that bear on a topic or a set of experimental results (*e.g.*, [34]). It could also help track the evolution of research questions over time as a longitudinal analysis (*e.g.*, [35]). Further, automatically identifying questions would allow us to query existing databases for information (*e.g.*, [36]). Thus, there is a need for such an automated system to capture questions or known unknowns.

Just as knowledge-bases rely on the representation of known knowns [4], such an automated system would rely on representations of known unknowns. Prior efforts exist to capture them through understanding the phenomenon [20, 21, 23–25, 27, 37–67], creating taxonomies where a hierarchy of terms was linked by specified relationships [43, 46, 68–73] and ontologies specifying relationships among controlled vocabularies [74–77], annotating literature to create corpora [40, 41, 50, 55, 61, 63, 78–81, 84, 85, 264], and automating the identification of known unknowns through classification tasks [20, 42–45, 47–49, 51–54, 56, 58–62, 64–67, 86–94]. Some efforts have also sought to capture known unknowns completely by creating theoretical frameworks, determining if the task was feasible for humans to perform, and automating it [17, 19–27, 46, 95, 96]. (See Chapter 3 for more details.) None of these works have created a knowledge-base of known unknowns that can provide summaries and visualizations based on an input topic or a set of experimental results.

Only one work [17] was close in their creation of a search engine, similar to a PubMed search, for new directions and challenges in the COVID-19 Literature. They aim to help researchers discover scientific challenges and directions (their two known unknown categories) by

providing relevant articles and sentences for each input biomedical MeSH (Medical Subject Headings - a controlled vocabulary that is part of the Unified Medical Language System from NLM used for indexing, cataloguing, and searching for biomedical information and documents [18]) term. However, their two categories of known unknowns were quite broad, when most other works included more nuanced categorizations (*e.g.*, Chapter 3 with 13 categories focused on scientific knowledge goals and [19–27]). Their input concepts were not grounded in ontologies, limiting the ability to connect their work to other ontology-based efforts (*i.e.*, many knowledge-bases [4]). Also, they cannot support queries by experimental results as standard methods for contextualizing experimental results (functional enrichment analysis) use knowledge-bases and ontologies [28–32]. Further, they did not go beyond the identification of sentences to provide summaries or visualizations of the immense amount of data outputted. Tables alone do not suffice to truly explore all the outputted data (thousands of sentences can be found per input concept). They “hope to build more tools to explore and visualize challenges and directions across science" in future work [17]. Thus, no prior work has created a knowledge-base to capture the nuance of the known unknowns drawing from prior work, connected them to ontologies for integration with other knowledge-bases, and provided summaries and visualizations of the outputs to help researchers find the most pertinent questions.

Lahav *et al.* [17] cannot support queries by experimental results. For contextualizing experimental results, methods for standard functional enrichment analyses use knowledge-bases and ontologies [28–32], and natural language processing (NLP) tools over the biomedical literature [34, 97–106]. Some of this prior work not only aimed to characterize genes but also to help define new research areas (*e.g.*, [34] as one of a few goals), generate new hypotheses (*e.g.*, [107]), and find information about genes of unknown function and fill gaps in knowledge (*e.g.*, a preprint [105] employs manual curation). Thinking beyond a gene list, if we consider pathway models as experimental results, tools exist to associate pathway models to the literature (*e.g.*, [108]) and some of these took uncertainty into account (*e.g.*, [44, 109]). However, these works focused on confidence and relevance, respectively, rather than explicitly representing statements

of known unknowns. Thus, to the best of our knowledge, there is no prior work that can facilitate the widespread search for questions to find new avenues for exploration from a set of experimental results.

Thus, our goal was to create and use a knowledge-base containing representations of known unknowns, formally defined as an **ignorance-base** (based on [1] and Chapter 3), to **explore by topic** and **explore by experimental results**, as done in the knowledge-bases. Identifying questions that need answers, would allow us to then look to other fields and knowledge-bases for answers. We aim to help students, researchers, funders, and publishers better understand the state of our collective scientific **ignorance** (known unknowns) in order to help accelerate translational research through the continued illumination of and focus on the known unknowns and their respective goals for scientific knowledge.

While these ideas and methods are generally applicable across biomedical research, we focused on the prenatal nutrition field. Due to ethical and legal considerations and complexities in studying pregnant mothers and fetuses, prenatal nutrition is understudied and serves to benefit from the identification of questions that are well studied in other fields [110–113]. Fetal development is a critical period and exposure to nutrition has a lifelong impact [114]. For example, the micronutrient vitamin D is very important for maternal and fetal health, affecting the immune and musculoskeletal systems, neurodevelopment, and hormones [115–119] (see Figure 4.1). Abnormal vitamin D levels are associated with gestational diabetes mellitus, preterm delivery, frequent miscarriages, adipogenesis, pre-eclampsia, obstructed labor, Cesarean sections, reduced weight at birth, respiratory issues, postpartum depression, and autism [115]. If we can identify the known unknowns in prenatal nutrition, even just with regard to the role of vitamin D, we can look to other fields to help us answer the questions raised, thereby avoiding the additional risks faced when studying pregnant women and their children. Also the prenatal nutrition field is a good case study for these ideas because it contains a diverse literature with all types of studies from all over the world, meaning there is a higher potential for generalizability to other fields. Thus, this work has the potential to generalize beyond prenatal nutrition, and more specifically to

facilitate new interdisciplinary interactions that could advance the study of an underserved population and potentially help accelerate translational research for mothers everywhere.

We present the ignorance-base as a formal representation of **statements of ignorance** using the example of the prenatal nutrition literature. We define statements of ignorance as statements of incomplete or missing knowledge categorized based on the entailed **knowledge goals** (*i.e.*, the next actionable step based on the given unknown), as defined in Chapter 3. We chose this formalization of known unknowns because of the knowledge goal focus in order to provide actionable next steps to the students, researchers, funders, or publishers. For example, "these inconsistent observations point to the complicated role of VITAMIN D in the IMMUNE modulation and disease process" (PMC4889866) was a statement of ignorance. The ignorance statement with its entailed knowledge goal was identified based on the underlined words that communicate knowledge is missing, **lexical cues**, which mapped to an **ignorance taxonomy**, a formal categorization of knowledge goals. The cue inconsistent was an *indication of alternative research options or controversy of research*, the ignorance category, and complicated was an *indication of difficult research task*. Thus, the entailed knowledge goal was to determine the correct role of vitamin D in the immune modulation and disease process by creating novel methods or conducting new experiments to study the complicated role. This knowledge goal is an action that researchers can take in future research. As Chapter 3 showed, most ignorance statements were implicit, as the example above. In order to also understand the biomedical subjects of these statements and be able to connect them to other knowledge-bases, we identified the biomedical concepts from the open biomedical ontologies (OBOs) (*e.g.*, the all caps words in the example sentence) using the concept recognition methods from Chapter 2. The ignorance-base combines the identification and classification of ignorance statements and biomedical concepts to hopefully provide researchers with knowledge goals to pursue in future research.

With the ignorance-base, we show that it is possible to automatically find statements of ignorance related to a topic motivated by a researcher's search for the pertinent questions in vitamin D. The goal is to find areas of research (biomedical concepts) with lots of questions

(ignorance statements) that are ripe for future research. Analogous to exploring by topic using knowledge-bases, we used the ignorance-base to find other related keywords or concepts and specific knowledge goals for the researcher. We determined if other concepts were enriched in vitamin D ignorance statements compared to all statements (**ignorance enrichment**). We found the concepts IMMUNE SYSTEM, BRAIN DEVELOPMENT and RESPIRATORY SYSTEM to be enriched in vitamin D ignorance statements compared to the standard approach of concept enrichment in all vitamin D documents (see Figure 4.1). All of these biomedical concepts with their corresponding ignorance statements provided lots of interesting questions and ideas. In a similar vein, to help the researcher understand the general landscape of questions surrounding a topic and narrow in on types of question to ask (ignorance category), we present a summary of the ignorance categories and mapped out how they changed over time. We identified the ignorance categories that were over-represented in a subset of ignorance statements as compared to all ignorance statements (**ignorance-category enrichment**). By narrowing the researcher's search to specific ignorance categories, we argue that this provided the researcher with a filtered set of knowledge goals to explore as potential questions for future work. For example, the researcher could choose a topic that was a complete unknown (*indication of unknown or novel research topic or assertion*) or a topic where there were alternate existing hypotheses to explore (*indication of alternative research options or controversy of research*). Further, we argue how this can help track emerging research areas or perform a longitudinal analysis of the evolution of research questions over time for funding agencies and publishers [33, 35, 132–139].

We also show that it is possible to contextualize experimental results in terms of statements of ignorance to understand what questions may bear on them, potentially from another field. Similar to exploration by topic, the goal was to find areas of research (biomedical concepts) that ideally imply another field, with lots of questions (ignorance statements) that are ripe for future research. We used a gene list connecting vitamin D and spontaneous preterm birth (sPTB) from the literature [6] as our example. If vitamin D plays a role in preventing sPTB, it would affect mothers everywhere. In comparing our ignorance-base approach to the standard methods

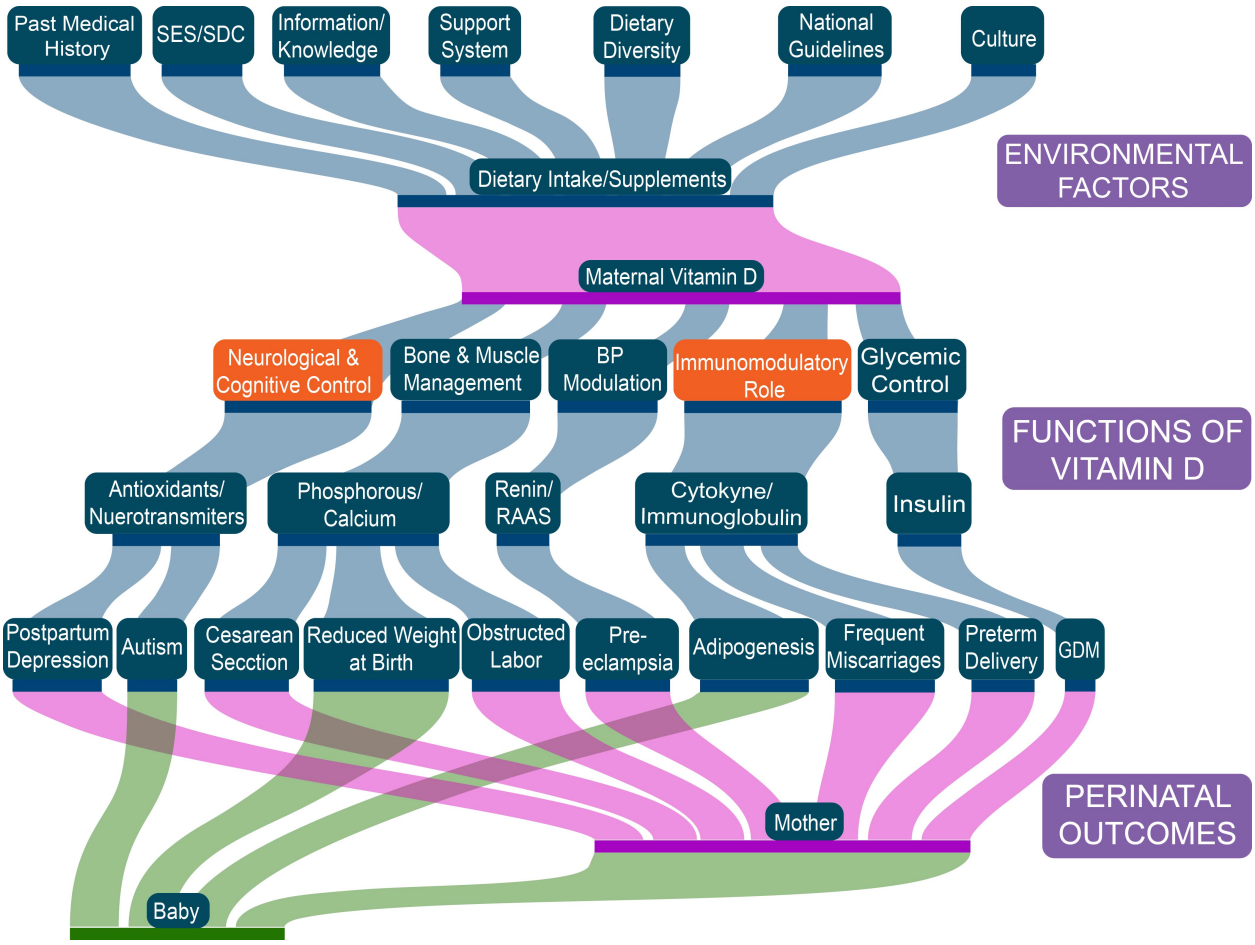


Figure 4.1: Relationship between society, maternal nutrition (vitamin D), and the effects on mother and offspring: a Sankey diagram created based on Figure 3 from [115]. The orange color represents the findings from the exploration methods that the concepts related to brain development and immune system were enriched in ignorance statements and possible novel avenues to explore. SES/SDC = socioeconomic status/sociodemographic characteristics; BP = blood pressure; GDM = gestational diabetes mellitus.

for the gene list, we found ignorance enrichment of the concepts IMMUNE SYSTEM and BRAIN DEVELOPMENT (see Figure 4.1). Yadama *et al.* [6] also found the immune system through their functional enrichment analysis and suggested it modulates the effects of maternal vitamin D intake and sPTB. They suggested increasing maternal vitamin D intake and that future work should to be done to fully determine the relationship between vitamin D, the immune system, and sPTB. Our results provide immune system ignorance statements that can be explored in future work. For the concept brain development, they did not mention anything related to the brain in their paper [6]. Thus, we found a novel concept that related to questions based on their gene list.

We provide all the ignorance statements and suggest questions for future exploration. Also, the concept brain development implies the field of neuroscience, providing a field for the researchers to look for answers. The ignorance-base provided a novel putative research area with specific ignorance statements or knowledge goals to pursue in future work based on a gene list.

The purpose of this work was to create the first ignorance-base with a more nuanced categorization of known unknowns and grounded in ontologies, going beyond a search engine based on MeSH terms with only two categories of known unknowns [17]. We highlight its power by providing summaries and visualizations (not just tables of sentences and articles) based on an input topic or a set of experimental results. Previous work [17] cannot support experimental results as an input. Thus, to the best of our knowledge, we created the first system that helps researchers explore the state of our collective scientific ignorance not only surrounding a topic but also experimental results. (A list of the formal terms we have introduced here and their definitions are shown in Table 4.1.)

Table 4.1: Term definitions.

Term	Definition
Ignorance	community/collective/scientific known unknowns
Knowledge-base	a database of known information
Ignorance-base	a knowledge-base, created from the literature, with additional annotations for the sentences that are ignorance statements
Statements of ignorance	statements of incomplete or missing knowledge categorized based on the entailed knowledge goal
Knowledge goal	the next actionable step based on the given unknown
Biomedical concept classification/recognition	automatically identifying and mapping biomedical entities to ontologies
Ontologies	controlled vocabularies with specified relationships
Open biomedical ontologies (OBOs)	an effort to create standardized ontologies for use across biological and medical domains
Lexical cue	words or phrases that signify a statement of ignorance
Taxonomy of ignorance	a categorization of ignorance statements based on the entailed knowledge goal
Exploration by topic	automatically find statements of ignorance related to a topic from the ignorance-base
Exploration by experimental results	contextualize experimental results in terms of statements of ignorance from the ignorance-base to understand what questions may bear on them
Ignorance enrichment	a method to identify biomedical concepts that are over-represented in a set of ignorance statements as compared to all sentences, and thus may be a new promising avenue to explore in relation to the input topic
Ignorance-category enrichment	a method to identify ignorance categories that are over-represented in a subset of ignorance statements as compared to all ignorance statements in order to illuminate the types of knowledge goals to pursue and to map out how they change over time

Methods

We combined the best-performing ignorance classifiers (Chapter 3) with state-of-the-art biomedical concept classifiers (Chapter 2) to create an ignorance-base that allowed us to explore it by a topic and by experimental results. The ignorance-base can be queried by ontology concepts, ignorance categories, specific lexical cues, or any combination of the three.

The rest of this section will be organized into the following subsections:

1. Creating the Ignorance-Base: Combining ignorance and biomedical concept classifiers
2. Ignorance-Base: Exploration by topic
3. Ignorance-Base: Exploration by experimental results

Materials

The inputs for all systems were scientific prenatal nutrition articles. We used full-text articles from the PubMed Central Open Access (PMCOA) subset of PubMed [159], allowing us more data and the ability to share it publicly. 1,643 prenatal nutrition articles (1939-2018) were gathered from querying PMCOA for 54 regular expressions determined in consultation with a prenatal nutrition expert, Teri L. Hernandez (T.L.H.) (same query and articles from Chapter 3). All articles were provided in XML format and converted to text format parsing through the XML using a script in Java. All subsequent computation was implemented in Python 3, with its associated packages. The classification frameworks and models were from Chapter 2 and 3.

To connect the ignorance statements to the biomedical concepts, the ignorance-base was built upon the PheKnowLator knowledge graph (PheKnowLator_v3.0.2_full _subclass_relationsOnly_OWLNETS_SUBCLASS_purified_NetworkxMultiDiGraph.gpickle), which semantically integrated eleven OBOs [276, 277]. A gene list (our motivating example of experimental results) was gathered from a PMCOA article (PMC6988958) [6]. We also used DAVID [278], a tool for functional annotation and enrichment analyses of gene lists, as a standard approach for gene list analyses to compare to our ignorance approach of ignorance enrichment.

Computation used a contemporary laptop (MacBook Pro 2019) and an NIH-funded shared supercomputing resource [235] that included:

- 55 standard compute nodes with 64 hyperthreaded cores and 512GB of RAM
- 3 high-memory compute nodes with 48 cores and 1TB of RAM
- GPU nodes with Nvidia Tesla k40, Tesla k20, and Titan GPUs
- A high-speed Ethernet interconnect between 10 and 40 Gb/s

We used both CPUs and GPUs to predict statements of ignorance and OBOs. Code for the ignorance-base and exploration methods can be found at:
`https://github.com/UCDenver-ccp/Ignorance-Base`.

Creating the Ignorance-Base: Combining ignorance and biomedical concept classifiers

Table 4.2: Ignorance Taxonomy: definitions, knowledge goals, example cues, and total cue count. The categories in bold were only narrow categories. Abbreviations are in italics. The ignorance-base was built upon this ignorance taxonomy. (Duplicate of Table 3.2 for reference in this Chapter)

Ignorance Category	**Definition**	**Knowledge Goal**	**Example Cues**	**Total Cues**
indication of *answered* research *question*	A statement of a goal or objective of a study that is attempted or completed during the study.	to find the answer(s) in the article; determine if the question(s) is (are) fully answered in the article	aim, goal, objective, our study, sought, to determine	64
indication of *unknown* or *novel* research topic or assertion	A statement that indicates something is not known (a lack of information), or information is presented for the first time (new or novel) and a significant amount of research is needed; not a statement about the absence of something.	to explore the unknown further to gain any insights	could not find, don't know, elusive, not...established, uncertain, still unclear	155

indication of *explicit* research *inquiry*	An explicit statement of inquiry (with a question mark or question word such as how, where, what, why).	to find answers to the question and/or discover methodologies that will help answer the question	?, what, where, wondered, why	19
indication of proposed or *incompletely understood* research topic or assertion	A positive or negative statement proposing a possible/feasible explanation for a phenomenon on the basis of limited evidence as a starting point for further investigation OR a statement that information is needed to support an assertion or claim, including both positive and negative statements. Either a statement that some evidence already exists, explaining how current findings support previous work, adding confidence to a claim OR a statement that information is limited, more research is needed or is ongoing including limitations – biases or short comings related to the study design and execution.	to gather more evidence to support the claim OR conduct more research to determine the validity of the claim; complete the partial picture; consider the short comings and biases for the next experiment and how it can be addressed.	a good understanding, believe, evidence...limited, has been suggested, hypothesis, no studies, possibly, preliminary stage, remains under investigation, still being discovered, support, trend	797

indication of *indefinite relationship* among research variables	A statement about a connection, link, or association between at least 2 variables; connectedness between entities and/or interactions representing their relatedness or influence.	to confirm the connection, link, or association between variables; determine the full underlying relationship between variables	affect, associated, correlate, factor, influence, interact, link, pattern, tend	198
indication of *largely understood* research topic or assertion	A statement staking a claim to the most likely explanation, relationship, or phenomenon; assumes that there is a good chance this understanding is correct.	to determine if the most likely option is correct or if another option is more feasible	almost all, assumed, concluding, evident, it is clear, most likely, thus	202
indication of *anomalous* or *curious* research finding	A statement of a surprising result, conclusion, observation or situation; the researchers were not expecting the result, conclusion, observation or situation but are intrigued by it.	to explore the surprising result, conclusion, or situation more and determine if the result, conclusion, observation, or situation is repeatable	appeared to be, interestingly, noteworthy, surprisingly	113

indication of *alternative* research options or *controversy* of research	Either an explicit statement of multiple (at least 2) choices, actions, approaches, or methods that need to be experimentally determined, including statements with an implied second option, such as "whether". This includes a statement of disagreement amongst researchers OR a lack of consensus OR at least two possible answers presented as results from different researchers - usually in reference to previous results and stated when results disagree with each other OR contradictions.	to determine the correct option or a better option and if there are disagreements, to determine the truth to break any disagreements	cannot rule out, claims, has been challenged, whether, whilst	221
indication of *difficult* research *task*	A statement of something not easily done, accomplished, comprehended, or solved; or a complicated thing with a multitude of underlying pieces or parts; heterogeneity; excludes medical complications.	to create methods to study the complicated system and to better understand any piece of the complicated system; potentially requires new experiments or better techniques	not feasible, remains...challenge, variability, rarely able to	98

indication of research *problem* or *complication*	A statement of issues, problems, mistakes, or medical complications that are cause for anxiety and/or worry.	to determine the gravity of the concern and determine if it needs to be dealt with before the next experiment or study	issue, error, insufficient, lack of reproducibility, publication bias, underestimated	98
indication of *future* research *work*	A statement of extensions, including next steps, directions, opportunities, approaches, or considerations of the described work that may be implemented at some future time point. This also includes a statement of suggestion or a proposal as to the next best course of action, especially one put forward by an authoritative body; advice telling someone the best action to take.	to determine the next course of action based on this future work proposal	additional research, are needed, continue to explore, further study, more...studies, recommend, warrants, worthy of closer attention	258
indication of *future* research *prediction*	A statement of extrapolation of given data into the future and/or from past observations, without reference to next steps.	to run the simulation or experiment to determine if the prediction is correct; publicize the outcomes of the study to the correct people	allow, expect, if so, serve as a basis, will	27

indication of *important consideration* for future research work	A statement calling for attention including an action needed to be taken immediately or information that needs to be disseminated immediately OR critical: being in or verging on a state of crisis or emergency OR urgently needed OR absolutely necessary.	to take the urgent action ASAP or distribute the knowledge ASAP	call for action, cautious, crucial, emphasis, global problem, high on the agenda, necessary, relevant to note, vital	263

The goal of the ignorance-base was to capture all ignorance statements and their biomedical subjects or concepts for the prenatal nutrition literature. Thus, to create the ignorance-base we combined the ignorance and biomedical concept classifiers (Chapter 3 and 2 respectively) over all 1,643 prenatal nutrition articles. The ignorance-base included all sentences from all articles to capture all biomedical concepts for comparison of our ignorance approach (only ignorance statements) to the standard approach (all articles). For ignorance classification, we used the 91 gold standard corpus articles, and ran the best ignorance classifiers (Chapter 3), the ensemble of 13 different binary classifiers, over the other 1,552 articles (see Table 4.2 for the ignorance taxonomy). Similarly, we ran our state-of-the-art biomedical concept classifiers (Chapter 2) over all 1,643 articles to automatically identify biomedical concepts represented in ten OBOs. These ten OBOs were the same ones used to manually annotate the CRAFT Corpus [166, 170], a corpus of mouse articles (see Chapter 2 for more details):

1. Chemical Entities of Biological Interest (ChEBI)
2. Cell Ontology (CL)
3. Gene Ontology (GO):
    a) Gene Ontology Biological Process (GO_BP)

   b) Gene Ontology Cellular Component (GO_CC)

   c) Gene Ontology Molecular Function (GO_MF)

4. Molecular Process Ontology (MOP)

5. NCBI Taxonomy (NCBITaxon)

6. Protein Ontology (PR)

7. Sequence Ontology (SO)

8. Uber-anatomy Ontology (UBERON)

For each of these ontologies, two sets of concept annotations were created for CRAFT (and appear in the public distribution): only proper classes of these OBOs and another adding in extension classes to better integrate the OBOs (created by the semantic annotation lead but defined in terms of proper OBO classes). We employed automatic concept recognition of our prenatal corpus with both the core OBOs and with the corresponding extended OBOs (suffixed with "_EXT"). Note that classification performance on the OBOs_EXT was lower in general compared to the OBOs, especially for PR and PR_EXT, so caution should be taken in interpreting those results. We focused on the proper OBO classes for this work, but have the data and results for both. (PheKnowLator does not currently have the OBOs_EXT, but it is easily extendable.) We used the best-performing models for each OBO (highest F1 score - harmonic mean between precision and recall), with most F1 scores ranging from 0.7-0.98 with the exception of PR at 0.53 (see Table 5 in Chapter 2). Note that even though all of the classifiers performed close to the state of the art for both tasks, they were still automated, so we draw conclusions cautiously. Because of this, we manually reviewed a random sampling of the identified biomedical concepts to check performance (a few hundred of each). Combining ignorance and biomedical concept classifiers automatically captured all ignorance statements and biomedical concepts for the 1,643 prenatal nutrition articles.

For clarification, the underlying data for the ignorance-base included sentences like the example above and another example: "it has an important role in BONE HOMEOSTASIS,

BRAIN DEVELOPMENT and MODULATION OF the IMMUNE SYSTEM and yet the impact of ANTENATAL VITAMIN D deficiency on infant outcomes is poorly understood" (PMC4072587). In terms of ignorance, the lexical cues important mapped to *important consideration*, role and impact mapped to *indefinite relationship*, yet mapped to *anomalous/curious*, and poorly understood to *unknown/novel*. For the biomedical concepts, BONE HOMEOSTASIS mapped to GO:0060348 (bone development), BRAIN DEVELOPMENT mapped to GO:0007420, BRAIN also mapped to UBERON: 0000955, MODULATION OF...IMMUNE SYSTEM mapped to GO:0002682, IMMUNE SYSTEM also mapped to UBERON:0002405, ANTENATAL mapped to GO:0007567, and VITAMIN D mapped to ChEBI:27300. This was all identified by the classifiers. Note that we also identified biomedical concepts in non-ignorance statements. The entailed knowledge goal was to explore the relationship between prenatal vitamin D deficiency and infant outcomes through the important role of vitamin D.

The power of the ignorance-base was in its exploration. Thus, in order to explore these data, we created a network representation of the ignorance-base to connect all sentences from these articles using both the ignorance lexical cues and biomedical concepts (see Figure 4.2). We combined all the literature data to connect sentences that have the same ignorance lexical cues, such as poorly understood, and then used PheKnowLator to combine all the sentences with the same biomedical concepts, such as VITAMIN D. The semantic integration of PheKnowLator allowed us to not only connect our sentences to the biomedical concepts, but also to each other; these connections were used in exploration by experimental results. This network can be used to search for all sentences that included the biomedical concept VITAMIN D, the lexical cue poorly understood, sentences with the ignorance category *unknown/novel*, or any combination of these features. Each sentence also related back to an article with its own metadata to be used for summaries and visualizations. For example, the publication date was used to map how ignorance categories change over time for a topic. Note that all sentences in all articles were included

whether or not they contained ignorance statements, allowing for the ignorance enrichment comparison to the background information.

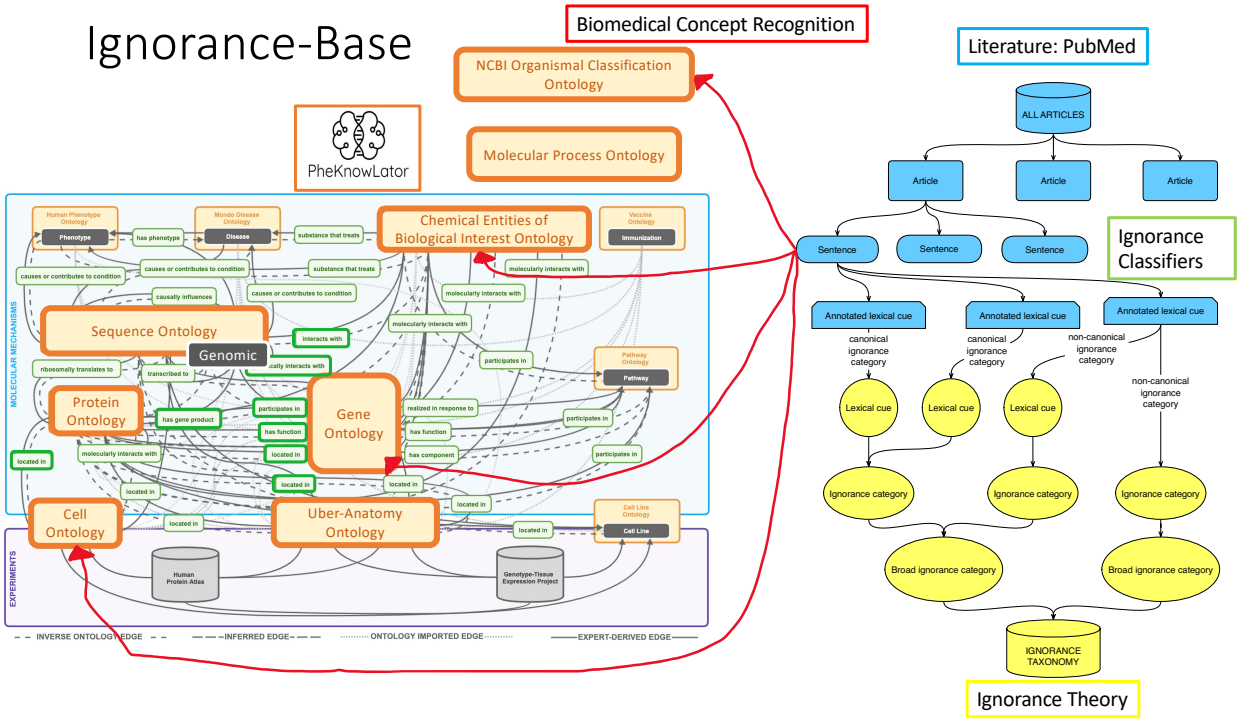


Figure 4.2: Network representation of the ignorance-base: The top right corner is the literature connecting the articles via tokenized sentences (in blue) to the ignorance taxonomy (in yellow) through the ignorance classifiers (the annotated lexical cues). Note that in order to capture lexical cues that map to the non-canonical ignorance category, we mapped the annotated lexical cue both to the canonical ignorance category and to the non-canonical one. The sentences also connect to the biomedical concepts on the left with PheKnowLator [276, 277] using the biomedical concept classifiers with the ontologies of interest in bold and larger font. Note that PheKnowLator [276, 277] does not include the Molecular Process Ontology or the NCBI Organismal Classification Ontology, which is why they are at the top not integrated with the rest.

Ignorance-Base: Exploration by topic

The goal of exploration by topic was to find biomedical concepts or keywords (areas of research) with lots of ignorance statements (questions) that are ripe for future research. To determine if the ignorance approach could do this, we compared the standard approach of finding biomedical concepts or keywords to our ignorance approach. Our motivating example is to help researchers find the most pertinent questions and topics in vitamin D to study in future research. An input topic consisted of a list of ontology concepts in PheKnowLator [276, 277]. In consultation with a prenatal nutrition specialist (T.L.H.), we mapped the topic of vitamin D to four

OBO concepts narrowed from 38 exact matches (280 partial matches): VITAMIN D (ChEBI:27300), D3 VITAMINS (ChEBI:73558), CALCIOL/VITAMIN D3 (ChEBI:28940), and VITAMIN D2 (ChEBI:28934). Note that going forward, when we refer to vitamin D, we mean all four search terms. For the standard approach, we gathered all sentences from the ignorance-base that included terms from this vitamin D OBO concept list. For the ignorance approach, we filtered these vitamin D sentences for only those that included an ignorance lexical cue. We analyzed the data to find areas of research to explore.

Our goal was to determine if the ignorance approach provided an alternative targeted exploration of a topic that was distinct from an unfiltered or standard approach to help researches find the most pertinent questions. Thus, we conducted two analyses for both approaches and compared them: (1) an analysis of the most frequent concepts in the subset of sentences not including vitamin D itself (most frequent) and (2) an analysis of concepts enriched in the subset of sentences as opposed to the background sentences in our ignorance-base, again not including vitamin D itself (concept or ignorance enrichment). We present both biomedical concept clouds and word clouds along with frequency tables to explore the most frequent terms. For enrichment, we used the hypergeometric test for over-representation with both Bonferroni (family-wise error rate) and Benjamini-Hochberg (false discovery rate) multiple testing corrections [28, 279]. In comparing the approaches (see Figure 4.3), if a concept was more frequent or enriched in the standard approach but not in ignorance, then it probably was established information. If a concept appeared in both approaches, then it may be currently studied. If a concept only appeared in the ignorance approach then it may be an emerging topic. Note that concepts that were not frequent nor enriched in either approach were not interesting to us. To find pertinent questions, the goal was to find concepts or keywords that were currently studied or emerging topics.

To also help the researcher understand the general ignorance landscape and narrow in on types of question to ask, we showed the types of questions asked (ignorance-category enrichment) and how they change over time for the ignorance approach. For the types of questions asked, we compared the percentage of each ignorance category between vitamin D ignorance statements and

all ignorance statements (a bar chart). Further, we determined which ignorance categories were enriched in vitamin D ignorance statements compared to all ignorance statements. To show how ignorance statements changed over time, we bubble plotted the ignorance categories per article over time, with the bubble size representing the percentage of sentences in an article scaled by the total number of sentences of that category. Using these methods, the researchers continued to deep dive into ignorance statements that included the topic, enriched concepts, and enriched ignorance categories to find knowledge goals to pursue for future research. In order to determine if we found truly **novel** avenues to explore using our ignorance approach, we consulted both our prenatal nutrition specialist, T.L.H., and PubMed to determine the state of the research.

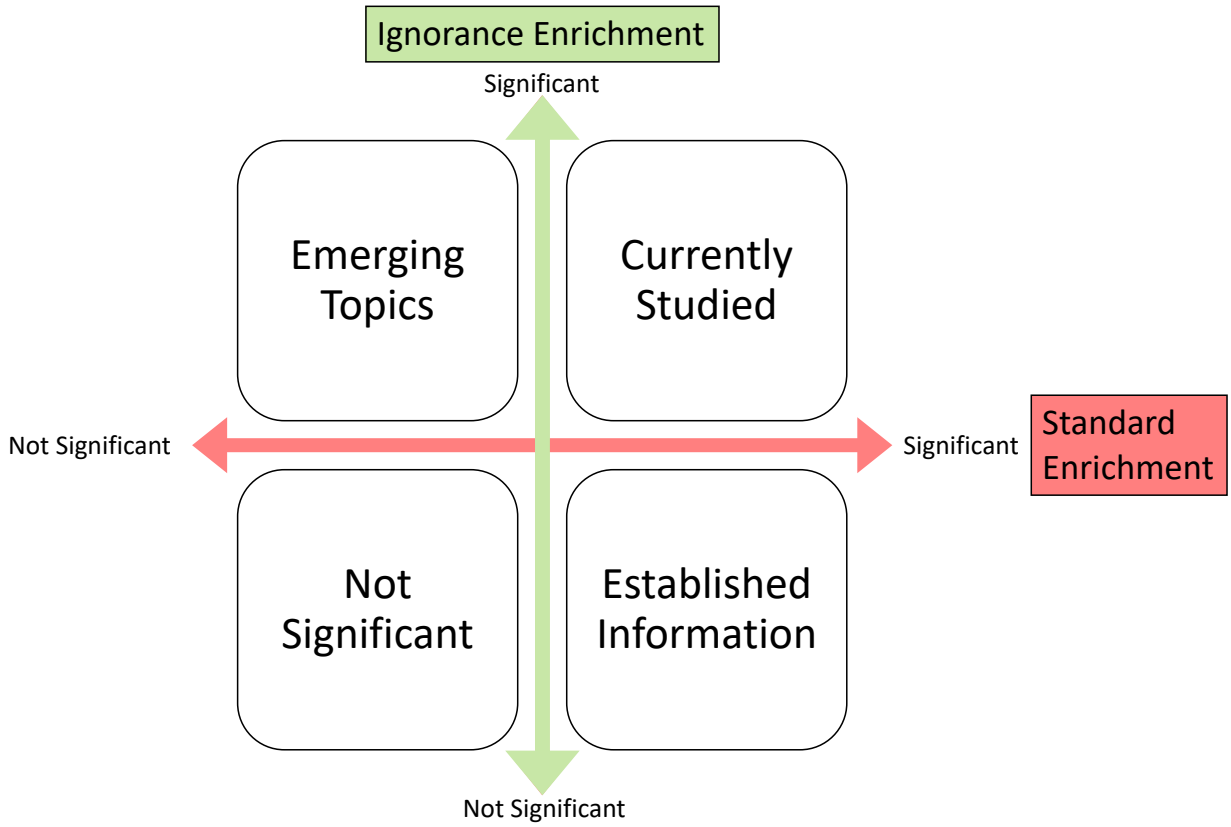


Figure 4.3: Ignorance vs. Standard Approach Results Chart: The interpretation of the results comparing the ignorance approach to the standard approach.

Ignorance-Base: Exploration by experimental results

Connecting experimental results such as a gene list to ignorance statements can identify questions that may bear on the results, providing new avenues for exploration, potentially from

other fields. The goal was to find biomedical concepts (areas of research) that ideally imply another field, with lots of ignorance statements (questions) that are ripe for future research, similar to exploration by topic. Thus, exploration by experimental results used the same methods as exploration by topic with some added pre-processing steps and analyses based on the relationship between the inputs. The input topic was still an OBO concept list, but the extra pre-processing step was connecting the experimental results to OBO concepts in PheKnowLator [276, 277]. In general, as long as the experimental results can be mapped to OBO concepts in and through PheKnowLator, we can connect them to the ignorance-base. Querying the ignorance-base with this input topic (gene list from [6]) provided lots of ignorance statements for exploration. We analyzed the data to provide the researchers with the ignorance context for the experimental results.

Using our motivating example, we mapped the vitamin D and sPTB gene list (Entrez genes) [6] to the genomic part of the sequence ontology (SO) and the corresponding proteins in the protein ontology (PR). This initialized the list of ontology terms to use for our search. To add more terms, we utilized the relations ontology (RO) that connected the different ontologies together in PheKnowLator. For example, the relation “interacts with" (RO:0002434) connected proteins or genes to chemicals (ChEBI). This yielded a large list of ontology terms; we then found all the sentences that contained these terms (gene list sentences - our sentences of interest). Note that not all OBO concepts connected to a sentence because our articles did not have examples of every concept. From here we performed all of the same analyses as we did for exploration by topic, including finding articles, sentences, ignorance categories, and concepts to investigate. Further, we added three more analyses: (1) gene list coverage (prioritizing the OBO concepts that connect to the most genes), (2) comparisons to other enrichment analyses such as DAVID, and (3) comparisons to any other findings about the gene list such as findings from a paper. (See figure 4.4 for the exploration by experimental results pipeline.) Our goal was to find emerging topics or concepts in relation to the gene list as new avenues for exploration.

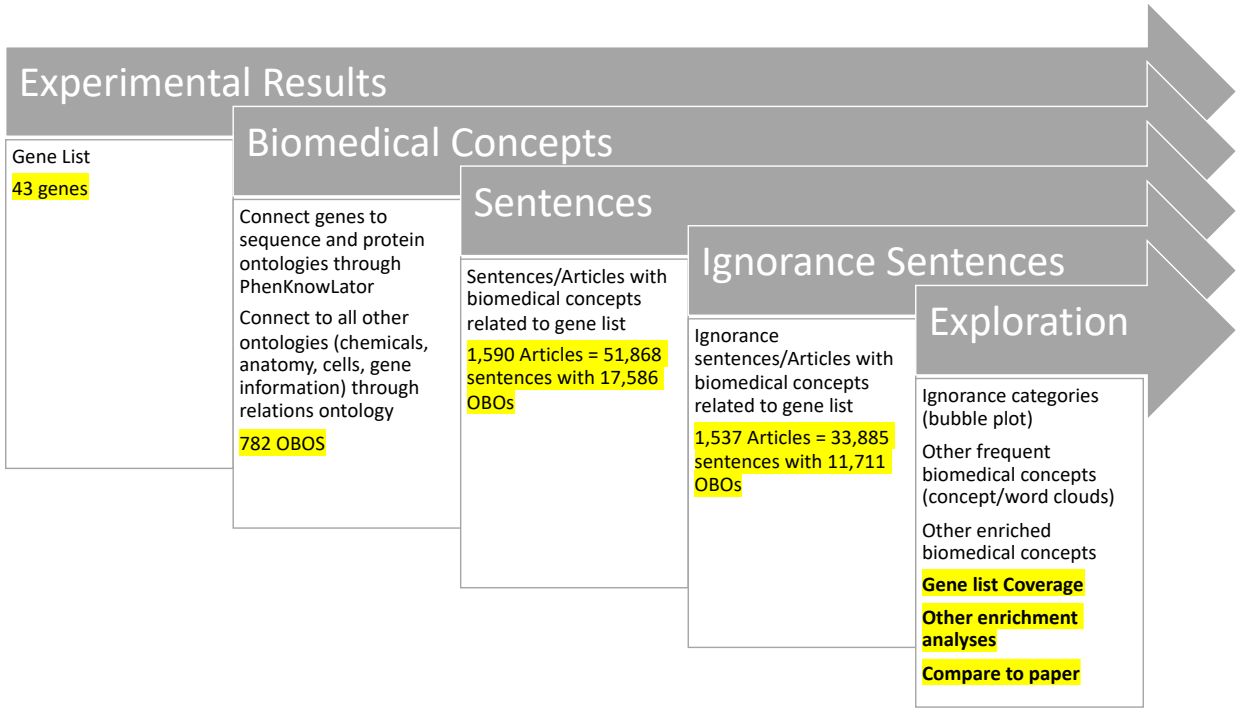


Figure 4.4: Exploration by experimental results (gene list) pipeline: The results are in yellow highlights for the example presented here. For exploration at the end of the pipeline, the three not highlighted are the same as exploration by topic and the three highlighted are the new additions based on a gene list.

Gene list coverage helped prioritize which OBO concepts highly related to the gene list. As we mapped the gene list to the OBO concepts, some OBO concepts had many genes map to them, implying that these OBO concepts were potentially more relevant to the gene list than concepts with fewer genes mapping to them. Thus, we sorted the OBO concept list by these high coverage OBO concepts and looked to see if those were enriched in all of the gene list sentences and/or in the gene list ignorance sentences. Of interest were currently studied and emerging topics concepts in comparing the ignorance approach to the standard approach (see Figure 4.3). This provided a smaller and more refined list to start exploring.

Canonical enrichment methods also helped prioritize OBO concepts. We compared our ignorance enrichment method to them, allowing us to both enhance the canonical methods and find new lines of investigation. From tools such as DAVID [278], we got a list of enriched OBOs (GO concepts) by using the gene list and functional annotations from their entailed knowledge-bases. We then found and examined any ignorance statements that contained the

concepts linked by DAVID. Our ignorance approach also provided a list of OBO concepts enriched in ignorance statements. Thus, we compared and examined these two lists to add the ignorance layer to the classic enrichment analysis in order to find new concepts that may be currently unknown to the functional annotation knowledge-bases, *i.e.,* potentially emerging topics related to the gene list. These concepts could be new avenues for exploration for both the functional annotation tool and for the researchers of the original gene list.

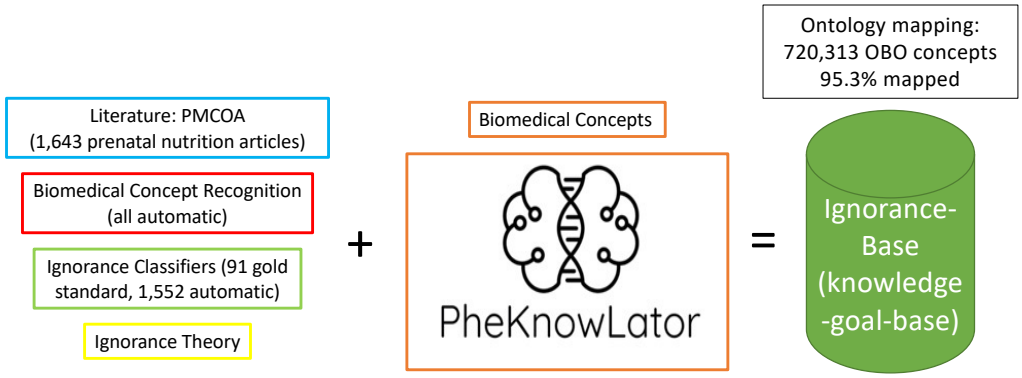
Lastly, given that our gene list came from a paper, we compared our ignorance-approach findings to the paper findings. We identified questions that may bear on it, providing new avenues for exploration from other fields. Thus we determined if the paper had similar findings or if we uncovered something new. Yadama et. al. [6] focused on the immune system as their main findings, and we looked for a new path for the researchers to explore based on concepts and knowledge goal statements that were not mentioned or cited in the article. We further consulted our prenatal nutrition expert (T.L.H.) and the literature to corroborate any findings. Our results explore these three added analyses.

Results

The Ignorance-Base: The power of combining ignorance and biomedical concept classifiers

The ignorance-base captured the connection between our collective scientific ignorance (ignorance taxonomy) and knowledge (PheKnowLator) through sentences from the literature. It contained a wealth of data (see Figure 4.5) ripe for exploration via the network (see Figure 4.2). The short manual review of some random sentences from the ignorance-base suggested that both the ignorance and biomedical concept classifiers generally correctly identified concepts (data not shown). The power of combining these two types of classifiers was in the exploration methods. However, to provide a base understanding of the ignorance-base, we present some summary statistics.

Our results suggest that ignorance statements proliferated throughout the ignorance-base. The 1,643 articles, spanning years 1939 to 2018 (see Figure 4.6), contained 327,724 sentences with over 11 million words. Just over half of those sentences had an ignorance lexical cue



Nodes	857,766	780,753	1,638,519
Edges	1,815,645	5,072,064	6,887,709

Figure 4.5: Summary information for the ignorance-base. The ignorance-base was a combination of biomedical concept classifiers and ignorance classifiers over a corpus of prenatal nutrition articles. The network representation connected the literature to the ignorance theory and biomedical concepts via PhenKnowLator [276, 277]. Note that 95.3% of the ontology annotations mapped to PheKnowLator.

(182,892), with articles averaging a total of 111 cues (with a median of 93). Every section of the articles had ignorance cues aside from the titles, with the most in the discussion and conclusion sections and the fewest in the abstract and results. These results agreed with our findings in Chapter 3 based on our smaller ignorance corpus. Our collective scientific ignorance was represented throughout the literature and was ripe for exploration.

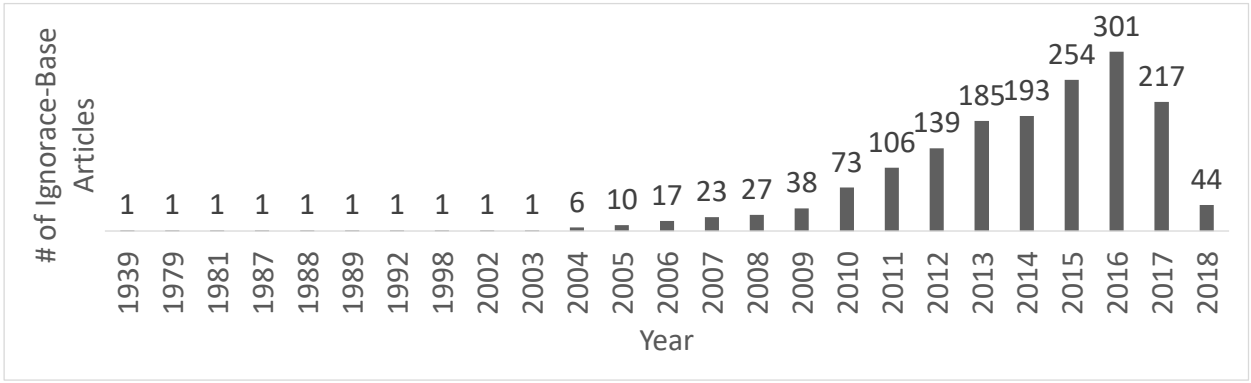


Figure 4.6: Article date distribution for the ignorance-base (1939-2018).

Further, our ignorance taxonomy was not only a categorization system of ignorance statements via knowledge goals, but also a depiction of the research life-cycle and how researchers discuss our collective scientific ignorance (see Figure 4.7 with proper definitions and example

lexical cues in Table 4.2). This can help researchers, including students, understand the research process and how it is discussed in the literature. Underneath the 13 categories of ignorance (with 3 broader ones in all caps in Figure 4.7) were 2,513 unique lexical cues both from related work and collected by our annotation tasks (see Chapter 3). 1,822 of them had examples in our ignorance-base. Further, the ignorance classifiers found 5,637 new unique lexical cues that signified ignorance, which were added to the ignorance-base and noted as such. Many of them were variations of the ones already captured and others were new, such as “not as yet", “interplay", and “have begun to illuminate". Our ignorance classifiers recognized more complex language than just a dictionary match, potentially hinting at generalizability. Overall, there were 517,445 ignorance annotations involving 7,459 unique lexical cues; this reinforced the diversity of ways that ignorance was expressed in the literature (agreeing with our results in Chapter 3). The ignorance taxonomy can help researchers understand the research life-cycle and how it is discussed in the literature.

Our ignorance-base also contained a wealth of different types of biomedical concepts, meaning our ignorance-base captured all different types of biomedical subjects ready to be explored in future work. Our biomedical concept classifiers (Chapter 2) identified 720,313 concepts involving 19,883 unique concepts from all of the ontologies and almost all of them (95.3%) mapped to PheKnowLator [276, 277] (see Figure 4.5). Note that we can only represent the biomedical concepts in PheKnowLator that were also captured by our biomedical concept classifiers. This overlap included six of our eight ontologies (missing MOP and NCBITaxon) or six of the eleven PheKnowLator ones (missing the human phenotype ontology, MONDO disease ontology, vaccine ontology, pathway ontology, and cell line ontology). In terms of errors, there were 850 predicted unique OBO concepts with non-existent OBO identifiers as our biomedical concept classifiers predicted identifiers character by character which allows for more errors (see Chapter 2 for more details). The other 1,432 unique OBO concepts that did not map seem to either be from the two ontologies not included in PheKnowLator (MOP and NCBITaxon) or were terms no longer used/depreciated from the ontologies. Even still, the ignorance-base captured a

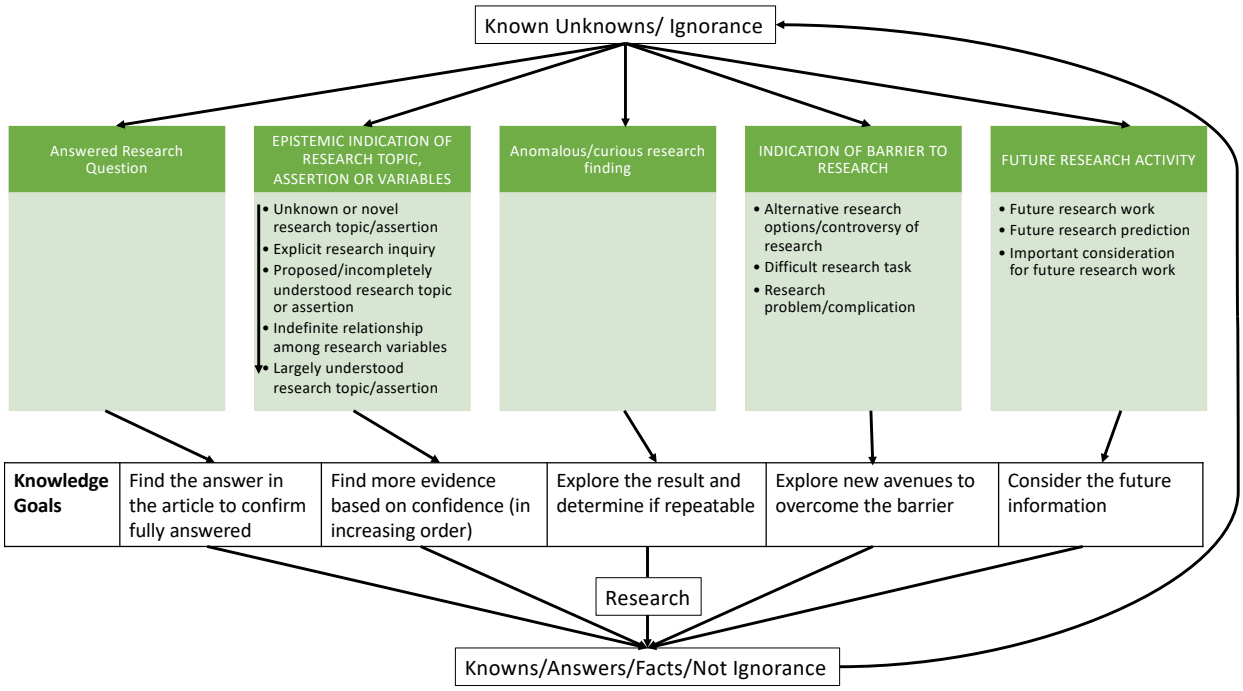


Figure 4.7: Ignorance taxonomy embedded in the research context: Starting from the top, research starts from known unknowns or ignorance. Our ignorance taxonomy is in green (an ignorance statement is an indication of each ignorance category) with knowledge goals underneath. Research is then conducted based on the knowledge goals to get answers; these then filter back to the known unknowns to identify the next research questions.

variety of biomedical concepts ready to be explored. Overall, the ignorance-base contained a great deal of data consisting of all different types of OBO concepts, ignorance categories, and lexical cues ready to be explored by topic or by experimental results.

Focusing on ignorance statements provided an alternative targeted exploration of a topic that was distinct from an unfiltered or standard approach

Focusing on ignorance statements provided the researcher with new avenues of exploration with specific ignorance statements to pursue (see Figure 4.8). We present results for both the ignorance approach (vitamin D ignorance statements), the standard approach (all vitamin D sentences), and a comparison of the two.

There was a great deal of research on vitamin D and more specifically a plethora of ignorance statements. Searching through the ignorance-base for the four vitamin D terms yielded 521 articles with 10,841 sentences mentioning vitamin D (9% of the 118,419 sentences in the 521

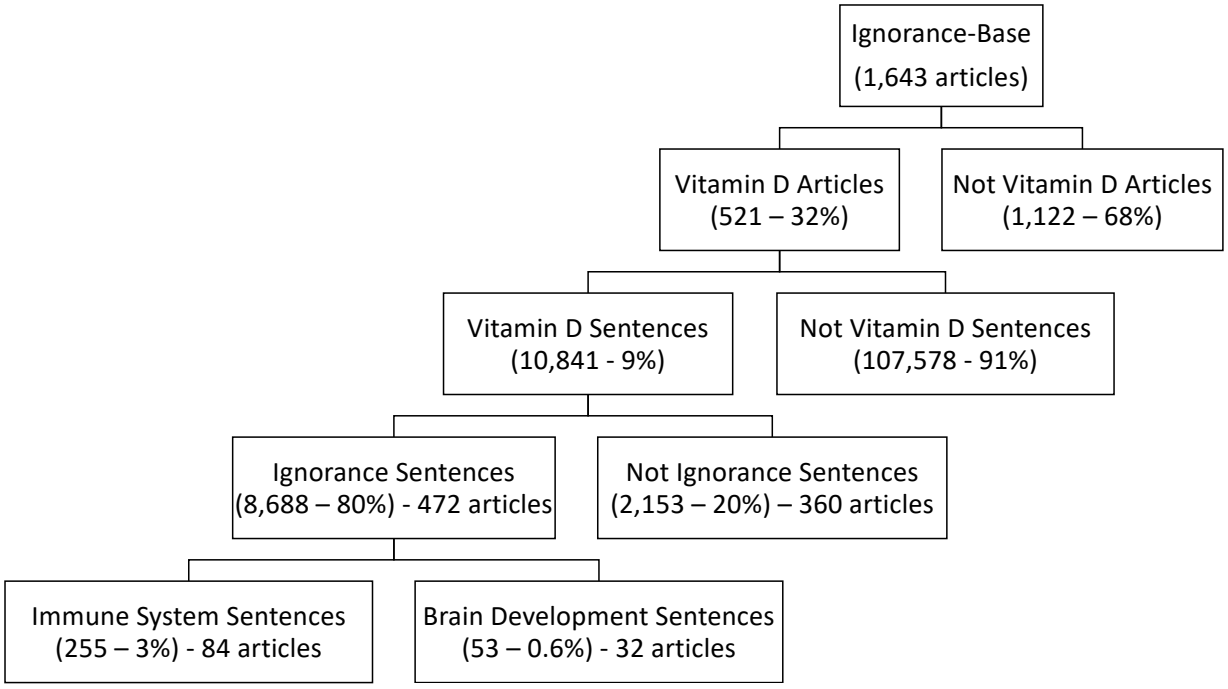


Figure 4.8: Exploring the ignorance-base by vitamin D: Searching the ignorance-base for vitamin D yielded many articles and sentences that can be explored using ignorance statements to find new research areas with lots of questions, including the IMMUNE SYSTEM and BRAIN DEVELOPMENT.

articles and 3% of all the sentences in the ignorance-base) (see Figure 4.8). Note that only the terms VITAMIN D and VITAMIN D2 pulled out sentences from the ignorance-base. These sentences included 17,584 unique biomedical concepts excluding the VITAMIN D concepts (88% of the total unique biomedical concepts). Of those VITAMIN D sentences, 8,688 sentences (80%) were ignorance statements spanning 472 articles. We explored the data to differentiate between knowns and unknowns to provide the researcher with biomedical concepts and corresponding ignorance statements for possible questions and topics for future work.

Focusing on term frequency provided some concepts of interest. The top five most frequent biomedical concepts for the ignorance approach included: FEMALE PREGNANCY, PARTURITION (giving birth), VERSICONOL ACETATE, BLOOD SERUM, and FEEDING BEHAVIOR (see Figure 4.9a). The first two aligned with the corpus theme of prenatal nutrition. VERSICONOL ACETATE is an intermediate in the biosynthesis of aflatoxin, which is a toxin produced by mold that may be toxic towards the vitamin D receptor in relation to rickets [280].

Vitamin D levels are mainly measured from the BLOOD SERUM, and FEEDING BEHAVIOR seems to highlight the importance of ingesting vitamin D. For the words, the most frequent terms were supplementation, maternal, status, levels, and women, and they also fit with the theme: supplements are suggested for many people, maternal and women fit with the corpus theme, and status and levels are measurement terms for vitamin D. None of these terms were surprising, which was a good sign that we captured meaningful information.

Term frequency can help prioritize areas to explore in relation to the input topic. Taking one of these terms or concepts provided avenues for the researchers to explore. For example, FEEDING BEHAVIOR (GO:0007631), defined as the behavior associated with the intake of food, was an interesting concept in relation to vitamin D. Vitamin D is naturally absorbed through sunlight and digestion. In searching for “vitamin D" and “feeding behavior" in the literature, we found that vitamin intake during pregnancy in general seems to affect both the metabolic system and food intake regulatory pathways in the offspring [281]. Therefore, frequent concepts that appear with a topic can provide keyword search terms for the researcher.

Further, frequent concepts can lead to some pertinent questions for the researchers. If the researchers chose FEEDING BEHAVIOR and VITAMIN D, they could look to the ignorance statements for research ideas. Most of these ignorance statements discussed the ingestion of vitamin D mainly via supplements, but also with some foods. The recommendations for ingestion all varied by study, agreeing with the findings from a systematic review [282]. One ignorance statement stood out specifically, “the high prevalence of Vitamin D deficiency in PREGNANT women is a worldwide health problem regardless of latitude, FOOD INTAKE or socio-economic status [93]" (PMC5941617) [283], citing a systematic review and meta-analysis that looked at vitamin D status globally [284]. All of these studies recommended vitamin D supplementation, but we could not find any studies that determine why supplementation is so low. How do supplements, specifically for vitamin D, fit into feeding behavior? A potential research topic could be to study what specific factors, beyond general socio-cultural factors, lead to women taking vitamin D supplements as part of their diet, especially for pregnancy. Studying this topic could

lead to novel methods that help mothers stay vitamin D sufficient throughout pregnancy, resulting in far less adverse outcomes for the offspring (see Figure 4.1). Thus, biomedical concept frequency can lead to a high impact research topic that could affect mothers everywhere.

Term frequency can also highlight terms that were more known vs. unknown. Comparing the ignorance-approach to the standard approach, the most frequent concepts in the standard approach were the same as the ignorance approach (see Figure 4.9b). This suggests that term frequency may not capture the difference in biomedical subjects between the two approaches. However, the top five words slightly differed between them: the word “deficiency" was the top most frequent term in the standard approach and the term “women" disappeared (see Figure 4.9). This may signify that vitamin D deficiency was established information, resulting in a lack of ignorance. At the same time, all these terms may be more unknown than known. Recall that 80% of the vitamin D sentences were ignorance statements, so it was possible that much of the context around VITAMIN D was still unknown in general. The ignorance frequency term list not only provided an avenue for exploration, FEEDING BEHAVIOR, with a potential research topic, but in addition may help distinguish between knowns and known unknowns, such as the word “deficiency".

To further distinguish between knowns and known unknowns, ignorance enrichment found at least three interesting new avenues to explore in relation to vitamin D that were captured by the standard approach, but buried amongst 275 concepts, and one avenue not captured by the standard approach at all. The ignorance approach found 11 ignorance enriched concepts, whereas the standard approach found 275, with an overlap of eight concepts (see Figure 4.10). Only focusing on the overlapping concepts, in the standard approach most of them were buried far down the list of enriched concepts ordered by enrichment p-value (indicated by the parentheses next to the overlapped concepts in Figure 4.10). Further, in comparing the two different approaches, the ignorance approach found concepts that were more systems or broad categories, including IMMUNE SYSTEM and BRAIN DEVELOPMENT, compared to the standard approach which were more specific entities, such as BLOOD SERUM and VITAMIN K. Ignorance enrichment

OBO Concept Cloud for Vitamin D Ignorance Statements

OBO ID	OBO ID Label	Frequency
go_0007565	female pregnancy	100.00%
go_0007567	parturition	36.06%
chebi_71657	versiconol acetate	29.21%
uberon_0001977	blood serum	22.45%
go_0007631	feeding behavior	14.73%

Word Cloud for Vitamin D Ignorance Statements

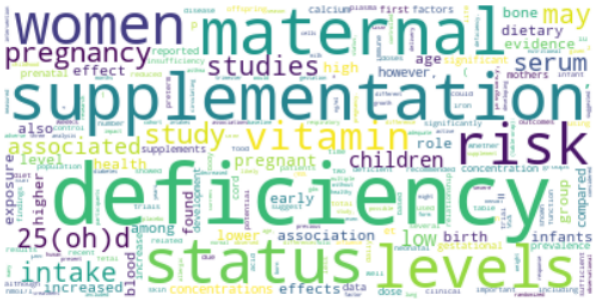
Word	Frequency
supplementation	14.34%
maternal	13.75%
status	13.31%
levels	12.36%
women	10.98%

(a) Vitamin D ignorance statements

OBO Concept Cloud for Vitamin D

OBO ID	OBO ID Label	Frequency
go_0007565	female pregnancy	100.00%
go_0007567	parturition	37.17%
chebi_71657	versiconol acetate	29.50%
uberon_0001977	blood serum	23.73%
go_0007631	feeding behavior	16.38%

Word Cloud for Vitamin D



Word	Frequency
deficiency	13.79%
supplementation	13.48%
maternal	13.39%
status	12.25%
levels	11.78%

(b) Vitamin D sentences

Figure 4.9: Term frequency results: Frequent Biomedical Concepts and Words in (a) vitamin D ignorance statements and (b) vitamin D sentences. Word clouds using biomedical concepts and words are on the left and right respectively. Also underneath are frequency tables of the top 5 most frequent concepts or words.

provided the researchers with a smaller list of targeted statements of knowledge goals to potentially pursue or spark ideas from.

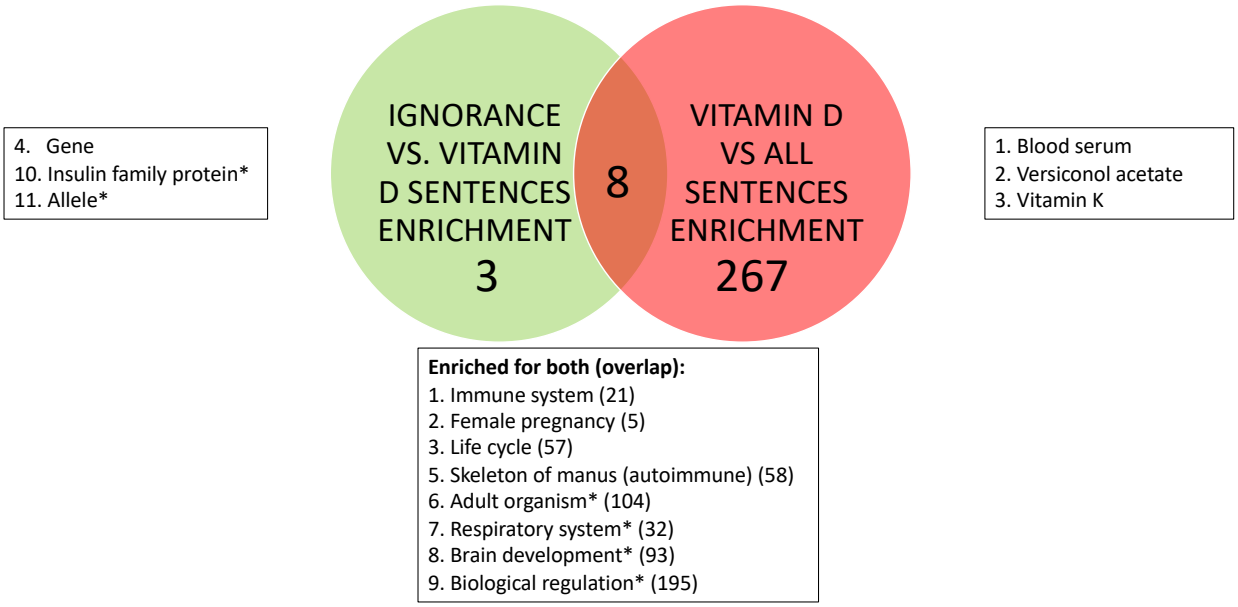


Figure 4.10: Comparison of standard and ignorance enrichment: A Venn diagram of biomedical concept enrichment between ignorance vitamin D (green) and just vitamin D (pink) sentences. Next to each bubble are concepts in their respective enrichment orders. The concepts in the middle are the overlap and the numbers correspond to the enrichment position for the ignorance vitamin D enrichment, with the standard enrichment position in parentheses. SKELETON OF MANUS is an error and is actually annotating autoimmune as in the parentheses. *Statistically significant with FDR but not family-wise error.

Focusing on the ignorance-enriched concepts also provided research topic ideas. T.L.H. and I determined that IMMUNE SYSTEM, RESPIRATORY SYSTEM, and BRAIN DEVELOPMENT (captured by the standard approach) were all interesting in relation to vitamin D. They were all currently studied concepts with more room for future work. We also found the insulin family protein (not captured by the standard approach) intriguing because many studies have attempted to determine the link between vitamin D and gestational diabetes mellitus [285]. Any of these concepts can be explored more thoroughly to find a research topic.

We explored the specific ignorance statements for the concepts of interest to narrow in on a research topic, just as with FEEDING BEHAVIOR. We chose to look at the IMMUNE SYSTEM ignorance statements, which provided the researcher with 255 ignorance statements with their entailed knowledge goals, spanning 84 articles (see Table 4.3 for the top eight articles with the

most ignorance statements). Note that only one article had no ignorance statements that included VITAMIN D or IMMUNE SYSTEM. Thus, only using the ignorance statements themselves, we have already found a set of articles and sentences for the researchers to review for a potential research topic.

Table 4.3: Articles with the most ignorance statements: The top eight articles for vitamin D and immune system in order of the most ignorance statements.

PMCID	Article Title	Date	# of ignorance statements	# of non-ignorance statements
PMC4448820	Inflammation and Nutritional Science for Programs/Policies and Interpretation of Research Evidence (INSPIRE)	5/15	22	0
PMC4251419	Vitamin D and immunity	12/14	17	0
PMC3717170	Vitamin D: beyond bone	5/13	13	1
PMC3277098	Vitamin D and allergic disease: sunlight at the end of the tunnel?	12/11	13	0
PMC4889866	Maternal Vitamin D Level Is Associated with Viral Toll-Like Receptor Triggered IL-10 Response but Not the Risk of Infectious Diseases in Infancy	5/16	11	0
PMC3347028	Vitamin D and its role during pregnancy in attaining optimal health of mother and fetus	3/12	10	0
PMC5489519	Vitamin D Modulation of TRAIL Expression in Human Milk and Mammary Epithelial Cells	6/17	8	0
PMC4302429	Vitamin D deficiency decreases survival of bacterial meningoencephalitis in mice	1/15	8	1

Choosing the IMMUNE SYSTEM was still quite a large topic with lots of ignorance statements and so we continued to narrow our search using ignorance-category enrichment. Understanding and tracing ignorance categories over time can both help the researchers narrow in on a research topic and show how the questions in a field were asked more broadly. VITAMIN D ignorance statements in general employed a wide-range of different ignorance categories (see Figure 4.11), spanning all the thirteen categories of ignorance with ten enriched in VITAMIN D

ignorance statements as compared to all ignorance statements (see the green highlights in Figure 4.11). For example, *unknown/novel* was enriched in this domain, pointing to large unknowns about the context of VITAMIN D in pregnancy and fetal development. To also understand how these questions changed over time, looking at the bubble plot for IMMUNE SYSTEM and VITAMIN D ignorance statements (see Figure 4.12) showed that *unknown/novel* was spread out amongst the different articles. This suggests that researchers have continued to discuss their full unknowns over time and it may be a good knowledge goal area for a research topic. Thus, the researchers can continue to narrow in on a research topic not only with a biomedical concept, such as IMMUNE SYSTEM, but also with an ignorance category, such as *unknown/novel*.

We chose to dive deeper into the *unknown/novel* category for VITAMIN D and IMMUNE SYSTEM, with the goal of finding the researchers the most pertinent questions as an example of exploration by topic. There were many *unknown/novel* ignorance sentences ripe for exploration. Below were some ignorance sentences (lowercase) from this set with the biomedical concepts capitalized and the ignorance lexical cues underlined:

1. “in the last five years, there has been an explosion of published data concerning the IMMUNE effects of VITAMIN D, yet little is known in this regard about the specific IMMUNE effects of VITAMIN D during PREGNANCY.” (PMC3347028)

2. “these results describe novel mechanisms and new concepts with regard to VITAMIN D and the IMMUNE SYSTEM and suggest therapeutic targets for the CONTROL of AUTOIMMUNE diseases.” (PMC3717170)

3. “however, findings regarding the combined effects of PRENATAL and POSTNATAL VITAMIN D status on fs [food sensitization], two of the most critical periods for IMMUNE SYSTEM DEVELOPMENT (19,20), are unclear.” (PMC3773018)

4. "it has an important role in BONE HOMEOSTASIS, BRAIN DEVELOPMENT and MODULATION OF the IMMUNE SYSTEM and yet the impact of ANTENATAL VITAMIN D deficiency on infant outcomes is poorly understood." (PMC4072587)

5. "background: VITAMIN D is known to affect IMMUNE function; however it is uncertain if VITAMIN D can alter the IMMUNE RESPONSE towards the persistent herpesviruses, EBV and CMV." (PMC4113768)

(Note that not all biomedical concepts were recognized by the biomedical concept classifiers.) The overall research topic or knowledge goal based on these statements was the need to explore the relationship between VITAMIN D and the IMMUNE SYSTEM especially in pregnancy. The same methods can be used for the other top enriched concepts including BRAIN DEVELOPMENT (data not shown). For BRAIN DEVELOPMENT, the overarching knowledge goal was the need to determine if VITAMIN D and BRAIN DEVELOPMENT were truly linked. Thus, from querying the ignorance-base for the topic VITAMIN D, the researchers now have research topic ideas based on ignorance statements and knowledge goals connecting VITAMIN D to the IMMUNE SYSTEM and BRAIN DEVELOPMENT (see Figure 4.8). The next step would be for the researcher to conduct research based on these ideas and publish papers with both some answers and some ignorance statements, continuing the cycle of research. Our exploration by topic methods provided multiple starting points for this research.

Connecting experimental results (a gene list) to ignorance statements can identify questions that may bear on it, providing new avenues for exploration, potentially from other fields

Similar to exploration by topic, exploration by experimental results provided the ignorance context for a gene list as possible future work for the researchers. Connecting a vitamin D and sPTB gene list from a paper [6] to ignorance statements found a new avenue for exploration, BRAIN DEVELOPMENT, that was not mentioned in the paper, and its implied field, neuroscience, could possibly help find answers. Following the exploration by experimental results pipeline (see Figure 4.4), the 43 genes mapped to 782 OBO concepts. These OBOs connected to 51,868 sentences (1,590 articles), gene list sentences, that included 17,586 unique OBO concepts

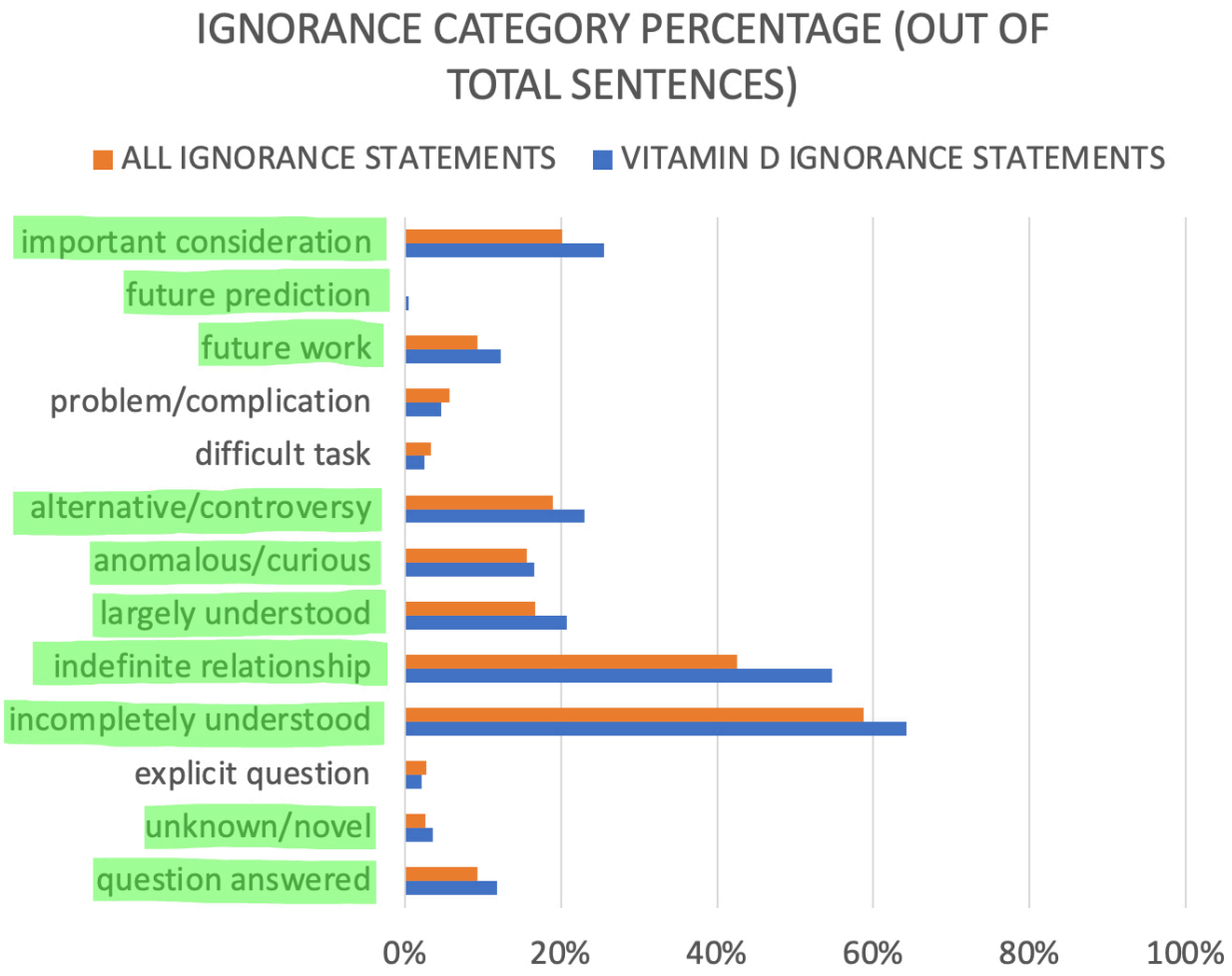


Figure 4.11: Ignorance-category enrichment: Ignorance vitamin D sentences compared to all ignorance sentences. The 10 categories highlighted in green were enriched for vitamin D ignorance statements compared to all ignorance statements.

(88% of the total unique OBO concepts), of which, 33,885 sentences (1,537 articles) were ignorance statements with 11,711 unique OBO concepts (59% of the total unique OBO concepts). This suggests that the majority of the gene list sentences were ignorance statements (65%). These data can be explored by topic using the OBO list similar to above, but we chose to instead focus on the three new analyses. The three new analyses were helpful to digest both the many OBO concepts and the many statements of ignorance connected to the gene list to provide areas of research to explore in future work.

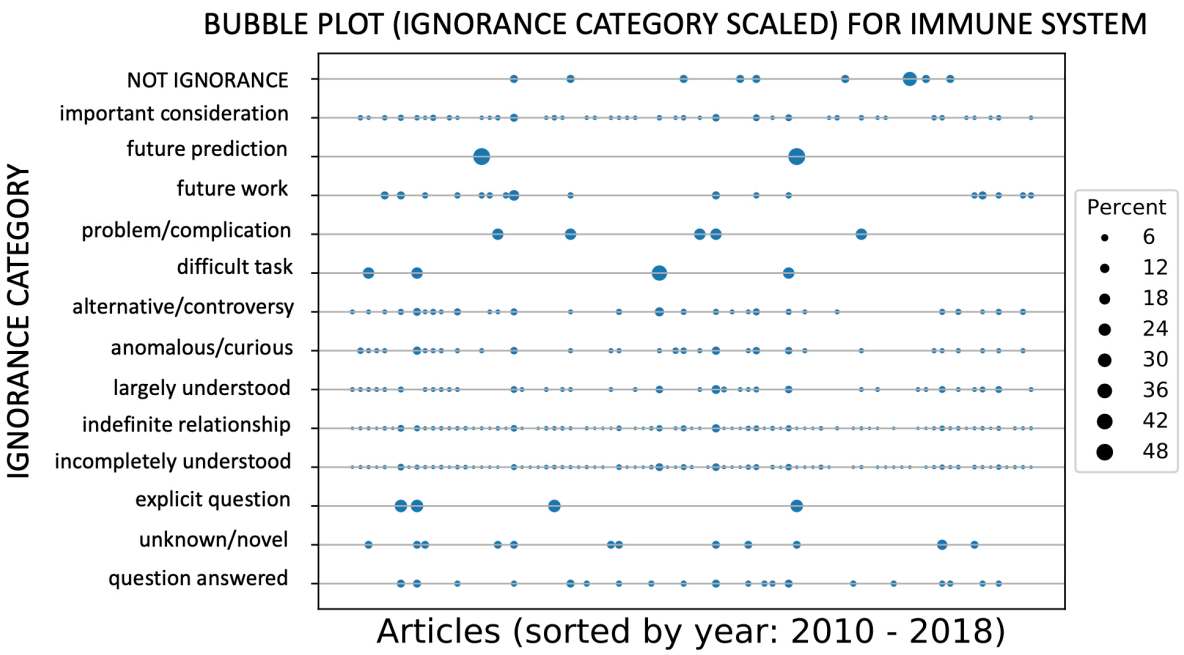


Figure 4.12: How ignorance changes over time: A bubble plot of vitamin D and immune system sentences (including non-ignorance sentences). The x-axis is the articles sorted by time. The y-axis is the ignorance categories. Each bubble represents the portion of sentences in each article in that ignorance category (scaled by the amount of total ignorance sentences in the category). For example, *future prediction* only appeared in two different articles and was basically split in half between both.

With the gene list connected to so many concepts (782), combining gene list coverage and ignorance enrichment helped prioritize concepts to explore (see Table 4.4). The highest covered OBO concept was PROTEIN CODING GENE (SO:0001217). In the top 25 most covered OBO concepts, all concepts were established information, currently studied, or not significant (see Figure 4.3). (None were emerging topics.) Note that some had no information from the literature-side, meaning no conclusions could be drawn based on the current information. In fact, all 18 concepts enriched in ignorance, were also enriched in all gene list sentences, including: PROTEIN CODING GENE, INNATE IMMUNE RESPONSE, GENE, IMMUNE RESPONSE, and BRAIN in the top 25. Note that there were some IMMUNE SYSTEM and BRAIN related concepts. These concepts were ripe for exploration in relation to the gene list to find knowledge goals that may bear on them.

Combining other canonical enrichment methods with ignorance enrichment also helped prioritize the many OBO concepts produced by our gene list (see Figure 4.13 focusing only on the

Table 4.4: Gene list coverage enrichment information: The top 25 OBO concepts sorted by the highest gene list coverage with enrichment information for all gene list sentences and for ignorance. NO INFORMATION means that the ontology term existed in PheKnowLator and was connected to our gene list, but there were no sentences that contained it on the literature side. *Statistically significant with FDR but not family-wise error.

OBO ID	OBO label	Gene Coverage	Enriched in all gene list sentences	Enriched in Ignorance
SO:0001217	protein coding gene	37	YES*	YES
GO:0005515	response to virus	23	NO INFORMATION	NO INFORMATION
CL:0000094	granulocyte	19	YES	NO
GO:0005886	plasma membrane	18	YES	NO
UBERON:0000178	blood	17	YES	NO
UBERON:0002371	bone marrow	15	YES	NO
CL:0000576	monocyte	14	YES	NO
GO:0005576	extracellular region	14	YES	NO
GO:0005615	extracellular space	13	YES	NO
CL:0000775	neutrophil	13	YES	NO
CHEBI:2504	aflatoxin B1	12	YES	NO
CHEBI:39867	kidney	11	NO INFORMATION	NO INFORMATION
GO:0045087	innate immune response	11	YES	YES*
GO:0070062	extracellular exosome	11	YES	NO
GO:0016021	bone marrow	10	NO INFORMATION	NO INFORMATION
SO:0000704	gene	9	YES	YES
GO:0005829	cytosol	9	YES	NO
CL:1001608	foreskin fibroblast	7	NO	NO
UBERON:0001332	prepuce of penis	7	YES*	NO
GO:0005737	cytoplasm	7	YES	NO
GO:0006955	immune response	7	YES	YES
CL:0000765	erythroblast	7	NO	NO
UBERON:0000955	brain	6	YES	YES
GO:0035580	lead(0)	6	NO INFORMATION	NO INFORMATION
CL:0000771	eosinophil	6	YES	NO

gene ontology). DAVID [278] found 42 of the 43 genes and mapped them to 159 GO concepts. 51 of those were enriched and 30 were contained in sentences found in the ignorance-base. Of those 30, 19 were contained in gene list statements and 11 had no information. Of those 19, 17 had at least one ignorance statement, and the concepts were mainly related to the immune system. (Two concepts, RESPONSE TO STRESS (GO:0006950) and MULTI-ORGANISM PROCESS (GO:0051704) had no ignorance statements.) The ignorance statements for the 17 concepts can be

explored to provide more information to the canonical enrichment methods and their respective knowledge-bases.

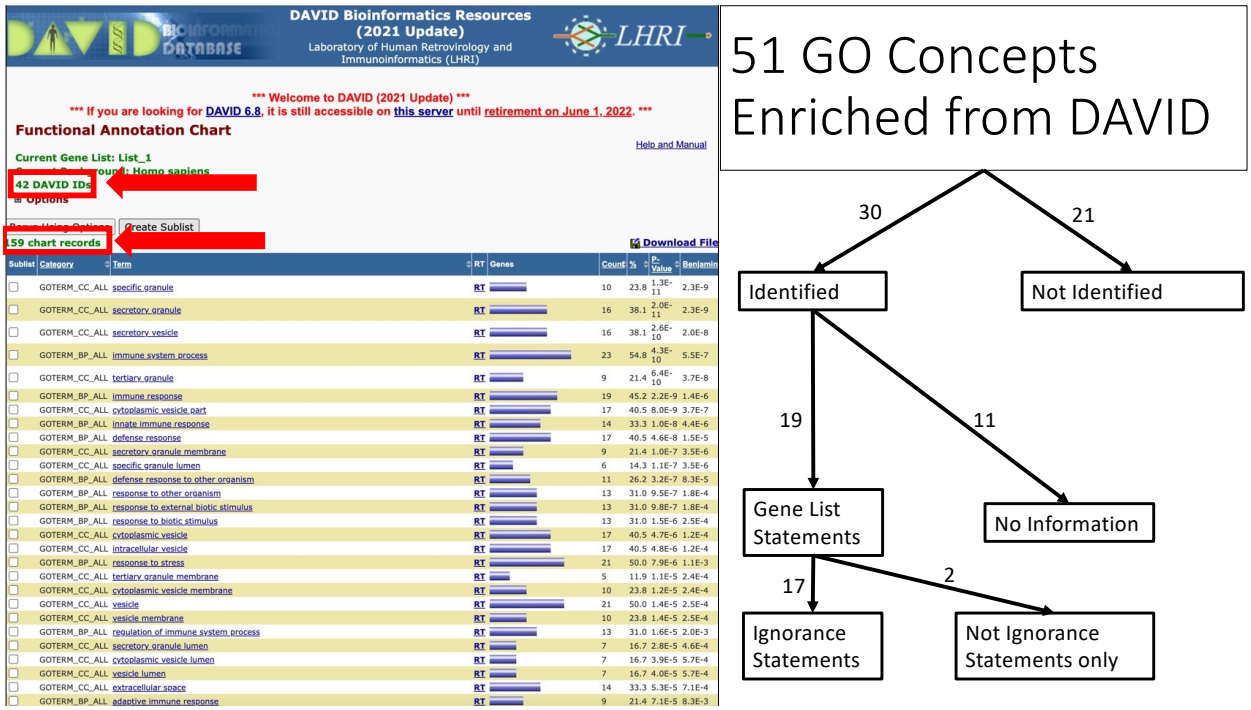


Figure 4.13: Enhancing canonical enrichment analysis using the ignorance-base: DAVID enrichment analysis for the gene ontology (GO) in relation to the ignorance-base. The DAVID initial analysis is on the left with 42 of the 43 genes found in DAVID mapping to 159 GO concepts. The right is a breakdown of where the 51 enriched GO concepts from DAVID fall within the ignorance-base.

To continue to add our ignorance lens to canonical methods, we compared our ignorance approach to the canonical approach. When comparing ignorance enrichment to DAVID, we found more ignorance in general compared to established information. 3,173 GO concepts were enriched in ignorance with 159 in DAVID (see Figure 4.14). Intriguingly, the overlap between the two analyses was small: 60. If we look at enrichment it was even smaller: only two concepts overlapped. Potentially this makes sense as we were enriching for the opposite things: ignorance vs. established knowledge. On the knowledge-side, most enriched concepts from DAVID pertained to the immune system, which was the main finding of Yadama *et al.* [6]. The two overlapping concepts, IMMUNE RESPONSE and INNATE IMMUNE RESPONSE, also pertain

to immunity and Yadama *et al.* [6] mention future work related to the immune system. These concepts were currently studied.

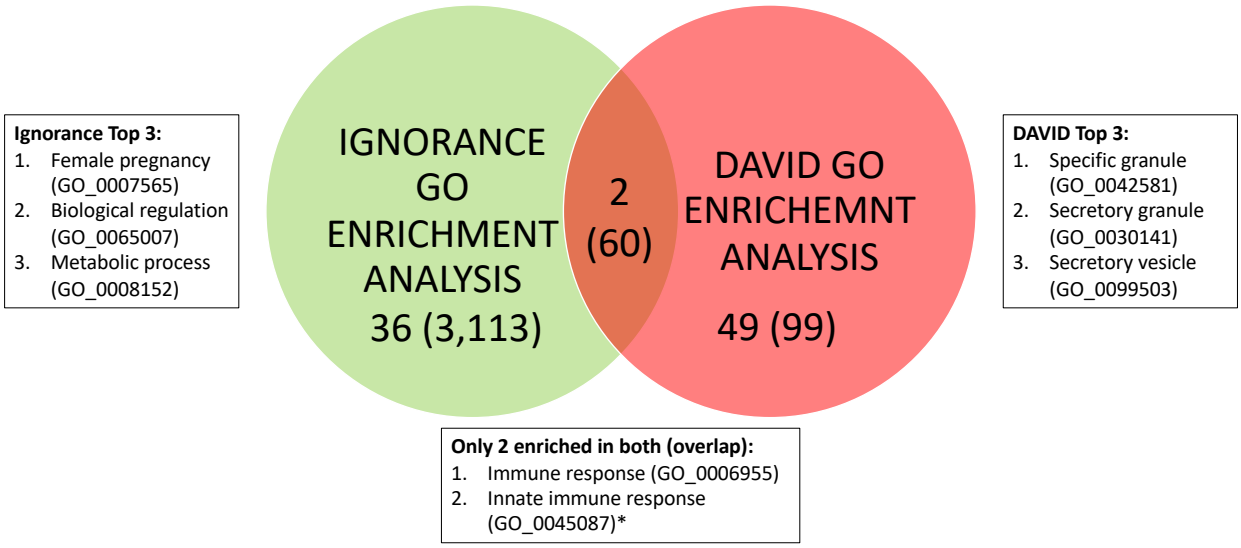


Figure 4.14: Comparison of DAVID and ignorance enrichment: A Venn diagram of gene ontology enrichment between the ignorance-base (green) and DAVID (pink). In parentheses are the total number of concepts found in each category without enrichment. Next to each bubble are the top three concepts for each enrichment method. The concepts in the middle are the overlap. *Statistically significant with FDR but not family-wise error.

Looking at the ignorance enriched concepts only, we achieved our goal of finding a new avenue to investigate that was not mentioned by the paper [6], namely the brain. Further, it also provided a different field to examine for answers, namely neuroscience. The top three ignorance-enriched GO concepts included FEMALE PREGNANCY, BIOLOGICAL REGULATION, and METABOLIC PROCESS (see Figure 4.14). Broadening this exploration beyond GO, there were 130 total ignorance enriched concepts for this gene list, including the 38 from GO. There were still some immune related concepts including (in order of enrichment): IMMUNE SYSTEM, IMMUNE RESPONSE, SKELETON OF MANUS (autoimmune), INTERLEUKIN-1 FAMILY MEMBER 7*, and INNATE IMMUNE RESPONSE*. (The * means that they were statistically significant with FDR but not family-wise error.) As mentioned above, we know there was still future work to be done on understanding the IMMUNE SYSTEM in relation to sPTB and VITAMIN D [6] and the ignorance approach provided specific ignorance

statements to explore for this future work. Even more striking though was the number of ignorance enriched concepts related to the BRAIN (12): BRAIN, BRAIN DEVELOPMENT, COGNITION, NERVOUS SYSTEM DEVELOPMENT, NEURON, LEARNING, SYNAPSE, NERVOUS SYSTEM, CENTRAL NERVOUS SYSTEM, NEUROTRANSMITTER*, NEURAL TUBE*, and NEUROGENESIS. There were no brain-related concepts both in the paper and in the DAVID analysis, signifying that this information was not yet established knowledge.

The brain may be a new emerging topic that the authors could examine more in relation to vitamin D and spontaneous preterm birth, and possibly look to the field of neuroscience for answers. We present some sentences (lowercase) from five papers to explore below (the biomedical concepts are capitalized and the ignorance lexical cues are underlined):

1. "this AREA OF the BRAIN, specifically the FRONTAL CORTEX, is important for LANGUAGE, MEMORY and higher order COGNITIVE functioning, including purposeful, goal-directed behaviours which are often referred to as executive functions.[4] the importance of adequate dha [docosahexaenoic acid] during this key period of BRAIN DEVELOPMENT is indicated in studies of preterm infants who are denied the full GESTATION period to accumulate DHA" (PMC4874207)

2. "discussion: we review relevant literature suggesting in utero inflammation can lead to PRETERM labor, while insufficient development of the GUT-BLOOD–BRAIN barriers could permit exposure to potential neurotoxins." (PMC3496584)

3. "a major Intake of DHA in the BRAIN happens in the last TRIMESTER of PREGNANCY; therefore, preterm infants are disadvantaged and have decreased BRAIN concentration of this vital lcpufa [long-chain polyunsaturated fatty acids]." (PMC3607807)

4. "at present, preterm infants have a limit of viability (50% survival rate) of around 23–24 weeks ga [gestational age] so post-NATAL nutrition will always be introduced during the second major phase of BRAIN growth, resulting in differences mainly in WHITE MATTER." (PMC3734354)

5. “the most likely explanation seems to be related to the timing of the nutrition event, since the infants were BORN at term rather than preterm when different developmental processes are occurring in the BRAIN.” (PMC3734354)

6. “the period between their PRETERM BIRTH and term BIRTH at 40 weeks, a time when the major BRAIN spurt is occurring), was spent ex utero in these infants; this exposure to environmental influences, at an early stage of BRAIN DEVELOPMENT, might be expected to increase their vulnerability to dietary effects.” (PMC3734354)

7. “the authors conclude by saying that reducing the energy deficit by improving early nutrition in preterms may improve the growth and maturation of the BRAIN.” (PMC3734354)

8. “iron status, more commonly assessed in PREGNANCY, is not only important in HEMATOPOESIS and NEUROLOGICAL and COGNITIVE DEVELOPMENT[9] but plays a crucial role in CARNITINE SYNTHESIS,[10] although CARNITINE precursors may be more important.[11]

   zinc is an important COFACTOR for more than 300 identified zinc metalloenzymes.[12] zinc insufficiency in late PREGNANCY disrupts NEURONAL REPLICATION and SYNAPTOGENSIS,[13] and maternal deficiency is associated with decreased dna, rna, and protein content of the f1 BRAIN.[14] zinc deficiency affects one in five world inhabitants.[14]ZINC supplementation reduces the risk of PRETERM BIRTH, though not sga [small for gestational age].[14]

   VITAMIN D deficiency is under investigation for its role in protection against dm [diabetes mellitus], cv [cardiovascular], some ca [cancers], osteoporosis, and optimization of IMMUNE function.[15] VITAMIN D might be an important mediator in GUT HOMEOSTASIS and in signaling between microbiota and host.[16] the INTESTINAL microbiome in both newborns and LACTATING mothers influences infant and childhood FOOD allergy and eczema.” (PMC4268639)

(Note that not all biomedical concepts were recognized by the biomedical concept classifiers. Also, the numbers in the sentences represent citations, which were superscript in the original article but were flattened for processing.)

In the paper, Yadama *et al.* [6] focused on the mother's immune system, but it is possible that this brain connection is instead focused on the effects of the spontaneous preterm birth on the offspring. Thus, a potential knowledge goal for the authors based on our analysis was to explore the connections between maternal VITAMIN D levels and spontaneous preterm birth through the maternal IMMUNE SYSTEM and the effects on the BRAIN DEVELOPMENT of the offspring. Exploring these connections would impact mothers everywhere. The ignorance-base provided a novel avenue (and field), BRAIN DEVELOPMENT (neuroscience), along with specific knowledge goal statements that the authors can explore for future work based on their initial gene list. Our exploration by experimental results method contextualized experimental results in the ignorance landscape, providing multiple avenues for future research, the immune system and the brain.

Discussion

Focusing on ignorance statements through our ignorance-base and exploration methods led to new research avenues that could help accelerate translational research. Further, the ignorance-base was not only a literature search engine similar to Lahav *et al.* [17], but also provided insights, summaries, and visualizations based on a topic and experimental results. The ignorance-base helped find areas of research with lots of questions ripe for exploration with the knowledge-goal focus and grounding in the OBOs. The ignorance-base predicted areas of research that were currently studied and an emerging topic with a corresponding field that may prove fruitful to help find answers. The exploration by topic method showed that vitamin D may play an important role in the immune system, respiratory system, and brain development (see Figures 4.8- 4.12 and Table 4.3). To corroborate these findings post 2018 when our corpus ended, recent review articles [115–119] included these areas as future work. These review articles required many hours of reading and synthesizing the literature article by article, whereas the

ignorance method automatically offered not only articles, but also specific sentences that discuss knowledge goals for future work. For example, “it has an important role in BONE HOMEOSTASIS, BRAIN DEVELOPMENT and MODULATION OF the IMMUNE SYSTEM and yet the impact of ANTENATAL VITAMIN D deficiency on infant outcomes is poorly understood" (PMC4072587) [286] showed that further research on the impact of vitamin D on infant outcomes was needed. The context of the sentence was also important: it was from a 2014 study in Rural Vietnam and appeared in the abstract objective section. Because we can sort our ignorance statements by time and section, we showed that since 2014, more research has been conducted on this topic [117] as found in our ignorance-base. Further, we can track how research questions emerge using our ignorance taxonomy (see Figure 4.12). Ideally, we would create an ignorance-base over the entire body of scientific literature to truly help researchers, students, funders, and publishers capture the landscape of our collective scientific ignorance and understand how research questions evolve over time. This smaller-scale effort demonstrated that this was possible.

We further demonstrated that the ignorance-base and exploration by experimental results method can find an emerging topic (see Figures 4.4, 4.13, 4.14, and Table 4.4). Ignorance enrichment of the 43 genes in common between vitamin D and spontaneous preterm birth (sPTB) [6] found many concepts that relate to the brain and some that relate to the immune system (as found by [6]). This suggested that brain development could be an emerging topic in relation to vitamin D and sPTB. For example, consider the ignorance statement: “discussion: we review relevant literature suggesting in utero inflammation can lead to PRETERM labor, while insufficient development of the GUT-BLOOD–BRAIN barriers could permit exposure to potential neurotoxins.” (PMC3496584) [287]. This sentence ties all the relevant concepts together: “vitamin D may be causing in utero inflammation which is leading to preterm labor, and due to the preterm labor there is insufficient development of the gut-blood-brain barrier leading to potential neurotoxins for the fetus". This article was not cited by Yadama et. al. [6], but may be a new knowledge area that needs to be explored further. Lastly, researchers could look to the field of

neuroscience to help find relevant information to some of these knowledge goals. Here are some potential questions that could be explored:

1. What is the association between development of the gut-blood-brain barrier and whole-body inflammation and neuroinflammation in the context of fetal development?
2. How do the 43 genes relate to offspring brain development? Are any of them specifically related to offspring brain function?
3. What are the effects of vitamin D on lifecourse brain development generally?
4. How does spontaneous preterm birth impact offspring brain development compared to those born at term? Are there any remedies for said effect? Does nutrition play a role?
5. How does the gestational timing of nutrition and supplement exposure affect offspring brain development? What role do iron and zinc play in brain development?

Looking at the literature further showed that vitamin D and brain development may in fact be an emerging topic since 2018, the last year of our data. The connection between vitamin D and brain development has only recently been studied extensively. Looking for recent papers on "vitamin D", "brain development", and "spontaneous preterm birth" in the literature (Google Scholar search on 9/13/2022), many review articles appeared (*e.g.*, [115–119]) that discussed the impact of vitamin D on maternal and fetal health (see Figure 3 in [115] and our adapted Figure 4.1). All of these studies agreed that there was a link between vitamin D, the immune system, and sPTB, and acknowledge at least one link between vitamin D and brain development. A 2022 review article stated that "recently, extensive scientific literature has been published determining the role of vitamin D in brain development" [115]. Note that these articles contained mentions of controversies and other types of ignorance statements, which also point to other areas of investigation. There appears to be room for exploration around the connections between vitamin D, sPTB, and brain development. Understanding the ignorance-context around a set of

genes in combination with the knowledge-context can help push the boundaries of our current understanding.

In general, we showed that ignorance-bases and knowledge-bases can enhance and complement each other. The ignorance-base itself was built upon a knowledge-base, PheKnowLator [276, 277]. We also utilized DAVID [278] as a comparison knowledge-base (see Figures 4.13 and 4.14) as well as other canonical methods (gene list coverage) to help prioritize the most relevant biomedical concepts (see Table 4.4). These analyses were made possible by grounding our ignorance-base in the OBOs, which allowed us to connect our ignorance-base to other knowledgebases. At the same time, we did not use any of these methods to their fullest potential. First, only six ontologies overlapped between the biomedical classifiers (Chapter 2) and PheKnowLator[276, 277], which limited the expansion of relevant concepts. Second, the method to create the OBO concept lists from both the vitamin D topic and the gene list was not very sophisticated (using only one step via the relations ontology). Finally, we did not use the knowledge-bases to determine if any ignorance statements have been answered. All of these limitations could be addressed in future work. Many other knowledge-bases and methods exist that can also be explored in relation to the ignorance-base.

The goal of this work was to demonstrate the feasibility and power of the ignorance methods, and we recognize that more improvements can be made. Our ignorance-base was created from both automatic ignorance and OBO classifiers run over 1,643 articles in the prenatal nutrition literature. Any automation of this kind added errors and all classification tasks can be improved upon to minimize it. For the ignorance classifiers, other parameter tunings and other algorithms, such as PubMedBERT [243], may yield improved results. For our biomedical concept classifiers, CRAFT [166, 170] was a corpus of mouse articles, not prenatal nutrition, and it only included ten ontologies. Applying biomedical classifiers with more similar training data and more ontologies (*e.g.*, MONDO disease ontology and the phenotype ontology) would be beneficial (*e.g.*, PubTator [288]), although all of them have their pros and cons. We ran these classifiers over only 1,643 prenatal nutrition articles. The scale of the ignorance-base was small; ideally we would

create an ignorance-base that included all articles (or at least PMCOA to start). We focused only on the prenatal nutrition literature, and future work can determine if the ignorance taxonomy and methods generalize outside of it. In terms of exploration methods, concept enrichment provided more fruitful concepts (see Figures 4.10 and 4.11) than concept frequency (see Figure 4.9), and there are other methods to explore including co-occurrence terms. The creation of a tool (similar to [17]) incorporating more data analyses and visualizations techniques into a user-interface that allows researchers to interact with the system could make the ignorance-base easier to adapt to new environments. Even with all these limitations, the current ignorance-base showed its power to find new research avenues to explore, providing insights, summaries, and visualizations beyond prior work [17]. We have just barely scratched the surface of what it can do. Collaborating with experts on vitamin D, delving into the topics introduced here, creating new methods, exploring other topics, and contextualizing other experimental results are obvious extensions of this work.

We demonstrated that a focus on ignorance statements through our ignorance-base and exploration methods can lead students, researchers, funders, and publishers to research avenues that are currently being studied or are emerging topics. Research begins from a foundation of established knowledge, and many knowledge-bases and ontologies exist to provide that. However, research continues through a process of posing questions and creating hypotheses to analyze and explore what is not yet understood. To facilitate that, we presented the first ignorance-base based on knowledge goals and OBOs, along with two new exploration methods that provided insights, summaries, and visualizations of statements of unknowns, controversies, and difficulties ripe for exploration in future work. Just as the literature contains both knowledge and ignorance, so too can both knowledge-bases and ignorance-bases help researchers navigate the literature to find the next important questions or knowledge gaps.

Overall, the ultimate goal of this work was to help students and researchers explore the literature through the questions and known unknowns. The scientific endeavor rests on our continuous ability to ask questions and push research farther as we learn more knowledge. The two methods presented here, exploration by topic and exploration by experimental results,

provided two ways to help students, researchers, funders, and publishers explore the state of our scientific ignorance.

Conclusion

The ultimate goal was to create an ignorance-base and exploration methods to help students, researchers, funders, and publishers find the next important questions or knowledge gaps. This can possibly help accelerate translational research over purely manual methods. The exploration by topic method not only found new avenues for exploration for researchers interested in vitamin D using our novel method of ignorance enrichment (the immune system, respiratory system, and brain development), but also elucidated how questions were asked and how that changed over time using our novel method of ignorance-category enrichment. Our exploration by experimental results method found an emerging topic (brain development) with specific knowledge goal statements to pursue that bear on a sPTB and vitamin D gene list. Further, the findings suggested a field (neuroscience) to look to for answers. These questions (and subsequent answers) have high potential to positively impact the health of pregnant women and their offspring globally. The importance of questions and knowledge goals in research is well established, and our ignorance-base and exploration methods brings these to the forefront to help researchers explore a topic and experimental results in the context of our collective scientific ignorance. The scientific endeavor rests on our continuous ability to ask questions and push research farther as we learn more knowledge. To paraphrase Confucius,“Real knowledge is to know the extent of one’s ignorance" (Analects 2:17).

Summary

Many knowledge-bases exist to capture what is known and they are used to find information and contextualize experimental results. Known unknowns are recognized as important and studied under different names and focuses, including a focus on knowledge goals. However, no prior work has created a knowledge-base to capture the nuance of known unknowns as knowledge goals, connected them to ontologies for integration with other knowledge-bases, and provided summaries and visualizations of the outputs to help researchers find the most pertinent

questions. Our goal was to create such knowledge-base to capture known unknowns from the literature to contextualize topics and experimental results in our collective scientific ignorance. We built an ignorance-base to explore it by topic to help researchers find the most pertinent questions related to a topic (*e.g.*, help graduate students find thesis topics), and by experimental results to contextualize them in the known unknowns. Based on our exploration by topic methods, we found multiple areas of research currently studied (the immune system, brain development, and the respiratory system) with lots of questions (ignorance statements) that are ripe for exploration for researchers interested in vitamin D. Based on our exploration by experimental results method applied to a vitamin D and spontaneous preterm birth gene list [6], we found both a currently studied research area (the immune system - the researchers' finding) and an emerging topic (the brain - not mentioned by the researchers) that are ripe for exploration in future research. We not only provided literature search results to find articles and sentences of interest, but also insights, summaries, and visualizations of the relevant data. The ultimate goal was that students, researchers, funders, and publishers can use our ignorance-base and methods to help find the next important questions or knowledge gaps.

Contributions And Acknowledgements

This work was submitted to the Journal of Bioinformatics with the preprint posted on bioRxiv [175].

- **Boguslav MR**, Salem NM, White EK, Sullivan KJ, Araki SP, Bada M, *et al.* Creating an Ignorance-Base: Exploring Known Unknowns in the Scientific Literature. 2022;:2022.12.08.519634. Available From: https://doi.org/10.1101/2022.12.08.519634

I would like to thank the other authors on the preprint including Nourah M. Salem, Elizabeth K. White, Katherine J. Sullivan, Stephanie P. Araki, Michael Bada, Teri L. Hernandez, Sonia M. Leach, and Lawrence E. Hunter. MRB, SML, and LEH conceived of the ideas. MRB created the ignorance-base and exploration methods. TLH was our prenatal nutrition expert.

TLH, SML, and LEH supervised the work. All authors helped write and reviewed the manuscript. I would like to acknowledge the BioFrontiers Computing Core for computing resources and support, especially Jonathon Demasi. I would like to thank Harrison Pielke-Lombardo for his updates to the Knowtator tool; William A. Baumgartner Jr. for his help with the literature and taxonomy work; and Tiffany J. Callahan for many discussions about this work.

CHAPTER V

DISCUSSION

Overview

Research begins with a question. It progresses through accumulating knowledge such that a previously unexplored subject (an unknown unknown) becomes an active research area exploring the questions (known unknowns), until a body of established facts emerge (known knowns). Many knowledge-bases exist to capture the known knowns and are used to explore topics or experimental results [4]. Equally as important are the questions surrounding a topic or a set of experimental results to help students, researchers, funders, and publishers better understand the state of our known unknowns [1–3]. But how do researchers find the most pertinent questions? These skills are developed in graduate school, where students learn both the questions in their field and how they are asked so that they can ask a question and provide some solutions for it as their thesis project [7–16]. Many of these questions or known unknowns are discussed in the scientific literature as statements about knowledge that does not exist yet, including goals for desired knowledge, statements about uncertainties in interpretation of results, discussions of controversies, and many others; collectively we term them **statements of ignorance**. A knowledge-base or an **ignorance-base** that provided insights, summaries, and visualizations of such statements could be useful to a wide variety of scientists ranging from graduate students looking for thesis projects as mentioned (*e.g.*, [12]) to funding agencies tracking emerging research areas (*e.g.*, [33]). It could help facilitate interdisciplinary interactions amongst researchers by finding questions from another field that bear on a topic or a set of experimental results (*e.g.*, [34]). It could also help track the evolution of research questions over time as a longitudinal analysis (*e.g.*, [35]). Further, finding pertinent questions automatically would allow us to query existing databases for information (*e.g.*, [36]). However, to the best of our knowledge, no prior work has created a knowledge-base or ignorance-base to capture the nuance of known unknowns as knowledge goals, connected them to ontologies for integration with other knowledge-bases, and provided summaries and visualizations of the outputs to help researchers

find the most pertinent questions, until now. We presented the necessary methods and tools to create the ignorance-base and explore it by a topic or a set of experimental results to provide insights into future research avenues and questions.

While all these ideas and methods are applicable across biomedical research, we chose to create the ignorance-base over the prenatal nutrition literature. Prenatal nutrition is understudied and serves to benefit from the identification of questions that are well studied in other fields due to the ethical and legal considerations and complexities in studying pregnant mothers and fetuses. Also the prenatal nutrition field is a good case study for these ideas because it contains a diverse literature with all types of studies from all over the world, meaning there is a higher potential for generalizability to other fields. Our goal was to help students, researchers, funders, and publishers better understand the state of our collective scientific ignorance (known unknowns) (Chapter 4) in order to help accelerate translational research through the characterization of known unknowns based on their respective goals for scientific knowledge (Chapter 3) and understanding their biomedical subjects (Chapter 2).

All of this work has resulted in state-of-the-art biomedical concept recognition models based on machine translation (Chapter 2), a taxonomy of ignorance with 13 knowledge goal categories, annotation guidelines, an ignorance corpus of 91 articles, high performing ignorance classifiers on the sentence- and word-levels (Chapter 3), and an ignorance-base with two types of exploration methods by topic and experimental results (Chapter 4). We showed that machine translation was a salient avenue to explore to improve concept recognition. The CRF and BioBERT paired with OpenNMT performed the best for concept recognition. The ignorance corpus and ignorance-base suggested that ignorance statements were rampant throughout the prenatal nutrition literature and employed a wide vocabulary. Further, we provided high performing ignorance classifiers that showed that automatically identifying ignorance statements was feasible. Finally, we showed the power of combining ignorance and biomedical concept classifiers to create the ignorance-base and explore it by a topic and a set of experimental results, providing insights into future research areas. We provided a researcher with specific knowledge

goal statements to help narrow in on a research topic (immune system, brain development, and respiratory system were enriched in vitamin D ignorance statements). Focusing on ignorance statements can help researchers find pertinent questions to pursue in future work. We also provided the ignorance context for a gene list [6], finding the concept immune system, as found by the original authors, as well as a novel putative relationship with the concept brain development. Both these concepts with their corresponding ignorance statements were ripe for exploration in future work.

The ultimate goal was that students, researchers, funders, and publishers can use this work to help find the next important questions or knowledge gaps. The importance of questions and knowledge goals in research is well established, and this work brings these to the forefront. The scientific endeavor rests on our continuous ability to ask questions and push research farther as we learn more knowledge.

Strengths

Overall, the main strengths of this work were the diversity, quality, and amount of data that allowed us to draw reasonable and actionable conclusions to explore in future work. All of our data for the ignorance work (Chapter 3 and 4) was from the very diverse prenatal nutrition literature. This literature included a wide range of study types from animal studies to retrospective studies (based on reviewing the articles), and had publications from all over the world (based on reviewing the articles). Thus, even though our focus was on one field, we believe that our results will generalize beyond prenatal nutrition because of these factors.

This work can further generalize because it was grounded in ontologies. The OBOs [155] are a biomedical research community effort to standardize the biomedical vocabulary for different types of biomedical entities. They are used in NLP systems and in the creation of knowledge-bases. Since we identified and normalized the biomedical concepts to the OBOs in the ignorance-base, we can connect our work to other NLP systems and knowledge-bases to potentially look for knowledge that helps answer or qualifies ignorance statements. Further, we also consulted an ontologist (Mike Bada) to help create the ignorance taxonomy to ensure our

formalization of ignorance was grounded in the same principles as other ontologies. Thus, our ontological grounding was a strength.

With this ontological grounding for the ignorance taxonomy, we created generalizable annotation guidelines that yielded a robust and high quality ignorance corpus and high performing classifiers (Chapter 3). We conducted two annotation tasks with five different annotators that maintained inter-annotator agreements above 80%. Further, we were able to annotate enough data for both a training set and a separate testing set for final evaluation on unseen data. These factors led to more trust in the high performing classifiers trained on this corpus. Similarly, the CRAFT corpus [166, 168–170] for the biomedical concept recognition task had the same setup and strengths (Chapter 2). Thus, both the quality and the amount of data in both corpora seemed to be sufficient to create high quality and high performing ignorance and biomedical concept classifiers.

With these high performing classifiers, applying them to the full set of prenatal nutrition PMCOA articles (1,643) for the ignorance-base seemed to be enough data to provide insights and discover novel findings for a topic of and a gene list related to vitamin D. The prenatal nutrition field is understudied and serves to benefit from the identification of questions that are well studied in other fields [110–113]. Our prenatal nutrition expert (Teri L. Hernandez) was also a strength of this work to provide actionable insights to the field. In consultation with her, our results provided possible new avenues and questions to explore for researchers looking for a research topic in vitamin D and researchers wanting to contextualize their vitamin D and spontaneous preterm birth gene list in the broader ignorance landscape. For the topic, our results matched the current literature showing that the immune system, brain development, and the respiratory system were all currently being studied in relation to vitamin D. Further, our approach provided specific knowledge goals for the researchers to pursue for each concept. For the gene list connecting vitamin D and spontaneous preterm birth, the results identified both the same finding as the researchers with a focus on the questions still surrounding it (the immune system), and a new avenue to study which was not mentioned in the paper, but was verified by the literature as an emerging topic (brain development). Based on the brain development ignorance statements, we

created an hypothesis of the connection between all concepts. We also provided five example questions that could be explored further based on these findings. For answers to these questions, researchers could look to the field of neuroscience, as it is the implied field of brain development. Thus, our ignorance-base has provided meaningful insights and novel findings based only on 1,643 articles. Even though our data was small, our methods, tools, and results seem to be generalizable.

Limitations

Even with all the strengths mentioned above, the biggest limitation surrounding this work was generalizability. Our focus on one field and the relatively small amount of data, bring the question of generalizability into focus. It is unclear if this work will generalize beyond the field of prenatal nutrition. Further, the ignorance-base was built on automatic classifiers for both the ignorance statements and the biomedical concepts, adding automation error. For the biomedical concept classifiers, there were especially errors because they were trained on a corpus focused on the mouse and not our field of interest. Both generalizability to another field and of our classifiers remain limitations.

In questioning generalizability, there are cases in which the methods and ideas presented here may not work. For example, if there is no ignorance to capture or it is discussed very differently then what we found here. Theoretical mathematics and physics mainly publish proofs and once it is proven there is no ignorance. Every field will differ in whether and how they discuss ignorance. In terms of how ignorance is discussed, our definition of ignorance and our task only focused on lexical cues as markers of statements of ignorance, but there could be other linguistic features. During our annotation task, we found at least one sentence that both annotators agreed was a statement of ignorance but could not agree on a lexical cue that signified it as such. Thus, the sentence was not included based on our task definition (see Chapter 3). Much prior work (*e.g.*, [41]) used lexical cues or keywords to help identify their specific phenomenon, including Lahav *et al.*[17]. They also tried to augment beyond just lexical cues by providing sentences to their annotators with none of their keywords [17]. During automation of the task, they tested keyword-based models and found good recall but not precision, while other non-keyword-based

models performed better [17]. They did not discuss in detail what the algorithms paid attention to for classification and the role that keywords may play in them [17]. More work needs to be done to understand what defines an ignorance statement or similar statements beyond just lexical cues, and this may differ depending on the field of interest.

Also, our ignorance taxonomy was much more granular than previous work. Not all categories may exist in other fields and we may be missing categories. For example, the ignorance categories *important consideration* or *indefinite relationship* may be minimal in theoretical mathematics where neither urgent calls to action nor associations tend to exist. At the same time, we may be missing an ignorance category or require more granularity. For example, for engineering a knowledge goal could be to build something with the ignorance category of implementation or something similar. In terms of situations requiring more granularity, in earlier versions of our ignorance taxonomy we had two separate categories for *indications of alternative options* and *indications of controversy of research* to separate disagreements from lists of options. During our annotation task, we could not always distinguish between the two more granular categories and we recognized that the knowledge goal was the same: to determine the correct option amongst the current options or find a better one. We thus saw reasons to combine the categories, but that may not have been the best choice. More work is necessary to determine the optimal ignorance taxonomy.

More work is also necessary to ensure we capture all the biomedical concepts of interest so that we can provide the ignorance context for any topic or experimental results. To fully understand the ignorance statements, we also captured their biomedical subjects as concepts so that we could explore the ignorance statements based on their subjects. If a biomedical subject or concept did not exist in our ignorance-base then we could not provide the ignorance context. Thus, for the ignorance-base, the exploration methods failed for biomedical concepts not captured by our biomedical concept classifiers. The CRAFT corpus only contained ten ontologies and so we only had classifiers for those. Very obvious missing ontologies were for diseases and phenotypes such as the MONDO disease ontology and the phenotype ontology, respectively.

Searching for terms such as preeclampsia (MONDO:0005081) was not possible. Due to this, many terms were not captured and thus could not be queried for in the explorations. This also included terms of interest based on our exploration examples such as spontaneous preterm birth (NCIT:C112864, EFO:0006917, or GSSO:009645), which was one focus of our gene list [6]. Future work needs to include more ontologies, more data, and better biomedical concept classifiers to help with this problem.

Even if a concept was in one of the ten ontologies from CRAFT, we were sometimes unable to draw conclusions about it because our biomedical concept classifiers may have missed it or it may not have been in our corpus of 1,643 articles. The concept could have been misidentified, altogether not identified, or the concept truly was not in the corpus. Either way, we could not draw any conclusions in relation to ignorance or not about that concept. Recall the concepts with no information about them in relation to gene list coverage (see Chapter 4 Table 4.4). The concepts existed in PheKnowLator [276, 277], but were not connected to any sentences from the literature corpus. Not only were we missing concepts based on missing ontologies, but also based on not having examples of them in the literature. Thus, both the possible explorations and results were limited based on the recognized concepts from the biomedical concept classifiers.

These limitations cast doubt on our exploration results in terms of both the ignorance statements and biomedical concepts that might be missing or possibly incorrect. For exploration by topic, our vitamin D query had no results for two of the four search terms. We do not know what ignorance statements exist about D3 VITAMINS (ChEBI:73558) and CALCIOL/VITAMIN D3 (ChEBI:28940). In terms of the results, it is almost impossible to know what biomedical concepts could be missing from our list of other frequent and enriched concepts in ignorance statements. These limitations also apply to our exploration by experimental results for our gene list[6]. It remains an open question of how best to map genes to ontology terms beyond the Gene Ontology. We provide a simple method using the relations ontology and PheKnowLator [276, 277], but there are more sophisticated methods to try including utilizing the ontology hierarchies to create expanded biomedical concept lists. In terms of the exploration results, we already

mentioned the issues with no information for certain concepts. Future work is necessary to confirm all of our exploration results.

Lastly, ideally we would create an ignorance-base over all the literature which relies on the scalability and usability of this work. Much of the computation was performed on a supercomputer with GPUs, which required lots of time and memory. This may not be feasible to run over all of PubMed, let alone all the literature. Further, there is currently no user-interface into the ignorance-base for students, researchers, funders, and publishers to explore. This work was distributed across several GitHub repositories as python and bash scripts. Both of these need to be addressed in future work.

Even with all these limitations, the goal of all this work was to spark ideas for future research and not necessarily provide the best pertinent question. We relied on the peer review process to ensure good science. We laid no claims to whether an ignorance statement was solved (closed) or unsolved (open). We rely on researchers to evaluate the results and continue to do more research. Even so, we hope that this pilot study demonstrated the importance and power of an ignorance-base to help researchers find the most pertinent questions based on identifying, categorizing, classifying, and exploring ignorance statements and their entailed knowledge goals.

Future Directions

The goal of my dissertation was to demonstrate the feasibility of identifying, categorizing, classifying, and exploring known unknowns in the scientific literature as driven by their entailed goal for scientific knowledge. Thus, there is a wealth of future work to conduct to go beyond just feasibility.

Biomedical Concept Recognition

For biomedical concept recognition, my main question is why does machine translation work for concept normalization? There should not be semantics in the OBO IDs and yet we found some. We conducted some experiments to both break the supposed semantics and to add in more, which both hurt and helped different OBOs. Future work could try to find a different mapping

from a concept to an ID that boosts performance for concept normalization using machine translation.

Our concept normalization work not only found some semantics in the OBO IDs, but also created new IDs that were non-existent in the ontologies. This was due to the character-by-character nature of the machine translation to allow for generalizability. These were automatic errors in our work, but future work could determine how to address these non-existent IDs. Further, they may help provide insights into the semantics the algorithm seems to detect.

For all our concept recognition work in both span detection and concept normalization, more ontologies can be added such as the MONDO disease ontology and the phenotype ontology. Also, more algorithmic development could be done in future work. All the algorithms can continue to be tuned with more hyperparameter searches and more algorithms can be tested such as PubMedBERT [243] for span detection and other machine translation algorithms [193] for concept normalization.

Ignorance Task

The main future work is to test the generalizability of the ignorance taxonomy and classifiers. Future work could first apply this work to another field, such as the database of Covid-19 literature, LitCovid [275]. To truly ensure generalizability to another corpus of literature, the automatic annotations need to be checked by experts to determine their reliability. If the annotations do not suffice, then we may not be capturing all types of ignorance or lexical cues. More annotation work can be conducted on a different field and more algorithms can be tested. Further, lexical cues may not be the only way to identify ignorance statements and so an exploration of other indicators of ignorance statements may prove fruitful.

Ignorance-Base

Ideally, we would create and maintain an Ignorance-base over the entire literature. This hinges on generalizability as discussed above and a user-interface for students, researchers, funders, and publishers to interact with ignorance statements. For the tool, more summarizing techniques and visualizations are necessary to help researchers digest the vast amounts of

information in the ignorance-base. I provided some visualizations for the researchers looking for a research topic and researchers with a gene list, but there is room for improvement. For example, the bubble plots we created to show how ignorance changed over time were somewhat difficult to read and interpret, especially with large amounts of ignorance statements. (We showed the bubble plot of vitamin D and immune system ignorance statements because just vitamin D ignorance statements was hard to read.) Further, we want to create new visualizations that would help funders and publishers especially determine the next big questions to fund or publish. New visualizations specific to them would be useful. Exploring new summarizing techniques and visualizations for the ignorance-base to incorporate into a tool and creating an ignorance-base over all of PubMed, would make our work more accessible and the insights more generalizable. This could come to fruition by combining our work with the COVID-19 search engine for challenges and directions [17].

With regards to insights, future work could test the insights we discovered in this work for the researchers looking for pertinent questions and researchers with a gene list. For the researchers looking for pertinent questions, are the ignorance statements with the immune system, brain development, and respiratory system interesting avenues to pursue for a research topic? For the researchers with a gene list, how is brain development related to the genes and does neuroscience have some answers? I hope that this work can help real researchers, especially students narrow in on research topics and real researchers contextualize their experimental results in the scientific ignorance landscape providing meaningful and testable insights for both.

Overall

All of this work could be the first step to creating a **question ecosystem** for all to explore to both understand the current state of our knowledge boundary and to push it. This ecosystem would entail an ignorance-base over all the scientific literature with tools for different audiences. To distinguish between open (unanswered) and closed (answered) questions, ideally we would also connect our current knowledge to the question ecosystem. We connected a knowledge-base, PheKnowLator [276, 277], to our ignorance-base but we did not utilize its knowledge to its fullest

potential (see [4] for the many applications of knowledge-bases). All of these tools could help people better understand science and the history of knowledge.

The current mistrust in science partially stems from people misunderstanding the complicated and entwined nature of ignorance and knowledge [289, 290]. To understand science is to understand the role of the question, which was the focus of this dissertation. I believe a refocus back to the question and history of knowledge could help bridge this gap. We need to educate our students at the minimum to understand science communications and at the maximum to become researchers themselves if they so desire [291]. We need to reexamine how scientists build trust with diverse audiences (*e.g.*, [292]). Scientists need to interact with diverse audiences more and help create digestible and accurate scientific communications. The stakeholders in all three areas need to work together to make changes to gain back the trust in science from diverse audiences [289] and I think my ignorance work can help.

There could be ignorance-base tools for education. An ignorance-base could provide teachers with a tool to teach the history of knowledge and foster curiosity. Students could visualize the history of questions similar to knowledge through timelines and networks. Citation networks would also be valuable to see who is generating the questions and knowledge. It could also help students with research projects, teaching them how researchers find pertinent questions. The ignorance-base is meant to spark ideas, curiosities, and more questions. If we can educate people on the research process, it could potentially better help them digest science in the future or pursue research careers if they so desire. Before implementing these ideas, I want to determine the role questions and ignorance play in education. For example, quantifying both the amount and types of ignorance statements in course materials. The results of this analysis would provide valuable insight into the types of ignorance-base tools needed for education to foster curiosity and teach the history of questions.

This dissertation work focused on ignorance-base tools for researchers to help them find pertinent questions to ask in order to conduct experiments to answer them. My dissertation work demonstrated that an ignorance-base can help find novel avenues to explore that have lots of

questions for a specific field. Even though we are proposing one large ignorance-base over all the literature, field specific ignorance-bases may still prove useful within each specific field. The Ignorance-base section above provides specific future work to create a larger research ignorance-base, including more summarizing techniques and visualizations.

There was a wealth of data in just our small ignorance-base, the challenge is how to present it, summarize it, and visualize it. The main summary of an article is its abstract. Would a "question summary" prove useful? Could ignorance statements be used to help summarize articles? For a one sentence summary, taking the first occurrence of statements with an *indication of answered research question* ignorance category seems promising (see Figure 5.1). Creating summaries using different ignorance categories and from different sections could prove useful and interesting. Thus, there is much work to be done focused on the question ecosystem.

Article Title	First *indication of an answered research question* Sentence
Dietary intake of fish, omega-3, omega-6 polyunsaturated fatty acids and vitamin D and the prevalence of psychotic-like symptoms in a cohort of 33 000 women from the general population (PMC2889879)	Our aim was to evaluate the association between the intake of different fish species, PUFA and vitamin D and the prevalence of psychotic-like symptoms in a population-based study among Swedish women
Season of Birth and Risk for Adult Onset Glioma (PMC2898025)	This paper reviews the plausibility of evidence focusing on the seasonal interrelation of farming, allergies, viruses, vitamin D, diet, birth weight, and handedness
Effect on the incidence of pneumonia of vitamin D supplementation by quarterly bolus dose to infants in Kabul: a randomised controlled superiority trial (PMC3348565)	Our aim was to assess whether oral supplementation of vitamin D3(cholecalciferol) will reduce the incidence and severity of pneumonia in a high-risk infant population
High Prevalence of Vitamin D Deficiency in Pregnant Women: A National Cross-Sectional Survey (PMC3427250)	The aim of this study was to estimate the prevalence of vitamin D deficiency among Belgian pregnant women and to assess the determinants of vitamin D status in the first and third trimester of pregnancy
Is Pet Ownership Associated with Higher Vitamin D? (PMC3513049)	To investigate potential relationships between household pet exposure and cord blood vitamin D concentrations, we analyzed information from a large, geographically-based, general risk birth cohort
Circulating Cathelicidin Concentrations in a Cohort of Healthy Children: Influence of Age, Body Composition, Gender and Vitamin D Status (PMC4859539)	This study sought to determine if circulating cathelicidin concentrations in healthy children are related to the age of the child, body composition and vitamin D status at birth and at the time of the study visit
A randomized controlled trial of vitamin D supplementation on perinatal depression: in Iranian pregnant mothers (PMC4992225)	This study evaluated the effect of vitamin D3supplementation on perinatal depression scores
Investigation of the vitamin D nutritional status in women with gestational diabetes mellitus in Beijing (PMC5273824)	To study the Vitamin D nutritional status of pregnant women with gestational diabetes mellitus (GDM) in the middle and late pregnancy and analyze the different sources of Vitamin D intake
Trajectory of vitamin D status during pregnancy in relation to neonatal birth size and fetal survival: a prospective cohort study (PMC5812027)	We investigated the associations between vitamin D status in early and late pregnancy with neonatal small for gestational age (SGA), low birth weight (LBW) and preterm delivery
Vitamin D levels in an Australian and New Zealand cohort and the association with pregnancy outcome (PMC6011374)	We aimed to compare vitamin D status in two distinct populations of pregnant women in Australia and New Zealand and to investigate the relationship between vitamin D status and pregnancy outcome

Figure 5.1: Summaries of articles example. A one sentence summary of each article based on the first ignorance statement that is an *indication of answered research question*, a statement of a goal or objective of a study that is attempted or completed during the study. The knowledge goal is to find the answer(s) in the article to determine if the question(s) is (are) fully answered in the article. The lexical cues are highlighted in yellow.

Research is disseminated to everyone through science communications such as science news articles. Ignorance-base tools focused on science communication and journalism could create a stronger connection between journalism and research. Tools could be created to help journalists create news stories and provide more context for stories. For example, it could have been valuable to see the progression of questions and subsequent knowledge around the Covid-19 pandemic as it developed (see [17] for a similar idea). Further, journalists could use an ignorance-base to find the most pertinent questions to publish. Future work could focus on methods and tools to help journalists better communicate the current science. This ecosystem would create connections between researchers, journalists, and people in general, helping to bridge the divide between them.

The general computational idea of an ignorance-base can be applied to anything with data that contains some form of a question, knowledge goal, or ignorance statement. Creating an ignorance-base in general helps organize and connect questions to ideally provide and spark new ideas for questions to pursue in the future. Another application of these ideas is the field of investing and market research. Investors want to know what ideas and products (questions) to invest in (answer/pursue). These same methodologies could be applied to patent documents, grant applications, and idea documents (*e.g.*, Shark Tank television transcripts) to find the next best idea to invest in. These tools could be a part of market research in the future. The difficulty in any application of these general ideas will be what data to use based on the goal.

The main difficulty and limitation of all of this work is data. Which data should be used to capture the known unknowns? These known unknowns begin as unknown unknowns that by definition cannot be captured in any form. The unknown unknowns come to light in the initial discovery of an idea. With the advent of natural language generation (NLG) [293, 294] and applying it to chatbots (*e.g.*, ChatGPT from OpenAI [295]), it may be possible for a chatbot to help researchers capture the unknown unknowns. ChatGPT from OpenAI [295] created a dialogue chatbot system that can answer follow-up questions and challenge incorrect premises. Future work could create a similar system but based on the ignorance-base to help researchers and

others generate and refine novel questions and ideas, getting at the unknown unknowns. Note many conversational systems, like ChatGPT, rely on reinforcement learning [296], which is quite difficult to evaluate. The systems cannot guarantee correct answers and can be biased quite easily [295]. Even still, using the current ignorance-base as starting data to create a chatbot could help generate new data as unknown unknowns. This would ideally also be a part of the question ecosystem.

Lastly, many ethical questions arise when studying questions or ignorance, and future work can tackle these challenges. For example, who asks the question matters in terms of any hidden agendas. Looking at the authors of scientific articles and their backgrounds could help elucidate this. What biases exist in the askers? Racial? Gender? Location? Can all questions be asked by everyone? Future work could look at both published articles and grant proposals to help answer these questions. Further, future work could apply these ignorance methods to the ethics literature itself and see what questions arise. All of this future work could impact the current understanding of science by fostering a questioning and curiosity driven society.

Reflections

This work has taught me how to both create and communicate a novel NLP task. Consistent and concise definitions of the phenomenon were necessary for both communication of the work and annotation of the phenomenon. It was an iterative process to determine these definitions along with examining many examples. To move from my own understanding of these definitions and examples to others understanding them was quite difficult and time consuming. Manual annotation tasks are very time consuming. Further, the adoption of a new task is slow even with a clear and concise definition, a corpus, and classification models. Honing in on the application and impact of the work was very important to convince publishers and funders of these ideas. Researchers also did not like that the name “ignorance" has negative connotations. In the beginning of this work, the main questions I received were about my use of the word “ignorance". It was difficult in general to choose a name, let alone define unknowns using current

knowledge and comparing it to current knowledge. I spent a lot of my PhD focused on how to communicate these ideas to the annotators, my advisors, publishers, and other researchers.

Consulting experts in my field of study early on helped me both communicate it better and determine the application and impact of it. From the beginning, I worked on my ignorance taxonomy in consultation with an ontologist to ensure my correctness in defining unknowns. Once I chose the field of prenatal nutrition to apply all of these ideas to, I found a prenatal nutrition expert to help with everything from collecting the prenatal nutrition articles to the insights I found from the ignorance-base. Domain experts were invaluable to this work.

Lastly, it is important to explain all methods and results (at least the plausibility of them). Many different types of methods exist for classification. Some can be optimized (*e.g.*, a CRF) and some cannot (*e.g.*, neural networks). For understanding, it is important to provide the reasons why a specific classification model was chosen. For reproducibility, it is important to provide all the resource constraints on the algorithms. Not everyone has equal access to resources and so providing alternative models with differing resource needs is helpful [249]. In terms of results, surprising results happen. (I even have an ignorance category to capture it - *indication of anomalous or curious research finding*.) For example, our results that machine translation from a biomedical concept to its ID on the character level improved concept normalization was unexpected because the IDs should not have any underlying semantics. The IDs were meant to be arbitrary and yet it seems we found some structure. When we noticed this worked, we did not fully trust the results at first. To provide more evidence for their feasibility, we conducted several experiments to both break the structure (shuffled and random assignments) and to add more (alphabetical assignments). These experiments gave further evidence to our results. Further, we convinced the reviewer of the paper that even though our results were unexpected, it seemed to be true. Thus, it is important to provide explanations for methods and enough evidence for results.

The main motivation for this work stemmed from my lack of a question to ask for my PhD dissertation because I did not know the questions in my field of computational bioscience. Now at the end of my PhD, I not only have many questions, but helped enumerate some of them for my

dissertation. This work has taught me the importance of questions and curiosity. I hope to use this work and these lessons to help society foster it in education and research to create both scientists and people who trust and can digest science communications.

CHAPTER VI

CONCLUSION

Research not only begins with a question but also ends with one. The first section of a research article, the introduction or background, introduces the main questions of the work. After all the experiments and results, the last sections of a paper, the discussion and conclusion sections, discuss the claims in the broader research context along with limitations and future work. The scientific method is also a cycle beginning where it ended back at the question, known unknown, or ignorance (see Figures 6.1 and 6.2). This dissertation aimed to bring these ideas to the forefront: to automatically help researchers stay up to date on the questions in a field and spark new ideas for future research. We created the first ignorance-base grounded in knowledge goals and the OBOs over the prenatal nutrition literature as a proof of concept of these ideas. We provided all the necessary tools and methods. This pilot study provided research avenues that could affect the health and well-being of mothers and offspring globally. This work revealed and explored the literature's known unknowns to show how ignorance drives science and research. "The only solid piece of scientific truth about which I feel totally confident is that we are profoundly ignorant about nature... It is this sudden confrontation with the depth and scope of ignorance that represents the most significant contribution of twentieth-century science to the human intellect." (Lewis Thomas [297]).

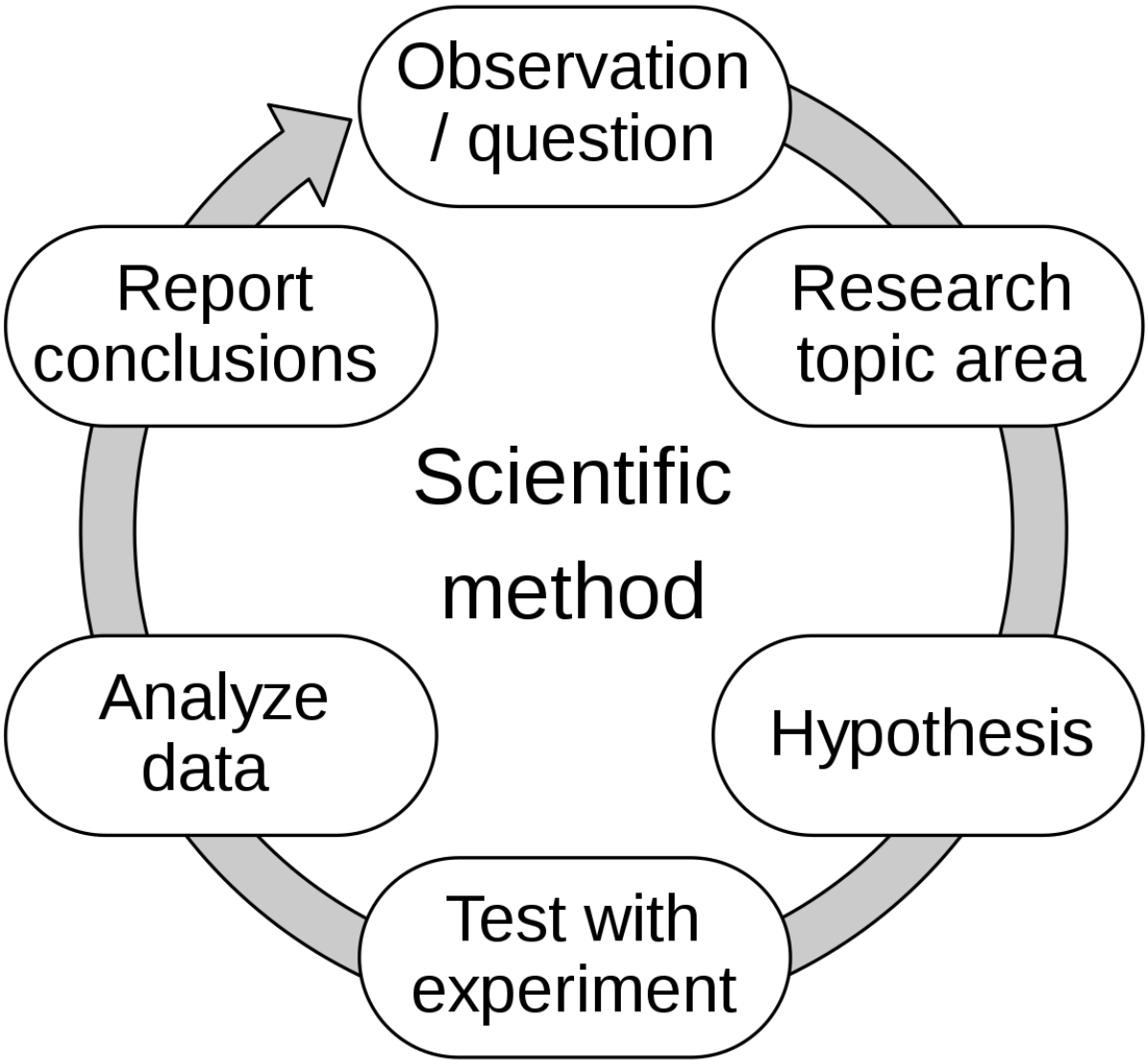


Figure 6.1: The scientific method from Wikipedia.

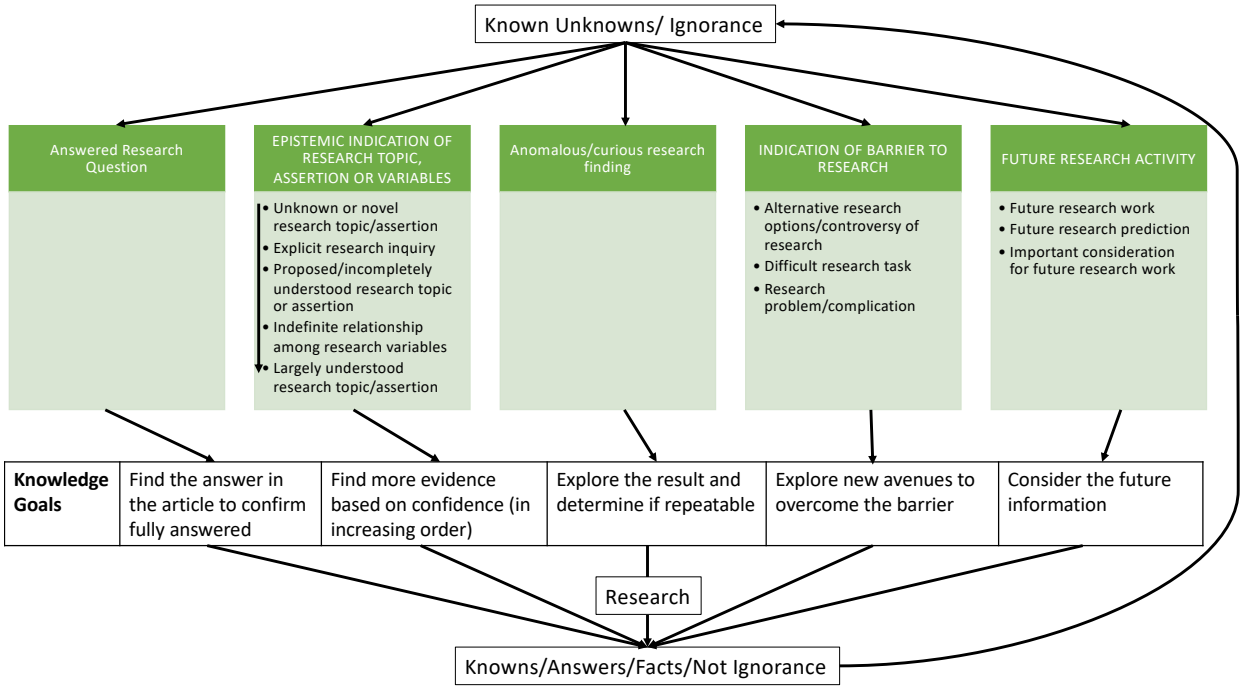


Figure 6.2: Ignorance taxonomy embedded in the research context: Starting from the top, research starts from known unknowns or ignorance. Our ignorance taxonomy is in green (an ignorance statement is an indication of each ignorance category) with knowledge goals underneath. Research is then conducted based on the knowledge goals to get answers; these then filter back to the known unknowns to identify the next research questions.

REFERENCES

1. Firestein S. Ignorance: How It Drives Science. Oxford University Press, USA, 2012.

2. Kuhn TS. The structure of scientific revolutions. [2d ed., enl. International encyclopedia of unified science. Foundations of the unity of science, v. 2, no. 2. Chicago: University of Chicago Press, 1970.

3. O'Leary Z. The Essential Guide to Doing Research. SAGE, 2004.

4. Callahan TJ, Tripodi IJ, Pielke-Lombardo H, and Hunter LE. Knowledge-Based Biomedical Data Science. Annual review of biomedical data science 2020;3:23–41.

5. US Department of Health and Human Services, National Institutes of Health, Office of Dietary Supplements. Dietary Supplement Label Database (DSLD).

6. Yadama AP, Mirzakhani H, McElrath TF, Litonjua AA, and Weiss ST. Transcriptome analysis of early pregnancy vitamin D status and spontaneous preterm birth. PLOS ONE 2020;15:e0227193.

7. Sternberg D. How to Complete and Survive a Doctoral Dissertation. St. Martin's Griffin Publishing Group, 1981.

8. Brause RS. Writing Your Doctoral Dissertation: Invisible Rules for Success. Routledge, 2012.

9. Burton S and Steane P. Surviving Your Thesis. Routledge, 2004.

10. Krathwohl DR and Smith NL. How to Prepare a Dissertation Proposal: Suggestions for Students in Education and the Social and Behavioral Sciences. Syracuse University Press, 2005.

11. Terrell SR. Writing a Proposal for Your Dissertation: Guidelines and Examples. Guilford Publications, 2022.

12. Tanaka ML. A Thesis Proposal Development Course for Engineering Graduate Students. Journal of Biomechanical Engineering 2020;142.

13. Madsen D. Successful Dissertations and Theses: A Guide to Graduate Student Research from Proposal to Completion. Tech. rep. Jossey-Bass Inc, 1983.

14. Lei SA. Strategies for Finding and Selecting an Ideal Thesis or Dissertation Topic: A Review of Literature. College Student Journal 2009;43:1324–32.

15. Ségol G. Choosing a Dissertation Topic: Additional Pointers. College Student Journal 2014;48:108–13.

16. Eze U, Adebayo O, Nnodim I, Adejo A, and Obazenu L. How to choose a dissertation topic. Nigerian Journal of Medicine 2021;30:123–3.

17. Lahav D, Falcon JS, Kuehl B, et al. A search engine for discovery of scientific challenges and directions. In: AAAI, 2022.

18. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research 2004;32:D267–D270.

19. Turner M, Ive J, and Velupillai S. Linguistic Uncertainty in Clinical NLP: A Taxonomy, Dataset and Approach. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Candan KS, Ionescu B, Goeuriot L, et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2021:129–41.

20. Szarvas G. Hedge Classification in Biomedical Texts with a Weakly Supervised Selection of Keywords. In: *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, 2008:281–9.

21. Cheng K, Baldwin T, and Verspoor K. Automatic Negation and Speculation Detection in Veterinary Clinical Text. In: *Proceedings of the Australasian Language Technology Association Workshop 2017*. Brisbane, Australia, 2017:70–8.

22. Zhang S, Kang T, Zhang X, Wen D, Elhadad N, and Lei J. Speculation detection for Chinese clinical notes: Impacts of word segmentation and embedding models. Journal of Biomedical Informatics 2016;60:334–41.

23. Medlock B. Exploring hedge identification in biomedical literature. Journal of Biomedical Informatics 2008;41:636–54.

24. Medlock B and Briscoe T. Weakly Supervised Learning for Hedge Classification in Scientific Literature. In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, 2007:992–9.

25. Liu F, Zhou P, Baccei SJ, et al. Qualifying Certainty in Radiology Reports through Deep Learning–Based Natural Language Processing. American Journal of Neuroradiology 2021;42:1755–61.

26. Islam J, Xiao L, and Mercer RE. A Lexicon-Based Approach for Detecting Hedges in Informal Text. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020:3109–13.

27. Light M, Qiu XY, and Srinivasan P. The Language of Bioscience: Facts, Speculations, and Statements In Between. In: *HLT-NAACL 2004 Workshop: BioLINK 2004, Linking Biological Literature, Ontologies and Databases*. Ed. by Hirschman L and Pustejovsky J. Boston, Massachusetts, USA: Association for Computational Linguistics, 2004:17–24.

28. Tipney H and Hunter L. An introduction to effective use of enrichment analysis software. Human Genomics 2010;4:202.

29. Shi Jing L, Fathiah Muzaffar Shah F, Saberi Mohamad M, et al. A Review on Bioinformatics Enrichment Analysis Tools Towards Functional Analysis of High Throughput Gene Set Data. Current Proteomics 2015;12:14–27.

30. Hung JH, Yang TH, Hu Z, Weng Z, and DeLisi C. Gene set enrichment analysis: performance evaluation and usage guidelines. Briefings in Bioinformatics 2012;13:281–91.

31. Curtis RK, Orešič M, and Vidal-Puig A. Pathways to the analysis of microarray data. Trends in Biotechnology 2005;23:429–35.

32. Wijesooriya K, Jadaan SA, Perera KL, Kaur T, and Ziemann M. Urgent need for consistent standards in functional enrichment analysis. PLOS Computational Biology 2022;18:e1009935.

33. Boyack KW and Börner K. Indicator-assisted evaluation and funding of research: Visualizing the influence of grants on the number and citation counts of research papers. Journal of the American Society for Information Science and Technology 2003;54:447–61.

34. Rihn B, Vidal S, Nemurat C, et al. From transcriptomics to bibliomics. Medical science monitor : international medical journal of experimental and clinical research 2003;9:MT89–95.

35. Balili C, Segev A, and Lee U. Tracking and predicting the evolution of research topics in scientific literature. In: *2017 IEEE International Conference on Big Data (Big Data)*. 2017:1694–7.

36. Faruqui M and Das D. Identifying Well-formed Natural Language Questions. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, 2018:798–803.

37. Lakoff G. Hedges: A study in meaning criteria and the logic of fuzzy concepts. In: *Contemporary research in philosophical logic and linguistic semantics*. Springer, 1975:221–71.

38. Hyland K. Hedging in scientific research articles. Vol. 54. John Benjamins Publishing, 1998.

39. Farkas R, Vincze V, Móra G, Csirik J, and Szarvas G. The CoNLL-2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*. Uppsala, Sweden: Association for Computational Linguistics, 2010:1–12.

40. Ganter V and Strube M. Finding Hedges by Chasing Weasels: Hedge Detection Using Wikipedia Tags and Shallow Linguistic Features. In: *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*. Suntec, Singapore: Association for Computational Linguistics, 2009:173–6.

41. Vincze V, Szarvas G, Farkas R, Móra G, and Csirik J. The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. BMC Bioinformatics 2008;9:S9.

42. Kilicoglu H and Bergler S. Recognizing Speculative Language in Biomedical Research Articles: A Linguistically Motivated Perspective. In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Columbus, Ohio: Association for Computational Linguistics, 2008:46–53.

43. Kilicoglu H, Rosemblat G, and Rindflesch TC. Assigning factuality values to semantic relations extracted from biomedical research literature. PLOS ONE 2017;12:e0179926.

44. Zerva C, Batista-Navarro R, Day P, and Ananiadou S. Using uncertainty to link and rank evidence from biomedical literature for model curation. Bioinformatics (Oxford, England) 2017;33:3784–92.

45. Al-Khawaldeh FT. Hierarchical Attention Generative Adversarial Networks for Biomedical Texts Uncertainty Detection. INTERNATIONAL JOURNAL OF ADVANCED STUDIES 2019;8:14.

46. Szarvas G, Vincze V, Farkas R, Móra G, and Gurevych I. Cross-Genre and Cross-Domain Detection of Semantic Uncertainty. Computational Linguistics 2012;38:335–67.

47. Velldal E. Detecting Uncertainty in Biomedical Literature: A Simple Disambiguation Approach Using Sparse Random Indexing. Semantic Mining in Biomedicine 2010:9.

48. Fujikawa K, Seki K, and Uehara K. A hybrid approach to finding negated and uncertain expressions in biomedical documents. In: *Proceedings of the 2nd international workshop on Managing interoperability and compleXity in health systems - MIXHS '12*. Maui, Hawaii, USA: ACM Press, 2012:67.

49. Ren Y, Fei H, and Peng Q. Detecting the Scope of Negation and Speculation in Biomedical Texts by Using Recursive Neural Network. In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2018:739–42.

50. Konstantinova N and C. M. de Sousa S. Annotating Negation and Speculation: the Case of the Review Domain. In: *Proceedings of the Second Student Research Workshop associated with RANLP 2011*. Hissar, Bulgaria: Association for Computational Linguistics, 2011:139–44.

51. Zavala RR and Martinez P. The Impact of Pretrained Language Models on Negation and Speculation Detection in Cross-Lingual Medical Text: Comparative Study. JMIR Medical Informatics 2020;8:e18953.

52. Apostolova E, Tomuro N, and Demner-Fushman D. Automatic Extraction of Lexico-Syntactic Patterns for Detection of Negation and Speculation Scopes. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, 2011:283–7.

53. Qian Z, Li P, Zhu Q, Zhou G, Luo Z, and Luo W. Speculation and Negation Scope Detection via Convolutional Neural Networks. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, 2016:815–25.

54. Fei H, Ren Y, and Ji D. Negation and speculation scope detection using recursive neural conditional random fields. Neurocomputing 2020;374:22–9.

55. AL-Khawaldeh FT. Speculation and Negation Annotation for Arabic Biomedical Texts BioArabic Corpus. 2016:4.

56. Khandelwal A and Britto BK. Multitask Learning of Negation and Speculation using Transformers. In: *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*. Online: Association for Computational Linguistics, 2020:79–87.

57. Al-Khawaldeh FT. Speculation and Negation Detection for Arabic Biomedical Texts. 2019:6.

58. Velldal E. Predicting speculation: a simple disambiguation approach to hedge detection in biomedical literature. Journal of Biomedical Semantics 2011;2:S7.

59. Agarwal S and Yu H. Detecting hedge cues and their scope in biomedical text with conditional random fields. Journal of Biomedical Informatics 2010;43:953–61.

60. Zhou H, Li X, Huang D, Li Z, and Yang Y. Exploiting Multi-Features to Detect Hedges and their Scope in Biomedical Texts. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*. Uppsala, Sweden: Association for Computational Linguistics, 2010:106–13.

61. Zhou H, Deng H, Huang D, and Zhu M. Hedge Scope Detection in Biomedical Texts: An Effective Dependency-Based Method. PLOS ONE 2015;10:e0133715.

62. Ji F, Qiu X, and Huang X. Detecting Hedge Cues and their Scopes with Average Perceptron. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*. Uppsala, Sweden: Association for Computational Linguistics, 2010:32–9.

63. Hanauer DA, Liu Y, Mei Q, Manion FJ, Balis UJ, and Zheng K. Hedging their Mets: The Use of Uncertainty Terms in Clinical Documents and its Potential Implications when Sharing the Documents with Patients. AMIA Annual Symposium Proceedings 2012;2012:321–30.

64. Zhao Q, Sun C, Liu B, and Cheng Y. Learning to Detect Hedges and their Scope Using CRF. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*. Uppsala, Sweden: Association for Computational Linguistics, 2010:100–5.

65. Clausen D. HedgeHunter: A System for Hedge Detection and Uncertainty Classification. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*. Uppsala, Sweden: Association for Computational Linguistics, 2010:120–5.

66. Morante R and Daelemans W. Learning the Scope of Hedge Cues in Biomedical Texts. In: *Proceedings of the BioNLP 2009 Workshop*. Boulder, Colorado: Association for Computational Linguistics, 2009:28–36.

67. Verbeke M, Frasconi P, Van Asch V, Morante R, Daelemans W, and De Raedt L. Kernel-Based Logical and Relational Learning with kLog for Hedge Cue Detection. In: *Inductive Logic Programming*. Ed. by Muggleton SH, Tamaddoni-Nezhad A, and Lisi FA. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2012:347–57.

68. Mowery DL, Velupillai S, and Chapman WW. Medical diagnosis lost in translation – Analysis of uncertainty and negation expressions in English and Swedish clinical texts. In: *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*. Montréal, Canada: Association for Computational Linguistics, 2012:56–64.

69. Han PK, Klein WM, and Arora NK. Varieties of uncertainty in health care: a conceptual taxonomy. Medical decision making : an international journal of the Society for Medical Decision Making 2011;31:828–38.

70. Pearl J and Mackenzie D. The book of why: the new science of cause and effect. New York: Basic Books, 2018.

71. Regan HM, Colyvan M, and Burgman MA. A Taxonomy and Treatment of Uncertainty for Ecology and Conservation Biology. Ecological Applications 2002;12:618–28.

72. Smithson M. Ignorance and Uncertainty: Emerging Paradigms. Springer Science & Business Media, 2012.

73. Walker VR. The Siren Songs of Science: Toward a Taxonomy of Scientific Uncertainty for Decisionmakers. CONNECTICUT LAW REVIEW 1991:61.

74. Bandrowski A, Brinkman R, Brochhausen M, et al. The Ontology for Biomedical Investigations. PLOS ONE 2016;11. Ed. by Xue Y:e0154556.

75. Bastian FB, Chibucos MC, Gaudet P, et al. The Confidence Information Ontology: a step towards a standard for asserting confidence in annotations. Database 2015;2015:bav043–bav043.

76. Brush MH, Shefchek K, and Haendel M. SEPIO: A Semantic Model for the Integration and Analysis of Scientific Evidence. CEUR Workshop Proceedings 2016;1747:6.

77. Chibucos MC, Mungall CJ, Balakrishnan R, et al. Standardized description of scientific evidence using the Evidence Ontology (ECO). Database 2014;2014:bau075–bau075.

78. Rindflesch TC and Fiszman M. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics. Unified Medical Language System 2003;36:462–77.

79. Thompson P, Nawaz R, McNaught J, and Ananiadou S. Enriching a biomedical event corpus with meta-knowledge annotation. BMC Bioinformatics 2011;12:393.

80. Bongelli R, Canestrari C, Riccioni I, et al. A Corpus of Scientific Biomedical Texts Spanning over 168 Years Annotated for Uncertainty. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), 2012:2009–14.

81. Cruz Díaz NP. Detecting Negated and Uncertain Information in Biomedical and Review Texts. In: *Proceedings of the Student Research Workshop associated with RANLP 2013*. Hissar, Bulgaria: INCOMA Ltd. Shoumen, BULGARIA, 2013:45–50.

82. Konstantinova N, Sousa SC de, Cruz NP, Maña MJ, Taboada M, and Mitkov R. A review corpus annotated for negation, speculation and their scope. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), 2012:3190–5.

83. Zhou HW, Yang H, Zhang J, Kang SY, and Huang DG. The research and construction of Chinese hedge corpus. Journal of Chinese Information Processing 2015;29:83–9.

84. Sanchez LM and Vogel C. A hedging annotation scheme focused on epistemic phrases for informal language. In: *Proceedings of the Workshop on Models for Modality Annotation*. London, UK: Association for Computational Linguistics, 2015.

85. Jiménez-Zafra SM, Taulé M, Martín-Valdivia MT, Ureña-López LA, and Martí MA. SFU ReviewSP-NEG: a Spanish corpus annotated with negation for sentiment analysis. A typology of negation patterns. Language Resources and Evaluation 2018;52:533–69.

86. Yang H, Roeck AD, Gervasi V, Willis A, and Nuseibeh B. Speculative requirements: Automatic detection of uncertainty in natural language requirements. In: *2012 20th IEEE International Requirements Engineering Conference (RE)*. 2012:11–20.

87. Jean PA, Harispe S, Ranwez S, Bellot P, and Montmain J. Uncertainty detection in natural language: a probabilistic model. In: *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*. Nîmes France: ACM, 2016:1–10.

88. Sergeeva E, Zhu H, Tahmasebi A, and Szolovits P. Neural Token Representations and Negation and Speculation Scope Detection in Biomedical and General Domain Text. In: *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*. Hong Kong: Association for Computational Linguistics, 2019:178–87.

89. Zhou H, Huang D, Li X, and Yang Y. Combining structured and flat features by a composite Kernel to detect hedges scope in biological texts. Chinese Journal of Electronics 2011;20:476–82.

90. Zhou H, Li X, Huang D, Yang Y, and Ren F. Voting-Based Ensemble Classifiers to Detect Hedges and Their Scopes in Biomedical Texts. IEICE TRANSACTIONS on Information and Systems 2011;E94-D:1989–97.

91. Zhou H, Xu J, Yang Y, Deng H, Chen L, and Huang D. Chinese Hedge Scope Detection Based on Structure and Semantic Information. In: *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*. Ed. by Sun M, Huang X, Lin H, Liu Z, and Liu Y. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016:204–15.

92. Georgescul M. A Hedgehop Over a Max-Margin Framework Using Hedge Cues. Proceedings of the 14th International Conference on Computational Natural Language Learning: Shared Task 2010:26.

93. Moncecchi G. Recognizing Speculative Language in Research Texts. PhD thesis. Université de Nanterre - Paris X ; Universidad de la República - Proyecto de Apoyo a las Ciencias Básicas, 2013.

94. Díaz NPC. Detección de la Negación y la Especulación en Textos Médicos y de. PhD Thesis. Universidad de Huelva, 2014.

95. Dalloux C, Claveau V, and Grabar N. Speculation and negation detection in french biomedical corpora. In: *RANLP 2019 - Recent Advances in Natural Language Processing*. Varna, Bulgaria, 2019:1–10.

96. Zou B, Zhu Q, and Zhou G. Negation and Speculation Identification in Chinese Language. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, 2015:656–65.

97. Vidal S, Ducloy J, and Houdry P. Mining medical data using multiple corpora interaction:the Transcriptomics investigation server experiment. 2003.

98. Chen J, Xu H, Aronow BJ, and Jegga AG. Improved human disease candidate gene prioritization using mouse phenotype. BMC Bioinformatics 2007;8:392.

99. Jourquin J, Duncan D, Shi Z, and Zhang B. GLAD4U: deriving and prioritizing gene lists from PubMed literature. BMC Genomics 2012;13:S20.

100. Jenssen TK, Lægreid A, Komorowski J, and Hovig E. A literature network of human genes for high-throughput analysis of gene expression. Nature genetics 2001;28:21–8.

101. Rebholz-Schuhmann D, Kirsch H, Arregui M, Gaudan S, Riethoven M, and Stoehr P. EBIMed–text crunching to gather facts for proteins from Medline. Bioinformatics (Oxford, England) 2007;23:e237–244.

102. Fontaine JF, Priller F, Barbosa-Silva A, and Andrade-Navarro MA. Genie: literature-based gene prioritization at multi genomic scale. Nucleic acids research 2011;39:W455–W461.

103. Börnigen D, Tranchevent LC, Bonachela-Capdevila F, et al. An unbiased evaluation of gene prioritization tools. Bioinformatics 2012;28:3081–8.

104. Grimes GR, Wen TQ, Mewissen M, et al. PDQ Wizard: automated prioritization and characterization of gene and protein lists using biomedical literature. Bioinformatics 2006;22:2055–7.

105. Modlin S, Gunasekaran D, Zlotnicki A, et al. Resolving the hypotheticome: annotating M. tuberculosis gene function through bibliomic reconciliation and structural modeling. 2018.

106. Yang L, Wang B, Xia G, Xia Z, and Xu L. Bibliomics-based Selection of Analgesics Targets through Google-PageRank-like Algorithm. In: *2007 Second International Conference on Bio-Inspired Computing: Theories and Applications*. 2007:98–101.

107. Hettne KM, Weeber M, Laine ML, et al. Automatic mining of the literature to generate new hypotheses for the possible link between periodontitis and atherosclerosis: lipopolysaccharide as a case study. Journal of Clinical Periodontology 2007;34:1016–24.

108. Miwa M, Ohta T, Rak R, et al. A method for integrating and ranking the evidence for biochemical pathways by mining reactions from text. Bioinformatics (Oxford, England) 2013;29:i44–52.

109. Zerva C. Automatic identification of textual uncertainty. PhD thesis. THE UNIVERSITY OF MANCHESTER, 2019.

110. Holdcroft A. Gender bias in research: how does it affect evidence based medicine? Journal of the Royal Society of Medicine 2007;100:2–3.

111. Slawson N. 'Women have been woefully neglected': does medical science have a gender problem? 2019.

112. Mastroianni AC, Henry LM, Robinson D, et al. Research with Pregnant Women: New Insights on Legal Decision-Making. The Hastings Center report 2017;47:38–45.

113. Meadows M. Pregnancy and the Drug Dilemma: (542682006-004). Tech. rep. American Psychological Association, 2001.

114. Hanson MA and Gluckman PD. Early Developmental Conditioning of Later Health and Disease: Physiology or Pathophysiology? Physiological Reviews 2014;94:1027–76.

115. Arshad R, Sameen A, Murtaza MA, et al. Impact of vitamin D on maternal and fetal health: A review. Food Science & Nutrition 2022;10:3230–40.

116. Tous M, Villalobos M, Iglesias L, Fernández-Barrés S, and Arija V. Vitamin D status during pregnancy and offspring outcomes: a systematic review and meta-analysis of observational studies. European Journal of Clinical Nutrition 2020;74:36–53.

117. Todorova M, Gerova D, and Galunska B. Vitamin D deficiency during pregnancy. Scripta Scientifica Medica 2022;54:19–28.

118. Dror D. Vitamin D in pregnancy. In: *Handbook of vitamin D in human health: Prevention, treatment and toxicity*. Ed. by Watson RR. Human Health Handbooks. Wageningen: Academic Publishers, 2013:670–91.

119. KIRCA AŞ. The Effect of Vitamin D Deficiency in Pregnancy on Maternal Results. Recent Studies in Health Sciences 2019:359.

120. Perera N, Dehmer M, and Emmert-Streib F. Named Entity Recognition and Relation Detection for Biomedical Information Extraction. Frontiers in Cell and Developmental Biology 2020;8.

121. Suman S, Dash A, and Rautaray SS. A Literature Survey on Biomedical Named Entity Recognition. In: *Advances in Power Systems and Energy Management*. Ed. by Priyadarshi N, Padmanaban S, Ghadai RK, Panda AR, and Patel R. Lecture Notes in Electrical Engineering. Singapore: Springer Nature, 2021:109–19.

122. Cohen KB and Demner-Fushman D. Biomedical Natural Language Processing. John Benjamins Publishing Company, 2014.

123. Demoulin NTM and Coussement K. Acceptance of text-mining systems: The signaling role of information quality. Information & Management. Big data and business analytics: A research agenda for realizing business value 2020;57:103120.

124. Leser U and Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. Briefings in Bioinformatics 2005;6:357–69.

125. Hirschman L, Yeh A, Blaschke C, and Valencia A. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 2005;6:S1.

126. Jin-Dong K, Claire N, Robert B, and Louise D, eds. Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. Hong Kong, China: Association for Computational Linguistics, 2019.

127. Lu Wang L, Lo K, Chandrasekhar Y, et al. CORD-19: The Covid-19 Open Research Dataset. ArXiv 2020.

128. Furrer L, Cornelius J, and Rinaldi F. UZH@CRAFT-ST: a Sequence-labeling Approach to Concept Recognition. In: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. Hong Kong, China: Association for Computational Linguistics, 2019:185–95.

129. Lafferty J, McCallum A, and Pereira FCN. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. 2001:10.

130. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2019. Ed. by Wren J:btz682.

131. Klein G, Kim Y, Deng Y, Nguyen V, Senellart J, and Rush A. OpenNMT: Neural Machine Translation Toolkit. In: *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*. Boston, MA: Association for Machine Translation in the Americas, 2018:177–84.

132. Balili C, Lee U, Segev A, Kim J, and Ko M. TermBall: Tracking and Predicting Evolution Types of Research Topics by Using Knowledge Structures in Scholarly Big Data. IEEE Access 2020;8:108514–29.

133. Sutherland WJ, Fleishman E, Mascia MB, Pretty J, and Rudd MA. Methods for collaboratively identifying research priorities and emerging issues in science and policy. Methods in Ecology and Evolution 2011;2:238–47.

134. Liu N, Shapira P, and Yue X. Tracking developments in artificial intelligence research: constructing and applying a new search strategy. Scientometrics 2021;126:3153–92.

135. Ostrowsky J, Arpey M, Moore K, et al. Tracking progress in universal influenza vaccine development. Current Opinion in Virology 2020;40:28–36.

136. Dinakar B, Boguslav MR, Görg C, and Dinakarpandian D. Semantic Changepoint Detection for Finding Potentially Novel Research Publications. In: *Pac Symp Biocomput*. World Scientific, 2020.

137. Antonio E, Alobo M, Tufet Bayona M, Marsh K, and Norton A. Funding and COVID-19 research priorities - are the research needs for Africa being met? AAS Open Research 2020;3:56.

138. Benner JS, Morrison MR, Karnes EK, Kocot SL, and McClellan M. An Evaluation Of Recent Federal Spending On Comparative Effectiveness Research: Priorities, Gaps, And Next Steps. Health Affairs 2010;29:1768–76.

139. Wallis S, Cole DC, Gaye O, et al. Qualitative study to develop processes and tools for the assessment and tracking of African institutions' capacity for operational health research. BMJ Open 2017;7:e016660.

140. Swanson DR. Searching Natural Language Text by Computer. Science 1960.

141. Hunter L and Cohen KB. Biomedical Language Processing: What's Beyond PubMed? Molecular Cell 2006;21:589–94.

142. Henry S and McInnes BT. Literature Based Discovery: Models, methods, and trends. Journal of Biomedical Informatics 2017;74:20–32.

143. Swanson DR. Undiscovered Public Knowledge. The Library Quarterly 1986;56:103–18.

144. Fukuda K, Tsunoda T, Tamura A, and Takagi T. Information extraction: Identifying protein names from biological papers. In: *In Proceedings of the Pacific Symposium on Biocomputing '98 (PSB'98*. 1998:707–18.

145. Blaschke C, Andrade MA, Ouzounis C, and Valencia A. Automatic Extraction of Biological Information from Scientific Text: Protein-Protein Interactions. 1999:8.

146. Craven M and Kumlien J. Constructing Biological Knowledge Bases by Extracting Information from Text Sources. 1999:10.

147. Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, and Weinstein JN. MedMiner: An Internet Text-Mining Tool for Biomedical Information, with Application to Gene Expression Profiling. BioTechniques 1999;27:1210–7.

148. Hoffmann R and Valencia A. A gene network for navigating the literature. Nature Genetics 2004;36:664–4.

149. Müller HM, Kenny EE, and Sternberg PW. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. PLOS Biology 2004;2:e309.

150. Maglott D, Ostell J, Pruitt KD, and Tatusova T. Entrez Gene: gene-centered information at NCBI. Nucleic Acids Research 2007;35:D26–D31.

151. Arp R, Smith B, and Spear AD. Building Ontologies with Basic Formal Ontology. MIT Press, 2015.

152. Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. Nature genetics 2000;25:25–9.

153. Harris MA, Clark J, Ireland A, et al. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 2004;32:D258–261.

154. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research 2019;47:D330–D338.

155. Smith B, Ashburner M, Rosse C, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature Biotechnology 2007;25:1251–5.

156. MEDLINE®/PubMed® Data Element (Field) Descriptions. Technical Documentation.

157. Cohen KB, Johnson HL, Verspoor K, Roeder C, and Hunter LE. The structural and content aspects of abstracts versus bodies of full text journal articles are different. BMC Bioinformatics 2010;11:492.

158. Lin J. Is searching full text more effective than searching abstracts? BMC Bioinformatics 2009;10:46.

159. PMC Open Access Subset - PMC. 2003.

160. Liu J, Kong X, Zhou X, et al. Data Mining and Information Retrieval in the 21st century: A bibliographic review. Computer Science Review 2019;34:100193.

161. Du XJ, Bathgate RAD, Samuel CS, Dart AM, and Summers RJ. Cardiovascular effects of relaxin: from basic science to clinical therapy. Nature Reviews Cardiology 2010;7:48–58.

162. Arms WY. TREC: Experiment and Evaluation in Information Retrieval. MIT Press, 2005.

163. Hripcsak G and Rothschild AS. Agreement, the F-Measure, and Reliability in Information Retrieval. Journal of the American Medical Informatics Association : JAMIA 2005;12:296–8.

164. Kim JD, Ohta T, Tateisi Y, and Tsujii J. GENIA corpus—a semantically annotated corpus for bio-textmining. Bioinformatics 2003;19:i180–i182.

165. Thompson P, Ananiadou S, and Tsujii J. The GENIA Corpus: Annotation Levels and Applications. In: *Handbook of Linguistic Annotation*. Ed. by Ide N and Pustejovsky J. Dordrecht: Springer Netherlands, 2017:1395–432.

166. Cohen KB, Verspoor K, Fort K, et al. The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation in the Biomedical Domain. In: *Handbook of Linguistic Annotation*. Ed. by Ide N and Pustejovsky J. Dordrecht: Springer Netherlands, 2017:1379–94.

167. Cohen KB, Lanfranchi A, Corvey W, et al. Annotation of all coreference in biomedical text: Guideline selection and adaptation. In: 2010:37–41.

168. Verspoor K, Cohen KB, Lanfranchi A, et al. A corpus of full-text journal articles is a robust evaluation tool for revealing differences in performance of biomedical natural language processing tools. BMC Bioinformatics 2012;13:207.

169. Bada M, Eckert M, Evans D, et al. Concept annotation in the CRAFT corpus. BMC Bioinformatics 2012;13:161.

170. Baumgartner W, Bada M, Pyysalo S, et al. CRAFT Shared Tasks 2019 Overview — Integrated Structure, Semantics, and Coreference. In: *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*. Hong Kong, China: Association for Computational Linguistics, 2019:174–84.

171. Huang CC and Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in Bioinformatics 2016;17:132–44.

172. Filannino M and Uzuner Ö. Advancing the State of the Art in Clinical Natural Language Processing through Shared Tasks. Yearbook of Medical Informatics 2018;27:184–92.

173. Boguslav MR, Salem NM, White EK, Leach SM, and Hunter LE. Identifying and classifying goals for scientific knowledge. Bioinformatics Advances 2021;1:vbab012.

174. Boguslav MR, Hailu ND, Bada M, Baumgartner WA, and Hunter LE. Concept recognition as a machine translation problem. BMC Bioinformatics 2021;22:598.

175. Boguslav MR, Salem NM, White EK, et al. Creating an Ignorance-Base: Exploring Known Unknowns in the Scientific Literature. 2022.

176. Boguslav M and Cohen KB. Inter-Annotator Agreement and the Upper Limit on Machine Performance: Evidence from Biomedical Natural Language Processing. Studies in Health Technology and Informatics 2017;245:298–302.

177. Ross BC, Boguslav M, Weeks H, and Costello JC. Simulating heterogeneous populations using Boolean models. BMC Systems Biology 2018;12.

178. Boguslav M, Cohen KB, Baumgartner WA, and Hunter LE. Improving precision in concept normalization. In: *Biocomputing 2018*. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC, 2018:566–77.

179. Tripodi I, Boguslav M, Hailu N, and Hunter L. Knowledge-base-enriched relation extraction. 2018.

180. Kanimozhi U and Manjula D. A Systematic Review on Biomedical Named Entity Recognition. In: *Data Science Analytics and Applications*. Ed. by R S and Sharma M. Communications in Computer and Information Science. Singapore: Springer, 2018:19–37.

181. Chang L, Zhang R, Lv J, Zhou W, and Bai Y. A review of biomedical named entity recognition. Journal of Computational Methods in Sciences and Engineering 2022;22:893–900.

182. Song B, Li F, Liu Y, and Zeng X. Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. Briefings in Bioinformatics 2021;22:bbab282.

183. French E and McInnes BT. An overview of biomedical entity linking throughout the years. Journal of Biomedical Informatics 2023;137:104252.

184. Natale DA, Arighi CN, Barker WC, et al. The Protein Ontology: a structured representation of protein forms and complexes. Nucleic Acids Research 2011;39:D539–D545.

185. Köhler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Research 2014;42:D966–D974.

186. Hailu ND. Investigation of traditional and deep neural sequence models for biomedical concept recognition. Thesis. University of Colorado at Denver, Anschutz Medical Campus. Health Sciences Library, 2019.

187. Ramshaw LA and Marcus MP. Text Chunking Using Transformation-Based Learning. In: *Natural Language Processing Using Very Large Corpora*. Ed. by Armstrong S, Church K, Isabelle P, Manzi S, Tzoukermann E, and Yarowsky D. Text, Speech and Language Technology. Dordrecht: Springer Netherlands, 1999:157–76.

188. Hochreiter S and Schmidhuber J. Long Short-Term Memory. Neural Computation 1997;9:1735–80.

189. Lyu C, Chen B, Ren Y, and Ji D. Long short-term memory RNN for biomedical named entity recognition. BMC Bioinformatics 2017;18:462.

190. Sheikhshabbafghi G, Birol I, and Sarkar A. In-domain Context-aware Token Embeddings Improve Biomedical Named Entity Recognition. In: *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*. Brussels, Belgium: Association for Computational Linguistics, 2018:160–4.

191. Dai X, Karimi S, Hachey B, and Paris C. An Effective Transition-based Model for Discontinuous NER. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020:5860–70.

192. Dai X. Recognizing Complex Entity Mentions: A Review and Future Directions. In: *Proceedings of ACL 2018, Student Research Workshop*. Melbourne, Australia: Association for Computational Linguistics, 2018:37–44.

193. Popescu-Belis A. Context in Neural Machine Translation: A Review of Models and Evaluations. arXiv:1901.09115 [cs] 2019.

194. Alan Ruttenberg, Melanie Courtot, and Chris Mungall. OBO Foundry Identifier Policy. 2019.

195. Keeping checks on machine learning. Nature Methods 2021;18:1119–9.

196. Funk C, Baumgartner W, Garcia B, et al. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. BMC Bioinformatics 2014;15:59.

197. Apache UIMA ConceptMapper Annotator Documentation.

198. Tanenblatt M, Coden A, and Sominsky I. The ConceptMapper approach to named entity recognition. In: *Proceedings of the seventh international conference on language resources and evaluation (LREC'10)*. 2010.

199. Tseytlin E, Mitchell K, Legowski E, Corrigan J, Chavan G, and Jacobson RS. NOBLE – Flexible concept recognition for large-scale biomedical natural language processing. BMC Bioinformatics 2016;17:32.

200. Wen A, Fu S, Moon S, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. npj Digital Medicine 2019;2:1–7.

201. Campos D, Matos S, and Oliveira JL. A modular framework for biomedical concept recognition. BMC Bioinformatics 2013;14:281.

202. Nunes T, Campos D, Matos S, and Oliveira JL. BeCAS: biomedical concept recognition services and visualization. Bioinformatics 2013;29:1915–6.

203. Groza T and Verspoor K. Assessing the impact of case sensitivity and term information gain on biomedical concept recognition. PloS one 2015;10:e0119091.

204. Basaldella M, Furrer L, Tasso C, and Rinaldi F. Entity recognition in the biomedical domain using a hybrid approach. Journal of Biomedical Semantics 2017;8:51.

205. Furrer L, Jancso A, Colic N, and Rinaldi F. OGER++: hybrid multi-type entity recognition. Journal of Cheminformatics 2019;11:7.

206. Manda P, SayedAhmed S, and Mohanty SD. Automated ontology-based annotation of scientific literature using deep learning. In: *Proceedings of The International Workshop on Semantic Big Data*. 2020:1–6.

207. Devkota P, Mohanty SD, and Manda P. A Gated Recurrent Unit based architecture for recognizing ontology concepts from biological literature. BioData Mining 2022;15:22.

208. Huang Z, Xu W, and Yu K. Bidirectional LSTM-CRF Models for Sequence Tagging. arXiv:1508.01991 [cs] 2015.

209. Lample G, Ballesteros M, Subramanian S, Kawakami K, and Dyer C. Neural Architectures for Named Entity Recognition. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016:260–70.

210. Ma X and Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In: *Proceedings of the 54th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers). Berlin, Germany: Association for Computational Linguistics, 2016:1064–74.

211. Gillick D, Brunk C, Vinyals O, and Subramanya A. Multilingual Language Processing From Bytes. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, 2016:1296–306.

212. Habibi M, Weber L, Neves M, Wiegandt DL, and Leser U. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics 2017;33:i37–i48.

213. Gridach M. Character-level neural network for biomedical named entity recognition. Journal of Biomedical Informatics 2017;70:85–91.

214. Zhao Z, Yang Z, Luo L, et al. Disease named entity recognition from biomedical literature using a novel convolutional neural network. BMC Medical Genomics 2017;10:73.

215. Korvigo I, Holmatov M, Zaikovskii A, and Skoblov M. Putting hands to rest: efficient deep CNN-RNN architecture for chemical named entity recognition with no hand-crafted rules. Journal of cheminformatics 2018;10:1–10.

216. Luo L, Yang Z, Yang P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. Bioinformatics 2018;34:1381–8.

217. Corbett P and Boyle J. Chemlistem: chemical named entity recognition using recurrent neural networks. Journal of Cheminformatics 2018;10:59.

218. Unanue IJ, Borzeshi EZ, and Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. Journal of biomedical informatics 2017;76:102–9.

219. Wang X, Lyu J, Dong L, and Xu K. Multitask learning for biomedical named entity recognition with cross-sharing structure. BMC Bioinformatics 2019;20:427.

220. Devlin J, Chang MW, Lee K, and Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of NAACL-HLT*. 2019:4171–86.

221. Beltagy I, Lo K, and Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019:3615–20.

222. Peters M, Neumann M, Iyyer M, et al. Deep Contextualized Word Representations. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018:2227–37.

223. Peng Y, Yan S, and Lu Z. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Florence, Italy: Association for Computational Linguistics, 2019:58–65.

224. Johnson HL, Cohen KB, Baumgartner Jr WA, et al. Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. In: *Biocomputing 2006*. World Scientific, 2006:28–39.

225. Li H, Chen Q, Tang B, et al. CNN-based ranking for biomedical entity normalization. BMC Bioinformatics 2017;18:385.

226. Liu H and Xu Y. A Deep Learning Way for Disease Name Representation and Normalization. In: *Natural Language Processing and Chinese Computing*. Ed. by Huang X, Jiang J, Zhao D, Feng Y, and Hong Y. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2018:151–7.

227. Tutubalina E, Miftahutdinov Z, Nikolenko S, and Malykh V. Medical concept normalization in social media posts with recurrent neural networks. Journal of Biomedical Informatics 2018;84:93–102.

228. Madankar M, Chandak MB, and Chavhan N. Information Retrieval System and Machine Translation: A Review. Procedia Computer Science. 1st International Conference on Information Security & Privacy 2015 2016;78:845–50.

229. Bahdanau D, Cho K, and Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473 [cs, stat] 2016.

230. Funk CS. RECOGNITION AND NORMALIZATION OF TERMINOLOGY FROM LARGE BIOMEDICAL ONTOLOGIES AND THEIR APPLICATION FOR PHARMACOGENE AND PROTEIN FUNCTION PREDICTION. PhD thesis. 2015.

231. Cambria E and White B. Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. IEEE Computational Intelligence Magazine 2014;9:48–57.

232. Chiu B, Crichton G, Korhonen A, and Pyysalo S. How to Train good Word Embeddings for Biomedical NLP. In: *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Berlin, Germany: Association for Computational Linguistics, 2016:166–74.

233. William A. Baumgartner Jr. The Colorado Richly Annotated Full-Text (CRAFT) Corpus.

234. Bossy R, Golik W, Ratkovic Z, Bessieres P, and Nédellec C. Bionlp shared task 2013–an overview of the bacteria biotope task. In: *Proceedings of the BioNLP shared task 2013 workshop*. 2013:161–9.

235. BioFrontiers IT. Fiji User Guide.

236. Goyal A, Gupta V, and Kumar M. Recent Named Entity Recognition and Classification techniques: A systematic review. Computer Science Review 2018;29:21–43.

237. Ogren P. Improving Syntactic Coordination Resolution using Language Modeling. In: *Proceedings of the NAACL HLT 2010 Student Research Workshop*. Los Angeles, CA: Association for Computational Linguistics, 2010:1–6.

238. Reimers N and Gurevych I. Optimal Hyperparameters for Deep LSTM-Networks for Sequence Labeling Tasks. arXiv:1707.06799 [cs] 2017.

239. Bushaev V. Understanding RMSprop-faster neural network learning. Towards Data Science 2018.

240. Hinton G, Srivastava N, and Swersky K. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent. Cited on 2012;14:2.

241. Kingma, D.P., Ba, L.J., and Amsterdam Machine Learning lab (IVI, FNWI). Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations (ICLR)*. arXiv.org, 2015.

242. Demuth HB, Beale MH, De Jess O, and Hagan MT. Neural Network Design. 2nd. Stillwater, OK, USA: Martin Hagan, 2014.

243. Gu Y, Tinn R, Cheng H, et al. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. arXiv:2007.15779 [cs] 2020.

244. Friedman C, Rindflesch TC, and Corn M. Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. Journal of Biomedical Informatics 2013;46:765–73.

245. Chafe WL. Meaning and the Structure of Language. The University of Chicago Press, Chicago, Illinois 60637 ($10, 1970.

246. Algeo J and Butcher CA. The Origins and Development of the English Language. Cengage Learning, 2013.

247. Gelderen E. A history of the English language. A History of the English Language 2014:1–358.

248. Hall D, Berg-Kirkpatrick T, and Klein D. Sparser, better, faster GPU parsing. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2014:208–17.

249. Strubell E, Ganesh A, and McCallum A. Energy and Policy Considerations for Deep Learning in NLP. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019:3645–50.

250. Bojanowski P, Grave E, Joulin A, and Mikolov T. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics 2017;5:135–46.

251. Mehrabi S, Krishnan A, Roch AM, et al. Identification of Patients with Family History of Pancreatic Cancer - Investigation of an NLP System Portability. Studies in health technology and informatics 2015;216:604–8.

252. Caporaso JG, Deshpande N, Fink JL, Bourne PE, Cohen KB, and Hunter L. Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. In: *Biocomputing 2008*. World Scientific, 2008:640–51.

253. Hoehndorf R, Schofield PN, and Gkoutos GV. The role of ontologies in biological and biomedical research: a functional perspective. Briefings in bioinformatics 2015;16:1069–80.

254. Grishman R. Information Extraction. IEEE Intelligent Systems 2015;30:8–15.

255. Bromberger, Sylvain. On What We Know We Don't Know. Explanation, Theory, Linguistics, and How Questions Shape Them. Philosophical Books 1994;35:38–9.

256. Martin, J.R. and White, P.R.R. The Language of Evaluation - Appraisal in English. Palgrave Macmillan UK, 2005.

257. Rubin V. Identifying Certainty in Texts. PhD thesis. 2006.

258. Schiebinger L and Proctor RN, eds. Agnotology: The Making and Unmaking of Ignorance. Stanford: Stanford University Press, 2008.

259. Perez E, Lewis P, Yih Wt, Cho K, and Kiela D. Unsupervised question decomposition for question answering. arXiv preprint arXiv:2002.09758 2020.

260. Shardlow M, Batista-Navarro R, Thompson P, Nawaz R, McNaught J, and Ananiadou S. Identification of research hypotheses and new knowledge from scientific literature. BMC medical informatics and decision making 2018;18:46.

261. Patrick J and Li M. An ontology for clinical questions about the contents of patient notes. Journal of Biomedical Informatics 2012;45:292–306.

262. Bongelli R, Riccioni I, Burro R, and Zuczkowski A. Writers' uncertainty in scientific and popular biomedical articles. A comparative analysis of the British Medical Journal and Discover Magazine. PLOS ONE 2019;14:e0221933.

263. Chen C, Song M, and Heo GE. A Scalable and Adaptive Method for Finding Semantically Equivalent Cue Words of Uncertainty. Journal of Informetrics 2018;12:158–80.

264. Konstantinova, Natalia. Review of Relation Extraction Methods: What Is New Out There? In: *International Conference on Analysis of Images, Social Networks and Texts*. Vol. 436. Springer, Cham, 2014:15–28.

265. Ogren PV. Knowtator: a protégé plug-in for annotated corpus construction. In: Association for Computational Linguistics, 2006:273–5.

266. Harrison Pielke-Lombardo. Knowtator-2.0: A text annotation plugin for Protege 5+. 2018.

267. Knublauch H, Fergerson RW, Noy NF, and Musen MA. The Protégé OWL plugin: An open development environment for semantic web applications. In: *International semantic web conference*. Springer, 2004:229–43.

268. Dalianis H. Evaluation Metrics and Evaluation. In: *Clinical Text Mining*. Cham: Springer International Publishing, 2018:45–53.

269. Pustejovsky J and Stubbs A. Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications. "O'Reilly Media, Inc.", 2012.

270. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. The Journal of Machine Learning Research 2011;12:2825–30.

271. Chollet, Francois and others. Keras. 2015.

272. Dreiseitl S and Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics 2002;35:352–9.

273. Chien C. Batch size selection for the batch means method. In: *Proceedings of Winter Simulation Conference*. IEEE, 1994:345–52.

274. Publication Types.

275. Chen Q, Allot A, and Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic Acids Research 2021;49:D1534–D1540.

276. Callahan TJ. PheKnowLator Human Disease Knowledge Graphs – Class-based Knowledge Model with Inverse Relations and OWL-NETS Abstraction. 2021.

277. Callahan TJ. Phenotype Knowledge Translator: A FAIR Ecosystem for Representing Large-Scale Biomedical Knowledge. Tech. rep. Zenodo, 2021.

278. Sherman BT, Hao M, Qiu J, et al. DAVID: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Research 2022:gkac194.

279. Lee S and Lee DK. What is the proper way to apply the multiple comparison test? Korean Journal of Anesthesiology 2018;71:353–60.

280. Costanzo P, Santini A, Fattore L, Novellino E, and Ritieni A. Toxicity of aflatoxin B1 towards the vitamin D receptor (VDR). Food and Chemical Toxicology 2015;76:77–9.

281. Sánchez-Hernández D, Anderson GH, Poon AN, et al. Maternal fat-soluble vitamins, brain development, and regulation of feeding behavior: an overview of research. Nutrition Research 2016;36:1045–54.

282. Harvey NC, Holroyd C, Ntani G, et al. Vitamin D supplementation in pregnancy: a systematic review. Health Technology Assessment (Winchester, England) 2014;18:1–190.

283. Saggese G, Vierucci F, Prodam F, et al. Vitamin D in pediatric age: consensus of the Italian Pediatric Society and the Italian Society of Preventive and Social Pediatrics, jointly with the Italian Federation of Pediatricians. Italian Journal of Pediatrics 2018;44:51.

284. Saraf R, Morton SM, Camargo CA, and Grant CC. Global summary of maternal and newborn vitamin D status – a systematic review. Maternal & Child Nutrition 2015;12:647–68.

285. Ojo O, Weldon SM, Thompson T, and Vargo EJ. The Effect of Vitamin D Supplementation on Glycaemic Control in Women with Gestational Diabetes Mellitus: A Systematic Review and Meta-Analysis of Randomised Controlled Trials. International Journal of Environmental Research and Public Health 2019;16:1716.

286. Hanieh S, Ha TT, Simpson JA, et al. Maternal vitamin D status and infant outcomes in rural Vietnam: a prospective cohort study. PloS one 2014;9:e99005.

287. Angelidou A, Asadi S, Alysandratos KD, Karagkouni A, Kourembanas S, and Theoharides TC. Perinatal stress, brain inflammation and risk of autism-Review and proposal. BMC Pediatrics 2012;12:89.

288. Wei CH, Allot A, Leaman R, and Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Research 2019;47:W587–W593.

289. National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, and Government-University-Industry Research Roundtable. Examining the Mistrust of Science: Proceedings of a Workshop—in Brief. The National Academies Collection: Reports funded by National Institutes of Health. Washington (DC): National Academies Press (US), 2017.

290. Nadeem R. Americans' Trust in Scientists, Other Groups Declines. 2022.

291. Nasr N. Overcoming the discourse of science mistrust: how science education can be used to develop competent consumers and communicators of science information. Cultural Studies of Science Education 2021;16:345–56.

292. Baron RJ and Berinsky AJ. Mistrust in Science — A Threat to the Patient–Physician Relationship. New England Journal of Medicine 2019;381. Ed. by Malina D:182–5.

293. Gatt A and Krahmer E. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. Journal of Artificial Intelligence Research 2018;61:65–170.

294. Dale R. Natural language generation: The commercial state of the art in 2020. Natural Language Engineering 2020;26:481–7.

295. OpenAI. Chatgpt: A large-scale generative pre-training model for conversation. 2022.

296. Uc-Cetina V, Navarro-Guerrero N, Martin-Gonzalez A, Weber C, and Wermter S. Survey on reinforcement learning for language processing. Artificial Intelligence Review 2022.

297. Thomas L. A Long Line of Cells Collected Essays. 1990.