UNIVERSITÄT
DES
SAARLANDES

Saarland University
Faculty of Mathematics and Computer Science
Department of Computer Science

Knowledge-driven Entity Recognition and Disambiguation in Biomedical Text

Amy Siu

A dissertation submitted towards the degree
Doctor of Engineering
of the Faculty of Mathematics and Computer Science of
Saarland University

Saarbrücken, May 2017

Dean:	Prof. Frank-Olaf Schreyer
Colloquium:	4 September 2017

Examination Board

Supervisor and First Reviewer:	Prof. Gerhard Weikum
Second Reviewer:	Prof. Klaus Berberich
Third Reviewer:	Prof. Ulf Leser
Chairman:	Prof. Dietrich Klakow
Research Assistant:	Dr. Luciano Del Corro

Abstract

Entity recognition and disambiguation (ERD) for the biomedical domain are notoriously difficult problems due to the variety of entities and their often long names in many variations. Existing works focus heavily on the molecular level in two ways. First, they target scientific literature as the input text genre. Second, they target single, highly specialized entity types such as chemicals, genes, and proteins. However, a wealth of biomedical information is also buried in the vast universe of Web content. In order to fully utilize all the information available, there is a need to tap into Web content as an additional input. Moreover, there is a need to cater for other entity types such as symptoms and risk factors since Web content focuses on consumer health.

The goal of this thesis is to investigate ERD methods that are applicable to all entity types in scientific literature as well as Web content. In addition, we focus on under-explored aspects of the biomedical ERD problems – scalability, long noun phrases, and out-of-knowledge base (OOKB) entities.

This thesis makes four main contributions, all of which leverage knowledge in UMLS (Unified Medical Language System), the largest and most authoritative knowledge base (KB) of the biomedical domain. The first contribution is a fast dictionary lookup method for entity recognition that maximizes throughput while balancing the loss of precision and recall. The second contribution is a semantic type classification method targeting common words in long noun phrases. We develop a custom set of semantic types to capture word usages; besides biomedical usage, these types also cope with non-biomedical usage and the case of generic, non-informative usage. The third contribution is a fast heuristics method for entity disambiguation in MEDLINE abstracts, again maximizing throughput but this time maintaining accuracy. The fourth contribution is a corpus-driven entity disambiguation method that addresses OOKB entities. The method first captures the entities expressed in a corpus as latent representations that comprise in-KB and OOKB entities alike before performing entity disambiguation.

Kurzfassung

Die Erkennung und Disambiguierung von Entitäten für den biomedizinischen Bereich stellen, wegen der vielfältigen Arten von biomedizinischen Entitäten sowie deren oft langen und variantenreichen Namen, große Herausforderungen dar. Vorhergehende Arbeiten konzentrieren sich in zweierlei Hinsicht fast ausschließlich auf molekulare Entitäten. Erstens fokussieren sie sich auf wissenschaftliche Publikationen als Genre der Eingabetexte. Zweitens fokussieren sie sich auf einzelne, sehr spezialisierte Entitätstypen wie Chemikalien, Gene und Proteine. Allerdings bietet das Internet neben diesen Quellen eine Vielzahl an Inhalten biomedizinischen Wissens, das vernachlässigt wird. Um alle verfügbaren Informationen auszunutzen besteht der Bedarf weitere Internet-Inhalte als zusätzliche Quellen zu erschließen. Außerdem ist es auch erforderlich andere Entitätstypen wie Symptome und Risikofaktoren in Betracht zu ziehen, da diese für zahlreiche Inhalte im Internet, wie zum Beispiel Verbraucherinformationen im Gesundheitssektor, relevant sind.

Das Ziel dieser Dissertation ist es, Methoden zur Erkennung und Disambiguierung von Entitäten zu erforschen, die alle Entitätstypen in Betracht ziehen und sowohl auf wissenschaftliche Publikationen als auch auf andere Internet-Inhalte anwendbar sind. Darüber hinaus setzen wir Schwerpunkte auf oft vernachlässigte Aspekte der biomedizinischen Erkennung und Disambiguierung von Entitäten, nämlich Skalierbarkeit, lange Nominalphrasen und fehlende Entitäten in einer Wissensbank.

In dieser Hinsicht leistet diese Dissertation vier Hauptbeiträge, denen allen das Wissen von UMLS (Unified Medical Language System), der größten und wichtigsten Wissensbank im biomedizinischen Bereich, zu Grunde liegt. Der erste Beitrag ist eine schnelle Methode zur Erkennung von Entitäten mittels Lexikonabgleich, welche den Durchsatz maximiert und gleichzeitig den Verlust in Genauigkeit und Trefferquote (precision and recall) balanciert. Der zweite Beitrag ist eine Methode zur Klassifizierung der semantischen Typen von Nomen, die sich auf gebräuchliche Nomen von langen Nominalphrasen richtet und auf einer selbstentwickelten Sammlung von semantischen Typen beruht, die die Verwendung der Nomen erfasst. Neben biomedizinischen können diese Typen auch nicht-biomedizinische und allgemeine, informationsarme Verwendungen behandeln. Der dritte Beitrag ist eine schnelle Heuristikmethode zur Disambiguierung von Entitäten in MEDLINE Kurzfassungen, welche den Durchsatz maximiert, aber auch die Genauigkeit erhält. Der vierte Beitrag ist eine korpusgetriebene Methode zur Disambiguierung von Entitäten, die speziell fehlende Entitäten in einer Wissensbank behandelt. Die Methode wandelt erst die Entitäten, die in einem Textkorpus ausgedrückt aber nicht notwendigerweise in einer Wissensbank sind, in latente Darstellungen um und führt anschließend die Disambiguierung durch.

Summary

In recent years, the amount of biomedical information disseminated in textual form is growing at an ever increasing pace. In response, text mining has emerged as a major research area to harness this information. A text mining system consists of a pipeline of tasks, of which entity recognition and disambiguation (ERD) are indispensable ones early in the pipeline. However, biomedical ERD is notoriously difficult since there are many entity types such as chemicals, genes, proteins, disease, and symptoms, etc., and each type features its own nomenclature. Furthermore, biomedical entities encompass proper nouns as well as compound noun phrases, many of which are long with numerous variations. Most existing works focus on named entities in individual entity types, as well as on PubMed scientific publications. As soon as text mining taps into Web content such as encyclopedic health portals and patient discussion forums, existing methods fall short of handling entity types beyond the molecular level such as symptoms and lifestyle risk factors. Longer noun phrases are mostly neglected, and, to the best of our knowledge, there are no ERD tools that can cope with large-scale corpora.

This thesis makes four contributions that address these limitations. A common theme of the contributions is leveraging of a knowledge base (KB), namely UMLS (Unified Medical Language System). This preeminent KB of the biomedical domain contains entity names, definitions, semantic types, and entity-entity relations that are essential ingredients throughout this thesis. Below we summarize each contribution.

Fast entity recognition. Of the existing entity recognition (ER) methods, few are applicable to all entity types. In order to devise a method that addresses all types, we turn to a comprehensive dictionary covering all aspects of biomedicine, and devise a dictionary lookup method. That dictionary is UMLS, which contains 3.4 million entities, and is rich in lexical variants of entity names. By minimizing time-consuming natural language processing (NLP), and by speeding lookups up with Locality Sensitive Hashing and MinHash, our method aims to maximize throughput while balancing the loss of precision and coverage. When compared to MetaMap, the de facto standard tool, our method is two orders of magnitude faster while maintaining comparable precision, though at the cost of losing 13% coverage.

Semantic type classification of common words. Long noun phrases are ubiquitous in biomedical text, but they are largely disregarded. We observe that not all words in a noun phrase are equally information-bearing. We also observe that some words even carry non-biomedical meanings; this phenomenon becomes more commonplace as we go beyond scientific literature and venture into Web content. For

information extraction tasks, it is important to consider common nouns only when they carry crucial biomedical meaning. Therefore we devise a method to classify the semantic type of common nouns. Besides biomedical meanings, the semantic types explicitly include the negative case when nouns are used in a generic, non-informative sense, as well as non-biomedical meanings. We demonstrate the usefulness of the method with 50 common nouns and a custom set of fine-grained semantic types.

Fast entity disambiguation in topically annotated texts. Most existing entity disambiguation (ED) methods focus on single entity types such as chemicals, genes, proteins, and diseases. Of the few methods that do address all entity types, MetaMap is the only practical option as it is publicly available with an easy setup. Unfortunately, MetaMap is known to employ heavy NLP machinery, and its disambiguation module has limited functionality. The former renders the tool unsuitable for large-scale use, and the latter negatively affects disambiguation quality. This spurs us to devise an entity-type-agnostic, heuristics method for topically annotated texts, such as MEDLINE abstracts. The method first exploits the expert-assigned indexing with MeSH (Medical Subject Headings) terms and the presence of unambiguous entity names in UMLS, to determine unambiguous text mentions. These intermediate results then become extra cues that can be leveraged by ED methods. Experiments demonstrate that our method is one order of magnitude faster as well as more accurate than MetaMap.

Corpus-driven entity discovery and disambiguation. Existing ED methods are KB-driven – a text mention is disambiguated to one of the candidates selected from a KB. These methods either implicitly assume that the correct entity is always one of the candidates, or they declare that the text mention maps to a NIL placeholder when no correct candidate is found. The implicit assumption is unsatisfactory since no KB can be complete; there are always emerging entities waiting to be incorporated into the KB. The NIL placeholder is also unsatisfactory since no further information is available about the description-less, out-of-KB (OOKB) entity. Contrary to existing approaches, we devise a corpus-driven method. The method first discovers the latent topics an ambiguous entity name expresses in a corpus; each topic describes a latent entity, whether it exists in the KB or not. Then the method maps a text mention to a latent entity, which is in turn mapped to an in-KB or OOKB entity. To the best of our knowledge, our method is the first attempt in the biomedical domain to represent OOKB entities via their latent descriptions. We further demonstrate the applicability of the method for the political domain.

Acknowledgments

Many people helped me reach the finish line. To all of them I say, “Thank you!”

The biggest thank you goes to my advisor Gerhard Weikum; your patient tutelage single-handedly made this thesis possible. I also thank Cecilia Arighi, Vijay Shanker, and Cathy Wu at the University of Delaware for kickstarting me in biomedical text mining. These wonderful teachers taught me a lot about research (and other things).

Members of the Databases and Information Systems Group made the workplace a fun place. Thank you, Patrick the teammate, for your ideas, collaboration, and advice on work-related things and beyond. Thank you, Dat and Sairam the officemates, for being more than just officemates. All of my colleagues gave me a lot of moral support.

Thank you to my family – parents Franko and Shirley, brother Webster, husband Timo, and Cookie Cat – for even more moral support. One needs a lot of it for the long-distance march of pursuing a PhD.

My gratitude also goes to the Max Planck Institute for Informatics for funding this thesis.

Contents

Chapter 1

Introduction

1.1 Motivation

Biomedical information is discovered and disseminated at an extraordinary pace, and much of this dissemination is accomplished by publishing unstructured text, or free text. Two communities spreadhead this trend. The first community is the scientific community, who publish their research results in scientific literature. PubMed, the preeminent indexing service dedicated to the biomedical domain, contains to date over 26 million citations, and continues to grow at over 1 million new citations per year. The second community is the online community, who comprises experts and medical doctors, as well as patients and layman consumers. This community publishes a range of contents, from professional resources such as UpToDate.com that are written by experts for experts, to encyclopedic health portals such as Mayo Clinic and Medline Plus that are written by experts for patients, to online discussion forums in which only patients participate. Buried within these two universes of texts is a wealth of knowledge, but due to their sheer size and unstructured nature, automated methods are required to extract this knowledge.

The biomedical natural language processing (BioNLP) community employs text mining as a solution to automatically analyze the aforementioned texts. Text mining spans a range of information extraction (IE) tasks, from extracting protein-protein interactions, to drug-disease relations, to pharmacogenomics networks. Common to all these tasks, entity recognition and disambiguation (ERD) are the indispensable first steps in a text mining pipeline. Not only does the quality of downstream processes depend on the quality of entities extracted, these downstream analyses may not even begin without the entities as input. Besides fully automated text mining pipelines, there are also semi-automated ones such as indexing PubMed citations and curating knowledge bases, where entities are first automatically extracted to be manually selected and refined by experts. There is a large body of existing work addressing ERD; however, limitations exist.

Coverage in entity types. Biomedical text mining has to date focused heavily on extracting information at the molecular level, such as information about proteins, genes, chemicals, drugs, their derivatives such as protein sequences and gene mutations, and their interactions such as protein pathways. This focus leads to two norms. First, existing ERD works target named entities, which are well cataloged in dictionaries. Common nouns, long noun phrases, and in general entities not found in a dictionary are largely neglected despite being ubiquitous in biomedical text. Second, the entity types addressed revolve around the few entity types corresponding to said proteins, genes, chemicals, and so on. Furthermore, the majority of existing ERD methods specialize in only one entity type. There is relatively little effort to develop methods that address all entity types simultaneously. The most prominent solution is MetaMap, the method implemented as a software tool that performs dictionary-based ERD. While it is the de facto standard tool for biomedical text mining, its limitations in processing speed and disambiguation power render it infeasible in PubMed-scale projects.

Coverage in text genres. Focusing on IE tasks at the molecular level naturally in turn focuses the source texts to scientific literature, where the relevant information abounds. Consequently, the abstracts and full-length articles curated by PubMed have been the de facto standard corpus for biomedical text mining. The vast universe of Web contents mentioned above has been largely neglected. While much information is contained only in these non-scientific texts, they also contain fragments of text that stray away from a strictly biomedical focus. As soon as one mines texts beyond PubMed, it becomes mandatory to distinguish the non-biomedical content in order to disregard them. Since existing ERD methods are developed only with scientific literature in mind, they implicitly assume that all the entities in text are of a biomedical nature, and do not address the presence of non-biomedical entities.

1.2 Problem Statement

In this thesis, we consider the problems of entity recognition and disambiguation for the biomedical domain. Entity recognition is the task of identifying, from unstructured text, mentions that refer to entities. Entity disambiguation is the task of selecting for these mentions the correct underlying entity from a pool of candidates. The proposed solutions should address the following requirements.

- The solution should go beyond single proper nouns for named entities and address compound noun phrases also.
- The solution should be applicable across all sub-domains of biomedicine. We intentionally depart from the norm of specializing in selected sub-domains, and opt to investigate sub-domain-agnostic approaches.
- The solution should be applicable to both scientific texts and Web content. Where the text contains non-biomedical entities, the solution should be able to distinguish them as such.
- The solution should be fast, so that it is feasible to apply it to large-scale corpora.

1.3 Challenges

In order to propose solutions addressing the above requirements, the following challenges must be overcome.

Diversity of entity names. Biomedical entity names are notoriously difficult to recognize and disambiguate because they are often long with many variations; this applies to named entities and noun phrases alike. A comprehensive ERD method must also go beyond nouns and consider verbs, as verbs also provide crucial cues and can be leveraged as an entity also.

Incomplete dictionary. No dictionary can be 100% complete with all named entities, and further cannot be expected to catalog biomedical noun phrases in all their myriad variations. UMLS (Unified Medical Language System), the largest metathesaurus boasting 3.4 million entities sourced from 199 individual dictionaries (as of version 2016AB), still lacks entries such as *meds* the colloquial term for medicines, and *arm* the physical branch of a protein molecule.

Heterogeneous assortment of resources. The preeminent lexical resource for biomedical text mining, UMLS, is actually a potpourri of many heterogeneous resources. Each of the 199 contributing dictionaries specializes on its sub-domain, with entities defined at different levels of granularity, and complete with dictionary-specific entity-entity relationships. Overarching this motley collection of entities is the UMLS semantic network, which is in itself three separate types of information: a set of 133 semantic types that form a taxonomy; 15 semantic groups that do not align with the taxonomy; and relationships between semantic types. Since each entity in UMLS is assigned at least one semantic type, any pair of entities can also be viewed through the lens of the semantic network, that is, orthogonal to the dictionary-specific relationships. Additional resources such as the BioLexicon and the symptom terms in OMIM (Online Mendelian Inheritance of Man) further add to this tangled landscape. Putting all these together, judicious selection of resources that best suits the needs of a specific task is required.

Transcending sub-domains. Entities in each sub-domain has different nomenclatures. For instance, genes are often a mixture of English letters and numerical digits, while chemicals are often formulaic expressions involving chemical symbols. In order to go beyond sub-domains, we must also go beyond, for instance, the lexicographic features that apply only to one sub-domain. In other words, we must seek approaches that generalize to all sub-domains.

Transcending text genres. Different text genres have different language styles. Scientific literature, especially abstracts, are terse with convoluted sentence structure and specialized jargon. Web contents, on the other hand, have a more relaxed, everyday English style. In addition, biomedically themed Web contents are often "contaminated" with non-biomedical content. We must seek approaches that work for a range of language styles and can compensate for the presence of non-biomedical content.

Minimizing manual intervention. Manual intervention, such as labeling data samples and selecting seed data, is expensive. Where a method cannot be completely automated, we aim to at least minimize the manual intervention needed.

Maintaining quality while being fast. Processing speed and output quality are goals at odds with each other; higher speed is achieved by sacrificing quality, and vice versa. As mentioned, high quality entities are crucial for text mining. Therefore, in the pursuit of fast processing speed, we need to at least maintain quality comparable to a more sophisticated but slower method. Specifically for ERD tasks, that translates to maintaining good precision and good recall while achieving high throughput at PubMed-scale.

1.4 Contributions

A common theme of the various methods devised in this thesis is going beyond conventional specializations on corpora and sub-domains. Another theme is that each method leverages knowledge in UMLS according its respective goal. Specifically, we make the following contributions.

Fast entity recognition. To address the lack of entity recognition software tool that can process texts fast enough to handle PubMed-scale corpora, we devise and implement a dictionary lookup method that uses UMLS as the underlying dictionary. The method also uses Locality Sensitive Hashing with MinHash to quickly estimate string similarity between text mentions and dictionary entries. When compared to MetaMap, this method achieves comparable precision at a throughput that is two orders of magnitude faster, though at the expense of losing 13% in entity coverage. The speedup is the critical improvement that makes other large-scale text mining projects possible, namely, KnowLife (knowledge base construction) [41] and DeepLife (search and analytics application for up-to-date health information) [40]. The resulting implementation has been released as open source software.

Semantic type classification of common words. As an effort towards disambiguating long and complex noun phrases, we propose to disambiguate the semantic types of common nouns found in such phrases. We develop fine-grained, custom semantic types that encompass biomedical and non-biomedical types, as well as the non-informative type that denotes nouns used only in a generic way without carrying critical information. Using label propagation, a semi-supervised graph-based method, we demonstrate that only a small percentage of labeled seed nodes suffices to successfully label the rest of the nodes, hence minimizing manual effort in identifying seeds. The node-node relatedness in the graph reflects the similarity between the corresponding noun phrases, which is in turn derived by leveraging UMLS semantic types. To the best of our knowledge, this contribution is the first work in BioNLP that addresses general-domain content mingled in a biomedical-themed document.

Fast entity disambiguation in topically annotated texts. We devise an automatic and light-weight entity disambiguation method that exploits two key characteristics of biomedical resources: the indexing terms, or MeSH (Medical Subject Headings) terms, assigned by experts in all PubMed citations, and the lexical richness as well as heterogeneity of UMLS. Using two heuristics, the method first identifies anchors, or non-ambiguous text mentions. Then using the anchors and five further heuristics based on linguistic cues, co-occurrence statistics, and prior distributions of entities estimated from UMLS, the method disambiguates the remaining mentions. When compared to MetaMap's disambiguation module, our method is up to 11% more accurate at a throughput one order of magnitude faster. Not only is this method particularly amenable to PubMed the de facto corpus, it is also sub-domain-agnostic. The implementation of the method has also been released as open source software.

Corpus-driven entity discovery and disambiguation. We address the incompleteness of dictionary by first performing entity discovery in a corpus-driven approach. Using dimensionality reduction, we model a corpus of text snippets all containing an ambiguous entity name as a low-dimension latent topic space. Each latent topic corresponds to an entity expressed in the corpus; in other words, each entity has a latent description. Such a latent entity can either be mapped to an entity in the dictionary, thereby achieving entity disambiguation, or can be declared out of knowledge base (OOKB). To the best of our knowledge, this work is the first in the biomedical domain to cater for OOKB entities by representing them via their latent descriptions. We further demonstrate the generalizability of the method via experiments in the politics domain.

1.5 Publications

Parts of this thesis have been published or are in the process of attaining publication. We list here the relationships between the contributions and their publications:

Contribution	**Publication title and authors**	**Publication venue or under review**
Fast entity recognition	Fast entity recognition in biomedical text [187] Amy Siu, Dat Ba Nguyen, Gerhard Weikum	Workshop on Data Mining for Healthcare (DMH) at the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2013
Semantic type classification of common words	Semantic type classification of common words in biomedical noun phrases [188] Amy Siu, Gerhard Weikum	BioNLP Workshop at the Annual Meeting of the Association for Computational Linguistics (ACL), 2015
Fast entity disambiguation in topically annotated texts	Disambiguation of entities in MEDLINE abstracts by combining MeSH terms with knowledge [186] Amy Siu, Patrick Ernst, Gerhard Weikum	BioNLP Workshop at the Annual Meeting of the Association for Computational Linguistics (ACL), 2016
Corpus-driven entity discovery and disambiguation	Corpus-driven entity discovery and disambiguation Amy Siu, Patrick Ernst, Gerhard Weikum	Under review

Table 1.1: Publications directly resulting from contributions in this thesis

Methods developed in this thesis have also contributed to further publications, for which the author of this thesis is a co-author:

Publication title and authors	**Publication venue or under review**
KnowLife: A knowledge graph for health and life sciences [39] Patrick Ernst, Cynthia Meng, Amy Siu, Gerhard Weikum	System Demonstration at the International Conference on Data Engineering (ICDE), 2014
KnowLife: A versatile approach to constructing a knowledge graph for biomedical sciences [41] Patrick Ernst, Amy Siu, Gerhard Weikum	BMC Bioinformatics, 2015
DeepLife: An entity-aware search, analytics and exploration platform for health and life sciences [40] Patrick Ernst, Amy Siu, Dragan Milchevski, Johannes Hoffart, Gerhard Weikum	System Demonstration at the Annual Meeting of the Association for Computational Linguistics (ACL), 2016
HighLife: Higher-arity fact harvesting Patrick Ernst, Amy Siu, Gerhard Weikum	Under review

Table 1.2: Publications using methods devised in this thesis

Finally, we also list here other prior publications of the author of this thesis:

Publication title and authors	**Publication venue**
eFIP: A tool for mining functional impact of phosphorylation from literature [4] Cecilia N. Arighi, Amy Siu, Catalina O. Tudor, Jules A. Nchoutmboube, Cathy H. Wu, Vijay K. Shanker	Book chapter in Methods in Molecular Biology – Bioinformatics for Comparative Proteomics, 2011
Knowledge discovery on incompatibility of medical concepts [57] Adam Grycner, Patrick Ernst, Amy Siu, Gerhard Weikum	International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), 2013
Findings of the WMT 2017 biomedical translation shared task [85] Antonio Jimeno Yepes, Aurélie Névéol, Mariana Neves, Karin Verspoor, Ondřej Bojar, Arthur Boyer, Cristian Grozea, Barry Haddow, Madeleine Kittner, Yvonne Lichtblau, Pavel Pecina, Roland Roller, Rudolf Rosa, Amy Siu, Philippe Thomas, Saskia Trescher	Conference on Machine Translation (WMT) at the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017

Table 1.3: Prior publications of the author of this thesis

1.6 Thesis Outline

The remainder of this thesis elaborates on the aforementioned contributions. In Chapter 2, we begin by visiting background concepts, preliminaries, and characteristics of the biomedical texts that lead to challenges specific to the domain. Chapter 3 presents the state of the art in both the general and biomedical domains. The next four chapters each presents one contribution: Chapter 4 presents the dictionary-lookup method for entity recognition; Chapter 5 presents the classification of semantic types in long and complex noun phrases; Chapter 6 presents the disambiguation of entities in PubMed abstracts via heuristics; and Chapter 7 presents the discovery as well as disambiguation of entities using a corpus-driven approach. Finally, in Chapter 8 we summarize all contributions presented in this thesis and give an outlook on future directions.

Chapter 2

Background

2.1 Domain and Sub-Domain

Domain. A *domain*, according to the Oxford English Dictionary and the Merriam-Webster Dictionary, is "a sphere of knowledge." In terms of research, a domain is often the topic of a research community. Biomedicine, computer science, linguistics, politics, and psychology are all example domains.

Sub-domain. Just as a research community has a subset of researchers focusing on a narrower topic within the larger, overarching topic, so a domain also has *sub-domains*. Within the biomedicine domain, in particular, are the sub-domains of biochemistry, diseases (of animals, humans, and plants), drugs, genetics, molecular biology, cell biology, systems biology, and more. In biomedical literature, sub-domains are also referred to as disciplines, fields, and specialties [28].

General domain. The term *general domain* describes when no particular domain is in focus.

Using Wikipedia[1] as an example, it is an encyclopedia in the general domain since it does not focus on any particular domain. The portals therein, however, such as the Biology, Medicine, Linguistics, Political Science, and Psychology Portals, correspond to domains. Some portals such as the Alternative Medicine Portal and the Dentistry Portals correspond to sub-domains as they are even more specialized than the Medicine Portal.

2.2 Repositories of Knowledge

Knowledge, be it in the general or biomedical domain, can be stored in different kinds of repositories.

Knowledge base. A *knowledge base* (KB) stores knowledge in some structured manner. A common structure is the triple, where a subject–predicate–object construct represents a binary relation between two entities. Major knowledge bases in

[1] `wikipedia.org`

the general domain such as DBpedia [104], Freebase [16] (now discontinued), Wikidata [216], and YAGO [199], as well UMLS (United Medical Language System) [15], the largest and most authoritative knowledge base in the biomedical domain, all follow this triple format.

Ontology. An *ontology* mandates a formal delineation between concepts as classes, and objects as instances of classes. Besides relations, additional formal constructs such as attributes, axioms, and rules enable logical reasoning to be performed upon the classes and objects [58].

Taxonomy. A *taxonomy* aims at arranging entities in categories in a tree-like manner, such that each category is subsumed by a more general category. Another way to characterize this subsumption is the is-A relation; for instance, lung cancer is a kind of cancer, which in turn is a kind of disease.

Dictionary. A *dictionary*, also called controlled vocabulary [27] and terminology [55], aims at exhaustively collecting all the words and names used in a domain or sub-domain.

The precise characterization of various kinds of knowledge repositories is a matter of ongoing discussion [55, 58]. What is certain, however, is that knowledge resources in the biomedical domain often exhibit characteristics of multiple kinds of repositories. OMIM (Online Mendelian Inheritance of Man) [63] is a typical example: Primarily, it is a repository of human genetic disorders and the relevant genetic information. However, the detailed catalog of disease symptoms have often been leveraged as a dictionary [31, 137]. Similarly, while MeSH (Medical Subject Headings)[2] is primarily a taxonomy and the NCI (National Cancer Institute) Thesaurus [185] a dictionary, they have been leveraged beyond their categories [35, 137] and dictionary entries [49], respectively.

Many sub-domains have their own ontologies, such as the Foundational Model of Anatomy [168], the Human Disease Ontology [174], the Drosophila Phenotype Ontology [142], among many others. The OBO Foundry[3] provides an online, one-stop service to search or browse the ontologies individually. Many of these ontologies as well as other knowledge resources – altogether 199 of them in the 2016 version – have been combined by human experts to produce UMLS, the authoritative metathesaurus for the biomedical domain. The major biomedical repositories have been surveyed and their contents compared by [76].

2.3 Entity Discovery, Recognition, and Disambiguation

Entity discovery. The task of determining whether a entity expressed in some text is not yet registered in some KB is *entity discovery*. Such "newly discovered" entities are often new from a temporal point of view, such as emerging entities, but that is

[2] `www.nlm.nih.gov/mesh`

[3] `www.obofoundry.org`

not the default case. Other entities may be absent from a KB because the entities belong to the "long tail" and therefore are neglected or even intentionally omitted.

Entity recognition. Given a piece of text, the words that express an occurrence of an entity is a text mention. *Entity recognition*, also known as entity tagging, is the task of identifying text mentions. Notice that entity recognition does not require a KB, though leveraging one for dictionary lookup is a common approach.

Entity disambiguation. A text mention is *ambiguous* when it may refer to multiple candidate entities; for instance, "FISH" may refer to the gene, the animal, or the laboratory technique called Fluorescence in situ Hybridization. The task of selecting the correct entity from multiple candidates is then *entity disambiguation*. Given a text mention, the candidates are often selected from a KB, for example based on high string similarity to entity names. In this case, the selected candidate is the canonical entity; the task is further known as entity linking, entity normalization, and entity resolution, to emphasize the making of a connection between the text mention and the canonical entity. When the KB (such as UMLS) contains further information about the entity (such as entity type and entity definition), reaching the canonical entity means attaining this extra information.

The inherent differences in texts in the general and biomedical domains lead to many differences in their handling of entities. Table 2.1 below compares these differences.

	General domain	**Biomedical domain**
What constitutes an entity?	Proper nouns for named entities.	Proper nouns and composite noun phrases.
	Named entities are generally mutually exclusive. For example, *President Bush* is ambiguous and may refer to either George Bush the senior or his son. However, when *President Bush* appears in context, only one of the two Bush'es can be the correct person intended.	Entities are not always clearly mutually exclusive. In UMLS, entities are very fine-grained such that different shades of the same underlying entity are cataloged as different entities. This phenomenon is often seen in entities with modifiers, such that, for example, *lung cancer at onset* and *lung cancer at end stage* are separate entities from the same underlying disease. Therefore, depending on the text mining task, the desired granularity in disambiguation results may differ.
Entity names	Named entities are easily distinguished since each word starts with an upper case letter.	Named entities may or may not feature upper case letters; for example, while commercial drug names (such as Tylenol) start with an upper case letter, the corresponding chemicals (acetaminophen) are written in all lower case.
	Synonyms are generally fixed and known.	Both proper nouns and composite noun phrases are written in many variations. This phenomenon comes in many styles. For example, the same wet lab procedure can have different word orders (*X-ray emission spectrometry* and *spectrometry of X-ray emission*). Chemical names have inconsistent word divisions (one or more of the hyphens in *18-Hydroxy-11-Deoxycorticosterone* can be omitted). Long entity names may be abbreviated in multiple ways (*DNC*, *D&C*, and *D and C* all mean dilation and cutterage). Clinical texts in particular feature many abbreviations that are often hospital- or even doctor-specific.
Vocabularies	Apart from names, regular English words are sufficient.	On top of domain-specific names, domain-specific jargon is used in addition to regular English words. Some words exist only in the biomedical vocabulary (e.g. *methylation*), while other English words take up additional biomedical meaning. For example, *expression* means a textual utterance in the regular English sense, but additionally means a gene causing some effect.
Resources	Off-the-shelf software tools are available to disambiguate named entities.	MetaMap is the de facto software tool for entity disambiguation that can handle all sub-domains. In addition, different sub-domains have their own specific software tool.
	A few large KB's such as DBpedia, FreeBase, and YAGO serve as the pool of entities.	UMLS is the de facto KB covering all sub-domains. Many sub-domains have their own individual and smaller KB's.
Entity disambiguation vs. word sense disambiguation (WSD)	Entity disambiguation and WSD are separate research topics.	There is no clear-cut distinction between entity disambiguation and WSD, since noun phrases are often long and complex and words therein have ambiguous, biomedical word senses. For instance, *expression* have two senses, namely, the process of a gene effecting changes, and the facial expression reflecting an emotional state.
Prior probability distribution	Some named entities have a strong prior probability.	Most entities have a prior distribution heavily lopsided towards the dominant entity. In the NLM WSD dataset [219], the average prevalence of the dominant word sense is 78%.
Determining correctness	Since entities are named and mutually exclusive, it is clear whether the disambiguated entity is correct.	Since UMLS is a heterogeneous KB with very fine-grained entities from different source dictionaries, and since both proper nouns and compound noun phrases require disambiguation, exact disambiguation is often difficult even for humans. For example, whether *children* is the the human being entity, age group entity, or family member entity can be difficult to discern even when textual context is provided.

Table 2.1: Comparisons regarding entities between general and biomedical domains

2.4 Text Mining

Text mining refers to a broad range of information extraction tasks, such as text categorization, sentiment analysis, linguistic trends analysis, entity extraction, and relation extraction. In the biomedical domain, text mining concentrates on relation extraction, where the relations are itself a broad range, such as protein-protein interactions, gene-disease correlations, drug-adverse-effect relations, and pharmacogenomics (i.e. gene-drug-response relations). Regardless of the kind of relations, biomedical text mining tasks share one common processing mode, namely, the processing of a large amount of input text in a pipeline fashion. Modules in a pipeline depends on the exact IE task, though most pipelines adopt the following order: segmenting the text into sentences, performing syntactic analyses such as part-of-speech and dependency parsing, recognizing and disambiguating entities, ending with analysis modules such as incorporating KB knowledge and reasoning that produce the final, desired information.

There are again differences between text mining in the general and biomedical domains. Table 2.2 below compares these differences.

	General domain	**Biomedical domain**
Text genres	There is no dominant focus on a single genre.	Scientific literature is the dominant genre. Since PubMed is the preeminent indexing service for biomedical literature, PubMed is the de facto corpus, which comprises MEDLINE abstracts and PMC (PubMed Central) full-length articles.
	Frequently mined genres include: • News • Encyclopedia, especially Wikipedia • Microblogs, especially tweets • Query logs • Fiction • Conversation transcripts	Other mined genres include: • Encyclopedic health portals on the Web • Patient discussion forums on the Web • News • Microblogs, especially tweets • Drug labels • Clinical texts such as patient records and clinical transcriptions • Clinical trial documents • Patents
Language style	Text is written in regular English prose.	Language style varies from genre to genre. Scientific literature deviates greatly from regular norm of English grammar. An extreme example is the use of more than 10 consecutive words in a compound noun phrase to describe a very specific cell line complete with relevant gene, mutation, species, and laboratory treatment information. In addition, MEDLINE abstracts feature compact and convoluted language to fit the limited document length. Other genres deviate from regular prose according to their communication formats. Drug labels often arrange texts in tables. Social media on the Web feature colloquial utterings, poor grammar, incomplete sentences, misspellings, and excessive use of punctuation.

Table 2.2: Comparisons regarding text mining between the general and biomedical domains

2.5 Natural Language Processing Preliminaries

Synonym, abbreviation, and acronym. Figure 2.1 depicts the relationship between synonyms, abbreviations, and acronyms, which are different ways to describe identical items.

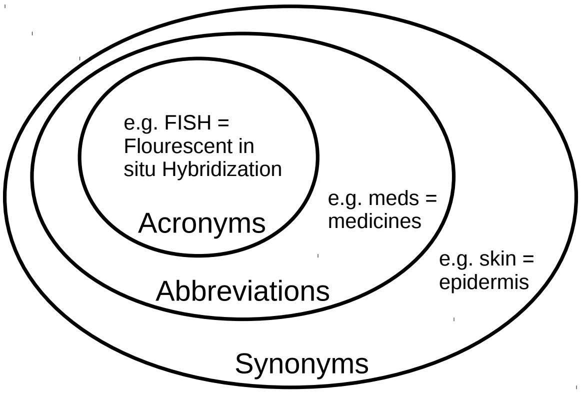

Figure 2.1: The relationships between synonym, abbreviation, and acronym.

Two textual expressions are *synonyms* if they refer to the same item, be it abstract or physical. For instance, *skin* and *epidermis* are synonyms.

Some synonyms are *abbreviations*, where a longer word or phrase is shortened. Abbreviations are typically noun phrases and entity names. For instance, *meds* is the abbreviation of *medicines*.

Some abbreviations are *acronyms* where, in a multi-word expression, the first letters in each word are singled out and taped together. For instance, *FISH* is the acronym of *Fluorescent in situ Hybridization*.

Hypernym and hyponym. Figure 2.2 depicts the relationship between hypernyms and hyponyms, which are hierarchical relations.

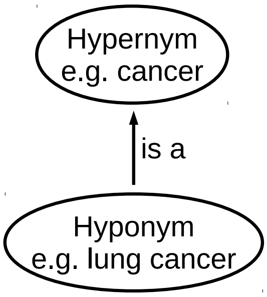

Figure 2.2: The relationships between hypernym and hyponym.

Hypernym is a broader category in which a *hyponym* is a member. For instance, *cancer* is the hypernym of *lung cancer*; equivalently. *lung cancer* is the hyponym of *cancer*.

Part-of-speech and dependency. Figure 2.3 shows the grammatical analysis of a sample sentence, computed by the Stanford CoreNLP tool[4] [116], the standard as well as state-of-the-art tool.

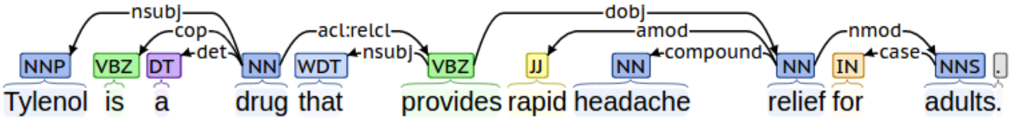

Figure 2.3: POS tags and dependencies of a sample sentence.

Each word in a sentence can be assigned a *part-of-speech* (POS) tag, which describes the word's syntactic function. Noun, pronoun, verb, adjective, adverb, and preposition are the main parts-of-speech. In the sample sentence, POS tags are the colored boxes above each word.

From a syntactic point of view, words within a sentence relate to each other in predefined relations called *dependencies*. In the sample sample sentence, dependencies are marked with arrows; for instance, *headache* depends on *relief* with a "compound" dependency. Dependencies are directed from one word to another, forming a tree. Dependency parsing is the process of determining such a dependency tree, and having first determined the POS tags is a prerequisite.

Morphology. The same underlying word in the English language can be spelled in different ways. Morphology describes how a root word (e.g. the verb *inject* in the infinitive form) should be modified in its various grammatical forms (*injected* as past tense and *injecting* as the gerund form). Modification can also be applied to a noun (e.g. *organization*) to obtain its adjective form (*organizational*). This modification process is called inflection and, in the case of verbs, it is specifically called conjugation.

Lemmatization. The lemma of a word is its prefix letters such that various forms all share that prefix. For instance, the lemma of *writes* and *written* is *writ*. As demonstrated by this example, lemmas are often themselves incomplete words.

Orthography. Orthography is the convention of using spelling, hyphenation, capitalization, word breaks, emphasis, and punctuation. For instance, *Non-Hodgkin's Lymphoma* and *non Hodgkin lymphoma* are the same disease written orthographically differently. Therefore one common strategy in unifying such differences is by transforming all letters to lower case and replacing non-letters with spaces. On the other hand, such differences are exactly the cues for recognizing certain entity types such as gene names. The combination of upper letters, digits, slashes, and italics (e.g. *EGFR/HER1*) is a strong indicator for gene name.

[4] `stanfordnlp.github.io/CoreNLP`

Chapter 3

State of the Art

Entity recognition and disambiguation are research problems in both the general and the biomedical domains. However, domain-specific characteristics of biomedical text lead to additional challenges, such that solutions in the general domain, if applicable at all, must be adjusted to address those challenges. In this chapter, we present the state of the art first in the general domain and then in the biomedical domain. Table 3.1 below provides an overview of this chapter.

	General domain	**Biomedical domain**
Entity recognition	Well studied for named entities.	Well studied for named entities in sub-domains. Noun phrases are also entities but under-explored.
Entity typing	Well studied for coarse-grained types (person, organization, geographic location, and miscellaneous). There are fewer methods for fine-grained types based on KB categories.	Entity typing is integrated into entity recognition. Explicit typing methods are scarce.
Entity disambiguation	Well studied for named entities.	Well studied for selected individual sub-domains. There are few methods that address all sub-domains.
Entity-centric competitions	The few events are auxiliary to the research community.	Plentiful events are core to the research community, as well as trailblazer of research directions.

Table 3.1: The state of the art for the general and biomedical domains

3.1 General Domain

3.1.1 Entity Recognition

Entities of interest in the general domain are named entities. Works in named entity recognition (NER) go as far back as the late 1990's [32], and progressed through the next decade as language-independent shared tasks at the CoNLL (Computational Natural Language Learning) conference [205, 206]. The bulk of proposed methods employ machine learning to assign, for each word in a passage, one of three labels – B to indicate the beginning of a named entity, I for "inside" or the continuation of a named entity, and O for "outside" of the named entity. Features for the machine learning include word-level ones (such as part-of-speech tags, orthographic patterns, lexical information, and previous word) as well as document-level ones (such as gazetteers and document meta-information). Yet still later, as social media became a genre available for text mining, there has been efforts [165, 198] to refine existing approaches to address the corresponding language style. Overall, this line of research recognizes entities and simultaneously classifies them into 4 types: persons, organizations, geographical locations, and miscellaneous [155]. The Stanford CoreNLP toolkit [116] has emerged as the de facto method as well as software tool that is currently the state of the art.

3.1.2 Entity Typing

Four coarse types. As alluded to above, the task of entity recognition already implicitly determines 4 coarse entity types (persons, organizations, geographical locations, and miscellaneous).

Fine-grained types. Other existing works explicitly determine fine-grained types, also often using entity recognition as a means of achieving that aim. From as few as 9 [164] to as many as 500 [234] types, various start-of-the-art methods address different levels of granularity.

FIGER [113] augments the 4 entity types by expanding each of them to more than a dozen types. It further introduces other types related to products, arts, and events to reach a total of 112 fine-grained types. In comparison, PEARL [133] and ClusType [164] use types that are more and less fine-grained, respectively. These two systems share two common themes in their methodologies. First, they both leverage relational phrases to infer the types of the entities involved in those relations. Second, they are both data-driven. By using efficient algorithms (integer linear programming for PEARL, and label propagation plus clustering for ClusType), both systems can process large amounts of data so that the system can accumulate a critical mass of information to draw meaningful conclusions.

Hierarchical types. Entity types do not need to be mutually exclusive. In particular, when they are arranged in a hierarchy of categories, an entity belonging to a more fine-grained sub-type (for instance, fruit) also belongs to the more coarse-grained super-type (food) simultaneously. Such is the premise for Rahman and Ng's method [153] and HYENA [234]. The former system arranges 92 sub-types in 29 super-types in a two-tier manner. The latter system goes as far as using 5 trees of 100 sub-types each, such that a tree can be up to 9 tiers deep. While both systems perform multi-label classification, the underlying algorithms employed are disparate. Rahman and Ng's method constructs factor graphs and performs inference, while HYENA formulates the problem as a Support Vector Machine (SVM) instance. Interestingly, despite the different choices of algorithm, the feature sets have much in common. Most notably, they include unigrams in the text mention, grammatical features such as part-of-speech and nearest verbs, as well as gazetteers.

3.1.3 Entity Disambiguation

Classic three-pronged approaches. There is a large body of prior works [112, 181] for named entity disambiguation (NED). State-of-the-art methods include AIDA [70], Babelfy [127], Spotlight [123], TagMe [46], and Wikifier [156]. All of these methods draw upon three types of ingredients, as summarized by [109]. First, popularity or prior probability or an entity assumes that the most prominent candidate is the most likely underlying entity. Such a choice can be described as an educated guess that does not require any further information from the text mention or the document in which it occurs. The second ingredient, context similarity, takes the document into account. It stipulates that similarity between the document containing the text mention and a definitional document that describes the candidate (for example, the candidate's Wikipedia article) must be high. The third ingredient, coherence, leverages further text mentions in the same document to stipulate that their underlying entities should all be related to the same topic. This is typically achieved by leveraging a KB, such that given a candidate, its categories, keywords, and related entities can be used to compare topic similarity.

Most of these state-of-the-art methods are also available as software tools. From a user's point of view, they offer a range of choices, where one can "shop around" for the most suitable tool or combine results of multiple tools.

Approach based on word embeddings. Another recent method [230] deviates from the above three-thronged approach, and opts for extending the popular skip-gram word embeddings to jointly model word-word, word-entity, and entity-entity similarities. A knowledge base (Wikipedia in this case) is still the backbone of the model, where link statistics are used to model the latter two similarities involving entities.

Joint NER and NED. Although NER and NED are separate tasks, each can provide cues for the other. Jointly performing entity recognition and disambiguation in a combined manner can therefore enhance the performances of both tasks. Such a joint approach is investigated in three methods, namely NEREL [183], a method by Durrett and Klein [37], and J-NERD [140]. While the NEREL uses a re-ranking model based on relations, Point-wise Mutual Information (PMI), and co-occurrence information extracted from Freebase and Wikipedia, the latter two methods both converge on using Conditional Random Field (CRF). Specifically, in the method by

Durrett and Klein, a classic linear chain CRF is enhanced with 3 more types of edges to model relationships introduced by NER, coreference resolution, and NED. On the other hand, J-NERD constructs a CRF to reflect the dependency parse tree of a sentence, and adds extra edges to model coreferences across sentences.

3.1.4 Competitions

In contrast to the biomedical domain, entity-related competitions are fewer and are less prominent in the general domain. The CoNLL shared tasks take place yearly covering a wide range of topics, though only the two aforementioned ones concern named entities. Another yearly-recurring event is the Text Analysis Conference – Knowledge Base Population (TAC KBP)[1], which offers the Entity Discovery and Linking (EDL) track. Apart from that, there are only one-off events such as the Cross-Device Entity Linking Challenge[2] at the 2016 Conference on Information and Knowledge Management (CIKM), and even smaller events targeting languages such as Chinese [44], Dutch[3], and German [11].

3.2 Biomedical Domain

The big picture of text mining in the biomedical domain has been comprehensively presented in surveys such as [71, 54, 167, 184, 241]. Under the umbrella of text mining is a range of tasks. We focus here on the entity-centric tasks.

3.2.1 Entity Recognition

Applicable to a Single Sub-domain

NED methods that target only a single sub-domain are the bulk of NED efforts in the community. Not only are there many methods proposed, the variety of sub-domains is also wide. Here we present a selection of the dominant sub-domains.

Genes and proteins. The two most prominent sub-domains are undoubtedly genes and proteins, which are often tackled simultaneously since, conventionally, a gene and its protein product bear the same name. Among the many methods developed, BANNER [100] is a prominent one based on CRF using orthographic and morphological features, and shallow syntax features. A recent work by Sheikhshab et al. [179] also uses CRF, but further combines it with techniques borrowed from graph-based label propagation to model prior knowledge. Recognizing gene names in patents [59] is another recent development that to date is an under-studied area.

Derivatives of gene and protein. These entities have a nomenclature different enough from regular genes and proteins that they require tailored methods to be recognized. In addition to genes, GNormPlus [222] also recognizes protein domains. For

[1] `tac.nist.gov/tracks`

[2] `competitions.codalab.org/competitions/11171`

[3] `wordpress.let.vupr.nl/clin26/shared-task`

gene mutations, tmVar [220] uses extensive word shape features to harness this nomenclature and train a CRF, followed by applying regular expressions as post-processing to both pick up mentions missed by the CRF and filter out false positives. SETH [203] instead harnesses nomenclature grammar guidelines and regular expressions to pick up the mentions.

Chemicals. Being entities at the molecular level interacting with genes and proteins, chemicals is another prominent sub-domain. The most popular underlying algorithm is CRF by far, as features such as word stem, prefix, suffix, word shape, and the use of Greek letters and Roman numbers are amenable to analyzing a chain of word tokens. Works such as the method by Grego et al. [56], ChemSpot [166], CheNER [212], and tmChem [103] all follow this line of methodology. Patents, as an under-studied text genre, are especially important for the pharmaceutical industry. Habibi et al. [60] investigate the efficacy of two of these tools (ChemSpot and tmChem) originally designed for regular biomedical text on this genre.

Drugs. Drug names have much overlap with chemicals, but they also have many standardized names curated in dictionaries such as the DrugBank [228]. Therefore Korkontzelos et al. [96] develop an ensemble system that aggregates results from individual NER methods that leverage a dictionary, word token-level features taken from chemical NER, regular expressions, as well as statistics gleaned from existing gold standard corpora. They report that results aggregated by a maximum entropy model generally outperform those by a perceptron classifier.

Cell lines. In order to evaluate the performance of existing methods against a new corpus, Kaewphan et al. [87] specifically include entities related to synthetic, lethal genes that appear in cancer literature. They also include a novel method based on CRF, and investigate the impact of dictionary and background corpus used in training the CRF.

Species. A closely related sub-domain to cell lines is species. However, unlike cell line names, species names are relatively few and well cataloged in dictionaries. Therefore systems such as LINNAEUS [51] and SF4GN [221] adopt a dictionary lookup approach.

Anatomical parts. Anatomical entities have been well established and cataloged in dictionaries. Moreover, their textual expressions have limited variations. AnatomyTagger [151] thus uses two dictionaries as the primary features for a CRF model, namely the species-independent Common Anatomy Reference Ontology (CARO) [61], and the human-centric Foundational Model of Anatomy (FMA) [168].

Diseases. Named entities such as *Down's Syndrome* and *Alzheimer's Disease* are well cataloged in dictionaries. On the other hand, unnamed noun phrases such as *inherited male breast metastatic cancer* are also abundant. Although their exact expressions vary greatly, part of the phrases are often cataloged in dictionaries as well.

Despite these observations, Sahu and Anand [171] take a decidedly different approach and investigate the feasibility of avoiding sub-domain-specific features. Their work investigates instead the contributions of character- and word-level embeddings to multiple types of neural networks.

Phenotypes. A phenotype is an organism's observable trait under the influence of the organism's genes and its environment. Like diseases, phenotypes encompass both named entities (for example *Crouzon syndrome*) and composite noun phrases (*abnormality of head and neck*). Therefore despite the existence of dictionaries such as the Human Phenotype Ontology [94], the ultimate goal of recent works is the construction of new dictionaries [2, 31] or expansion of existing ones [93], for which recognizing phenotypes in text is a means to achieving that goal.

Symptoms. Symptoms are closely related to phenotypes, since a symptom is a particular bodily trait that surfaces due to a disease or some environmental disturbance. There are relatively few existing works that focus on recognizing symptom entities. As a step towards fully recognizing text mentions of symptoms, Sondhi et al. [190] propose to identify whole sentences that contain symptom descriptions using word- and sentence-level features in separate CRF and SVM approaches. The recent introduction of the Micromed corpus [82] will be useful for future symptom recognition works since it provides annotations of symptom text mentions in tweets.

Applicable to Multiple Specific Sub-domains

Common sub-domain combinations. The set of sub-domains targeted by one single NED method is often influenced by gold standards, which in turn reflects the focus of the research community at the molecular level. One popular corpus is the GENIA corpus [90], which contains text mentions annotated with 6 entity types: proteins, DNA"s, RNA"s, chemicals, cells, and cell lines. Another popular corpus is the CRAFT corpus [215], which contains text mentions annotated with entities from 7 ontologies about proteins, genes, chemicals, cells, and taxonomy. Proteins, chemicals, and cells are therefore a staple combination.

Funk et al. [48] compare two such methods, the ConceptMapper [201] and the NCBO Annotator [178], against the CRAFT corpus. Another two prominent systems are ABNER [177] and Gimli [20], both of which use CRF to recognize the 6 GENIA entity types. A survey by Neves and Leser [139] examines further annotation methods that target the same line of entity types.

Since PubMed abstracts are the standard corpus, and the sub-domains of interest are commonly shared amongst the research community, the team in the University of Turku has taken the initiative to provide a public service to the rest of the community [62]. They have applied NERsuite[4] to the entire PubMed corpus and made the results publicly available. NERsuite is a tool built upon a fast CRF implementation called CRFsuite[5] and, in this case, CRF models specifically trained for the few selected biomedical sub-domains are plugged into the tool.

[4] `nersuite.nlplab.org`

[5] `www.chokkan.org/software/crfsuite`

Ensemble Web service. Very recently (in 2016), the National Library of Medicine (NLM) constructed a Web service [224] that combines 5 existing state-of-the-art NED methods targeting separate sub-domains. These individual methods have already been mentioned above (DNorm, GNormPLus, SF4GN, tmChem, tmVar). As a result, it is a one-stop Web-based tool that caters for the usual molecular-centric entity types (chemicals, genes, gene mutations, proteins) as well as diseases and species.

The BeCalm initiative. The "next big thing" may be the BeCalm online tool[6] which, at the time of writing this thesis, is in the initial stage of the project. BeCalm TIP (Technical Interoperability and Performance of annotation servers) is a task for recognizing chemicals, genes, and proteins in the upcoming BioCreative 5 competition. Task participants place their NED implementation into a Web server, and different implementations will be compared against the same gold standard. The vision is that future users can request NED results from one or more implementations.

Applicable to All Sub-domains

MetaMap. There is no question that MetaMap [5, 6] is the de facto software tool for the last decade. Its dominance can be attributed to its availability as a standalone, easy-to-use software, and to its applicability to all sub-domains. The latter is achieved by taking UMLS, the largest metathesaurus, as the dictionary. A major disadvantage of MetaMap is its slow speed. The publisher of the PubMed corpus and the owner of MetaMap, the National Library of Medicine (NLM), has responded by preprocessing the entire corpus with MetaMap[7].

Other dictionary-based approaches. Besides MetaMap, other NED implementations are also available. The BioPortal API[8] is a Web service that offers to annotate text mentions against dictionaries chosen by the user. MaxMatcher [238] implements an approximate entity name matching method with UMLS as the dictionary, taking into account that an entity in UMLS has multiple, often similar lexical variants.

Dictionary-less approaches. There are also methods proposed to perform NED without attempting to tie the entities to any entity type. These methods generally forego the dictionary, and instead aim to more accurately delineate the text mentions that truly express named entities. For instance, Kim et al. [91] focus on linguistic cues such as head words and word patterns. Their method uses a SVM classifier, and considers noun phrases as named entities.

3.2.2 Entity Typing

As an integration in NER or NED. The study of explicitly typing biomedical entities is scarce. We observe that entity type is often implicitly determined as an integrated part of an NED or NER method. In the case of NED, recall that a text

[6] `www.becalm.eu`

[7] `ii.nlm.nih.gov/MMBaseline`

[8] Accessible via `http://data.bioontology.org/annotator?text=[insert text]`

mention is often recognized for one the 6 GENIA entity types. For many biomedical text mining tasks focused on the molecular level, those 6 types are already sufficient. In the case of NER, text mentions are generally mapped to UMLS entities. Every entity has at least one expert-assigned UMLS semantic type. Therefore by virtue of using UMLS as the KB, NER always provides the corresponding entity types.

Explicit typing approaches. Nevertheless, there exist works that study entity typing as the primary concern. Stenetorp et al. [192] propose a machine-learning-based method and enhance it with approximate string matching. Depending on the gold standard, their method can disambiguate up to 97 entity types, which are therefore fine-grained like *sub-cellular structure*, *multi-tissue structure*, and *developing structure*.

The DIEBOLDS method [13] uses label propagation to disambiguate entities between only two types, namely diseases and drugs. In this work, the source texts are semi-structured Web pages; a typical page contains headings like *Symptoms*, *Side Effects*, and *Precautions* which provide important cues to the entity types therein. The label propagation graph harnesses these document structures and relations in a KB to build the edges.

3.2.3 Entity Disambiguation

Difficulty aspects. That biomedical entity disambiguation is a difficult task is only a blanket statement. Besides the better known reasons such as numerous synonyms and long noun phrases with diverse lexical variations, the task is also difficult in other aspects. For instance, different sub-domains have different levels of difficulty, which is reflected in highly varied disambiguation performances across sub-domains under the same method. Besides the inherent characteristics between entity names, other contributors include the abundance and distribution of training data [81, 235] as well as the quality and completeness of the KB being harnessed [79, 242]. From another point of view, textual style such as terms with regular English words and abbreviations also cause a rift in disambiguation performance [194].

Applicable to Single Sub-domain

Genes and proteins. Similar to entity recognition, genes and proteins are the most prominent sub-domains for entity disambiguation. Rebholz-Schuhmann et al. [161] review various methods trained on different KB resources against multiple gold standards. The methods fall into two broad categories, namely machine-learning-based and lexical-resources-based.

Diseases. As mentioned, named diseases are well cataloged in dictionaries. DNorm [101] and its clinical-text variant DNorm-C [102] exploit this unique arrangement by using pairwise learning to learn similarities between text mentions and dictionary entity names. Another work that targets clinical texts is a method by Zuccon et al. [240]. It first applies MetaMap to gather a first round of results. Knowing that many disease mentions are missed by MetaMap, it further applies CRF to pick up any remaining mentions. Jimeno-Yepes et al. [80] also leverage MetaMap, and use

MetaMap as one of three voting modules in an ensemble. The second module is a dictionary lookup method, and the third module is adapted from an information theory-based method [50] that models the specificity of KB entities. In a completely different approach, Islamaj Doğan and Lu [75] use a series of rules in a decision-tree-like manner, comparing a text mention's similarity to an entity's primary name and synonyms, as well as leveraging scoring functionality provided by Lucene[9], a standard indexing and search engine software tool.

Clinical texts. Although clinical texts are not an entity type, as a sub-domain they have recently seen a spate of works aimed at tackling the peculiar textual style. Besides the two aforementioned works [102, 240] that disambiguate disease names, other works focus on entities especially abundant in clinical texts. For instance, Wu et al. [229] address abbreviations, and they show that word embedding features improve the performance of an SVM using only conventional unigram and orthographic features. Kreuzthaler and Schulz [97] go as far as disambiguating only period characters. When the context to the left and right of a period is fed to a decision tree, the rules decide whether the period ends a sentence, ends an abbreviation, is part of a number, or is part of some special code from a controlled vocabulary.

Gradable terms are characterizations such as *normal* and *severe*, which have different meanings when applied to different entity types (for example measurement vs. disease), as well as different meanings between entities of the same type (for example *normal systolic dysfunction* vs. *normal anemia*). Shivade et al. [182] propose a probabilistic model not only to cluster related gradable terms, but also to order them by ordinal relationships and provide concrete numerical ranges of the corresponding clinical observations.

Applicable to Multiple Specific Sub-domains

Unlike entity recognition, entity disambiguation methods either address one sub-domain, or have no sub-domain limitations; to the best of our knowledge, there are no existing methods that specifically target a certain set of sub-domains. One likely reason is the specificity of individual sub-domains. As mentioned, different sub-domains present different challenges, so that it is sensible to tackle one sub-domain at a time, and combine separate solutions as needed.

Another likely reason is the availability of gold standards. To the best of our knowledge, the CRAFT corpus [215] is the only multi-sub-domain gold standard that contains annotations with KB entity identifiers pooled from multiple KB's including UMLS. Of the sub-domains featured (chemicals, genes, proteins, cells, and taxonomy), only a small portion of the annotations point to UMLS entities.

Applicable to All Sub-domains

MetaMap and alternatives. Despite its limited functionality in the disambiguation module, MetaMap remains the de facto standard software tool for the task due to its availability and easy setup. Other proposed methods employ a variety of ap-

[9] `lucene.apache.org`

proaches. For instance, Zheng et al. [236] build a graph for each document using the rich semantic information and structure of many KB's, perform collective inference, and finally rank the entity candidates. Zwicklbauer et al. [243] use a query-based approach. Documents describing entities are transformed into a document-centric KB, so that a text mention and its context can be turned into a query and retrieve KB items via Learning to Rank. Kim and Yoon [92] focus on disambiguating abbreviations by modeling word-topic, document-topic, and word-link distributions with an adaptation of Latent Dirichlet Allocation (LDA).

Designed for novel corpora. Texts from social media on the Web, in particular tweets and blog posts, are the focus of Limsopatham and Collier [110]. Their work compares the efficacy of Convolutional and Recurrent Neural Networks (CNN and RNN, respectively), where word embeddings are the primary feature. The CLEF-ER laboratory shared task [158] provides parallel corpora in 5 European languages, and the participating methods must disambiguate entities using the English corpus plus one or more non-English corpora.

Word Sense Disambiguation

Recall that entities of interest for the biomedical domain include noun phrases, which contain words that carry multiple biomedical meanings. Therefore word sense disambiguation (WSD) is a closely related problem to entity disambiguation, and the line between these two tasks is blurry.

Two gold standards. Most existing works are driven by two prominent gold standards, NLM WSD [219] and MSH WSD [83], which both contain judiciously selected and ambiguous words to reflect a range of underlying entity types (such as molecular processes, anatomical parts, and therapies) and word types (spelled-out terms and abbreviations).

Methods evaluated against the two gold standards. A wide variety of approaches have been evaluated against these two gold standards. Lin and Verspoor [111] construct an n-gram language model that incorporates semantic information. DALE [148] uses word lemmas to build feature vectors for candidates and text mentions before comparing them via cosine similarity. Stevenson et al. [193] specifically target PubMed abstracts for their indexing terms, or MeSH (Medical Subject Headings) terms, from which that the topic of the abstract can be determined and leveraged in the subsequent Personalized PageRank method. Jimeno-Yepes et al. [84] extract multiple kinds of word collocation statistics from text mentions, and use them as features for a Naïve Bayes classifier and separately for a vector-based similarity comparison method. Finally, Jimeno-Yepes [77] investigates the combination of local features derived from a text mention's context and global features in the form of word embeddings, and apply them to a recurrent neural network with long short-term memory.

3.2.4 Competitions

In the biomedical text mining community, competitions are important events and they occupy a significant proportion of the community's collective effort. As a further effect, these competitions impact the research directions of the community over the years not only by proposing new research problems, but also by providing gold standards as staple benchmarks. Driven by real-life information needs of "wet lab" scientists and professional knowledge base curators, various competitions elicit solutions that aim to satisfy those needs. While a survey by Huang and Lu [72] presents a comprehensive list of competitions, here we highlight the recent ones that emphasize working with entities.

For Biomedical Text Genre

BioASQ[10] is a competition about finding answers to biomedical questions. Of particular interest are factoid questions, whose answers are named entities. Other types of questions have answers in the form of text snippets and RDF (Resource Description Framework) triples from KB's.

BioCreative[11] hosts over the years entity recognition competitions for chemicals and genes, as well as entity disambiguation competitions for genes, Gene Ontology terms, and genes with corresponding species.

BioNLP shared tasks feature named entity recognition tasks for very specific sub-domains, and further requires participating methods to connect the recognized entities into relations. For instance, in the bacteria biotope task [34], entities of types hosts, body parts of the host, environments (food, medical, soil, and water), and geographical locations are to be recognized with the correct type. Another task, SeeDev [21], is about plant seed development. There are 16 entity types spanning from genes, metabolic pathways, genotypes, to environmental factors for recognition, to be followed by extraction of genetic and molecular mechanisms.

CALBC Challenge [162] uses the CALBC silver standard [159] as its evaluation corpus. Results of participating systems are evaluated against a portion of the annotations for 4 specific sub-domains.

For Clinical Text Genre

ShARe / CLEF eHealth Challenges include tasks that focus on the recognition and disambiguation of diseases [147], and the disambiguation of abbreviations [128].

i2b2 features an entity recognition task [214] that focuses on 3 entity types, namely medical problems, tests, and treatments. In another task [213], medications are to be recognized as named entities, together with the extraction of related information such as dosage and duration.

[10] `www.bioasq.org`

[11] `www.biocreative.org`

SemEval features one task [146] for the recognition and disambiguation of diseases.

Chapter 4

Fast Entity Recognition

MetaMap, the de facto standard software tool for biomedical entity recognition, employs much Natural Language Processing (NLP) machinery to recognize entities in UMLS (Unified Medical Language System), the largest metathesaurus. Knowing that NLP machinery is time-consuming, and that UMLS is rich in lexical variations, we investigate whether a fast, string-similarity-based method can achieve results comparable to those of MetaMap. We implement an NLP-light method that performs fast MinHash lookups via character trigram features. When compared to MetaMap, our method achieves comparable precision and 13% less coverage using less than 1% of the time.

4.1 Introduction

4.1.1 Motivation

In recent years, the amount of biomedical information has grown tremendously, and much of this information is disseminated in a variety of free texts. Besides PubMed[1], the preeminent resource for scientific literature, the Web features many health portals and patient discussion forums for the layman. The Biomedical Natural Language Processing (BioNLP) community has responded by developing and applying various text mining techniques in order to extract this information buried in the texts.

Biomedical text mining spans a range of information extraction (IE) tasks such as extracting protein-protein interactions [105], drug-adverse-effect correlations [89, 226, 227], and pharmacogenomics networks [33, 154]. Regardless of the goal of the IE task, text mining is a pipeline of processes or sub-tasks, where the final sub-task produces the desired information. Entity recognition (ER) often serves as an early sub-task, upon which other downstream sub-tasks depend. A typical sub-task immediately following entity recognition is entity disambiguation. Extracting relations between two entities in a dictionary, for instance, requires text mentions to be mapped to entities in the dictionary before relations between them can be analyzed. Downstream sub-tasks not only depend their success on the quality of the entities recognized, they may not even begin before the entity recognition task is completed. Therefore, it is crucial that a biomedical entity recognition method provides high quality results.

[1] www.ncbi.nlm.nih.gov/pubmed

ER in the biomedical domain presents unique obstacles uncommon to the general domain, such that providing high quality results is already a challenge. Biomedical texts, especially scientific literature geared towards professionals, are steep in specialized jargon. Biomedical concepts are often expressed in long phrases with a large number of variations. In the general domain, entities are often named and easily distinguished as noun phrases; in the biomedical domain, however, even verbs may be considered entities when they, for example, describe specific medical procedures as in *diagnosed* and *injected*.

Since the amount of aforementioned free texts is published at an ever increasing pace, a second challenge in biomedical ER is to provide high throughput. PubMed comprises over 26 million citations and is growing at more than one million new citations per year. As for Web content, existing health portals constantly have their contents updated and expanded, while patient discussion forums naturally only grow in size. Indeed, maintaining high quality while yielding high throughput is a prerequisite for any ER method that aims to support text mining at PubMed-scale.

To counter these challenges, ER has been a major area of research within the BioNLP community. Since entity names in different sub-domains exhibit characteristics consistent within its own sub-domain that are different from those in other sub-domains, the vast majority of efforts address entities of individual sub-domains such as chemicals, genes, and proteins. Of the efforts that address all sub-domains, there are relatively few works. Since its launch in 2001, MetaMap [6] has become the de facto standard software tool for general-purpose biomedical ER. MetaMap is a software installed on the user's own local machine. It takes the entire UMLS[2] as the dictionary; in other words, MetaMap considers all entity names from all sub-domains present in UMLS. Employing much Natural Language Processing (NLP) machinery, MetaMap trades off higher quality with lower throughput. As text collections grow in size, this lower throughput gradually becomes the bottleneck of a text mining pipeline. In our experience, without parallelization, processing 600k PubMed abstracts – a small portion of over 16m English abstracts in the entire collection – takes 26 days (3.8s per abstract using a single instance of MetaMap). Apart from MetaMap, few publicly and freely available alternatives exist. One possibility is the BioPortal API[3], an entity annotation service accessible over the Web, though the dependence on network connectivity again limits throughput and renders the option unsuitable for large-scale use.

Knowing that MetaMap takes UMLS as its dictionary of entities, we further observe that UMLS is the largest metathesaurus, rich in entity names, their synonyms, and their lexical variations. This observation spurs us to investigate an alternative ER method that is fast and exploits this lexical richness.

4.1.2 Contribution

We devise a string-similarity-based method for biomedical ER aimed at minimizing the use of NLP machinery and thus processing time. Specifically, the method achieves high quality and high throughput.

[2] `www.nlm.nih.gov/research/umls`

[3] Accessible via `http://data.bioontology.org/annotator?text=[insert text]`

High quality is achieved by exploiting the rich collection of lexical variations of entity names in UMLS, as this collection is amenable to a string-similarity-based approach. We observe, however, that this collection is still incomplete. We augment the collection with the missing variations, namely the plural forms of existing nouns as well as the full set of conjugations of existing verbs.

High throughput is achieved by turning to MinHash [19] as the key ingredient. MinHash is a variant of the Locality Sensitive Hashing (LSH) [22] algorithm that transforms a dictionary lookup into a hash lookup with high probability of success. In terms of NLP processing, the method uses at most part-of-speech tagging, which is a fast process, and avoids any further processing such as dependency parsing.

Together with judicious selection of a subset of the UMLS dictionary and simple heuristics in selecting which text mentions to perform lookups for, the method achieves up to 83% precision and 78% coverage under a strict rating scheme that penalizes failure in Word Sense Disambiguation (WSD), at a throughput of 1,720 PubMed abstracts or 175 Web pages per minute. The resulting code has been released as an open source software, and has made it possible to process large corpora in other biomedical text mining works [39, 40, 41].

4.2 Related Work

Biomedical ER for a Single Sub-domain

ER in the biomedical domain often focuses on a specific sub-domain. Proteins and genes are the most popular sub-domains, and the BioCreative initiative has been driving the BioNLP community with various gene mention recognition [189, 233] and normalization tasks [67, 115, 126]. Out of a large body of works, there are a number of software tools publicly available, where Gimli [20] and ABNER [177] are two notable ones. Since gene and protein names are written in a highly specific but non-standardized manner, recognizing their text mentions continues to be a research challenge. As recent as 2016, Sheikhshab et al. [179] propose to use a graph-based method to leverage a two-word window around a gene name in order to improve precision. Recognizing protein-centric entities such as sequence variants has been studied as well [220].

Chemicals are another popular sub-domain for ER because chemical names are written in a completely different but just as specific and non-standardized manner. Chemical names are a mixture of established names (e.g. *ferric oxide*), chemical elements and their symbols (Fe_2O_3), and other established words such as prefixes and suffixes, all jumbled up as multi-word or long formulaic expressions (*Amylo-(1,4,6)-transglycosylase*). Therefore for chemicals, orthographic features are an important ingredient for recognizing entities, as evidenced in two existing works [9, 103]. And since chemical names are too variable, no dictionary can hope to exhaustively list all possible names. As a result, there are existing works [9, 166, 212] that leverage a dictionary as a starting point, and refine the intermediate results with more sophisticated machinery such as Conditional Random Fields (CRF's). There has also been effort [95] to draw upon multiple methods and combine their results in an ensemble manner.

Besides recognizing proteins, genes, and chemicals, there are also works focusing on other sub-domains such as anatomical parts [151], cell lines [87], diseases [106, 138, 171, 240], drugs [96], malignancies [86], organisms and species [51, 132], as well as entities related to a single, highly specific biological system (bacterial type IV secretion system) [3].

Biomedical ER for Multiple Sub-domains

A number of existing approaches [91, 135, 176, 180, 191, 208, 235] are in principle applicable to all entity types. In practice, however, these approaches study their performances using one dominant gold standard, the GENIA corpus [90]. This corpus contains annotated text mentions belonging to 6 entity types: proteins, DNA's, RNA's, chemicals, cells, and cell lines. As a result, how generalizable these approaches are beyond these 6 entity types remain to be studied. A review by Funk et al. [48] provides a detailed analysis of further ER approaches targeting these entity types, using the larger and more recent CRAFT corpus [7].

Biomedical ER methods that truly tackle all sub-domains are relatively sparse. The seminal work by Frantzi et al. [47] propose the C-value / NC-value method to recognize multi-word terms in an unsupervised manner. BANNER [100], a method based on CRF that decidedly forgoes a dictionary, is another milestone contribution that becomes a building block for later, bigger systems. For general-purpose biomedical ER, however, MetaMap remains the most widely used software tool and is widely regarded as the de facto standard tool. Other alternatives such as the BioPortal API, MaxMatcher [238], and NOBLE [209] do exist to tackle any text genre, while cTakes [173][4] is specifically designed to tackle clinical text. A survey by Neves and Leser [139] offers a comprehensive overview of entity annotation tools. The League Table [160] was an effort to supply an online platform to compare and benchmark different annotation tools against multiple gold standards; this service seems to have been decommissioned. At the time of writing of this thesis, BeCalm[5] has just begun as a new initiative to provide an annotation metaserver.

Dictionary Construction and Enrichment

Apart from the task of looking up entities, there has also been efforts to enrich the dictionary upon which lookups are performed. Such is the contribution of BioLexicon [204] – this linguistic resource is a catalog of over 2.2m lexical entries featuring entity names, as well as words specific to the biomedical domain in all their lexical variations. Other sub-domain-specific efforts include enriching a dictionary for abbreviations [141], and constructing dictionaries from scratch for chemical names [66] and human phenotypes [31, 93].

[4] Latest version available at `ctakes.apache.org`

[5] `www.becalm.eu`

Approaches Employing String Similarity

String-similarity-based methods are frequently employed to perform various text-oriented tasks in the biomedical domain. Yamaguchi *et al.* [231] compare the performances of four different string similarity metrics for the task of clustering chemical and non-chemical abbreviations. Wellner *et al.* [225] combine an adaptive string similarity model with CRF's to pick up protein names in free text. String similarity metrics can be cast as a machine learning problem, as Tsuruoka *et al.* [210] propose – given protein names, learn the metric via logistic regression. The resulting metric is later used to look strings up from a dictionary of protein names. In the general domain, SpotSigs [202] extracts word signatures as delimited by determiners, and apply a Jaccard-similarity-based algorithm on these signatures to detect near duplicates in a large Web archive.

Approaches Employing LSH

LSH is a proven technique used in numerous applications, especially when one requires speed in working with large datasets. Ravichandran et al. [157] present an NLP example, where nouns from a Web corpus are clustered based on cosine similarity. More recently, Boytsov et al. [18] uses LSH to approximate k-nearest neighbor search to increase recall in an information retrieval task. Chum et al. [26] present another prominent example in the area of computer graphics. The authors extend the hashing algorithm with weighted set similarity measures. The resulting algorithm is capable of detecting near duplicate images and videos, and is highly scalable. As for the biomedical domain, however, no other ER method employs LSH to the best of our knowledge.

4.3 Methodology

4.3.1 Dictionary Construction

UMLS is made of up entities, called *concepts* in UMLS documentation, where one entity is represented by one or frequently multiple entity names bearing the lexical variations from different dictionary sources. Since the ultimate goal is to look up entity names via a hash table, where a hash signature is based on the entity name's exact spelling, the intermediate goal is to compile a collection of entity names complete in its lexical variations.

Preprocessing. Being a potpourri of different dictionaries with heterogeneous norms, many entity names in UMLS are unsuitable for a string-similarity-based method because they are highly unlikely to appear verbatim in free text. Therefore we apply two rounds of preprocessing in order to filter out unsuitable entity names.

In the first round of preprocessing, each entity name is checked against some 177 suffixes and any suffix occurrence is removed. All but 7 of these suffixes are enclosed in brackets, which already denote their supplementary nature; removing these suffixes does not affect the information content of the entity names. The suffixes belong to a few categories. Table 4.1 shows the categories and sample suffixes, and the entire list can be found in Appendix A.

Category	**Sample suffixes**
Entity type	(cell structure) [Chemical/Ingredient] (person)
Detailed species	(H1N1) (H3N2)
Discoverer and year of discovery	(Hensel, 1867) (Linnaeus, 1758)
Measurement unit	(__ degrees) (GRAMS)
Dictionary-source-specific attributes	[Ambiguous] [dup] -RETIRED-

Table 4.1: Categories of UMLS entity name suffixes

In the second round of preprocessing, entity names featuring long strings are discarded. For instance, *stage I Hodgkin's lymphoma lymphocyte depletion type below the diaphragm* and *pyrithione zinc 2% topical application shampoo* are long names that are highly unlikely to appear as-is in any scientific or Web document. Since we aim to construct a hash table for looking up entity names, including such long names in the hash table clogs it with unproductive signatures and increases collisions. Therefore, we take only those entity names in UMLS that are 5 words or shorter, and 100 characters or shorter. Using the freely available (category 0) portion of the 2012AB dataset (the STR column in MRCONSO table), the subset of UMLS thus obtained features 2.7m unique <entity, entity name> pairs.

Augmenting lexical variations. Many entities in UMLS already provide ample lexical variations such as singular and plural forms (for example in entity C0020974 *immunoglobulin injection*, *immunoglobulin injections*, and *immunoglobulins injection*), verb conjugations (C0021107 *implant*, *implanted*, and *implanting*), and different word orders (C0021943 *chromosome inversion* and *inversion chromosome*). Some entities, however, do not contain such information. Since the implementation of hash table lookup relies heavily on exact spelling and hence the completeness of lexical variations, we augment our subset of UMLS by generating missing variations.

The first augmentation concerns plural forms of nouns. We use WordNet [43] to detect entity names that end with an English noun, and then use MorphAdorner[6] to generate the noun's plural form. The entity C0751248 *M'Naghten rule*, for instance, is augmented with the additional entity name *M'Naghten rules.* This procedure generates 439k entity names.

[6] `morphadorner.northwestern.edu`

The second augmentation concerns verb conjugations. Starting from entities featuring a single word, we first use WordNet to check that it is an English verb. Then we verify that the corresponding entity belongs to UMLS-defined semantic groups that feature verbs. We choose *Activities & Behaviors*, *Concepts & Ideas*, *Phenomena*, *Physiology*, and *Procedures* as the qualifying semantic groups, such that, for instance, a gene named *CASH* is disqualified. After passing these tests, we apply MorphAdorner to conjugate the verbs. Care is taken not to incorporate a generated variation when such an entity name already exists in UMLS, because a pre-existing entity may well represent a semantically different entity. For instance, although from C0175735 *shear* the medical device we could generate *shearing*, that entity name already exists as entity C0205013 the therapeutic procedure. In this case, *shearing* is not used to augment the medical device entity. This procedure generates 4,614 entity names.

In summary, the entire dictionary construction procedure compiles a total of 3.1m unique <entity, entity name> pairs.

4.3.2 Locality Sensitivity Hashing and MinHash

Locality Sensitivity Hashing (LSH) [22] is a probabilistic method that reduces the dimensions of a high-dimensional dataset. Intuitively speaking, similar items in the dataset are hashed with high probability to the same bucket. Formally speaking, an LSH scheme is a distribution on a family $\mathcal{F}$ of hash functions π's operating on a collection of objects, such that for two objects x and y:

$$P_{\pi \in \mathcal{F}}[h(x) = h(y)] = sim(x, y)$$

where $sim(x, y) \in [0, 1]$ is some similarity function defined on the collection of objects.

MinHash (min-wise independent permutations) [19] is often employed as the hash functions π's since one MinHash scheme is itself a family of permutations. In addition, these permutations operate on objects that are sets. Formally speaking, $\mathcal{F}$ is a min-wise independent if, for any set $S \subseteq [n]$ and any $s \in S$, when π is chosen at random in $\mathcal{F}$, the following condition holds:

$$P\Big(min\big\{\pi(S)\big\} = \pi(s)\Big) = \frac{1}{|S|}$$

In other words, the condition requires that all the elements of any fixed set S have an equal chance to become the minimum element of the image of S under π.

[19] shows an additional property of LSH, namely, that the probability of two sets S_1 and S_2 being hashed to the same bucket is exactly their Jaccard similarity:

$$Jaccard(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

In our implementation of MinHash, we encode a string as the set S of character trigrams of the string. Let π_1, π_2, $\ldots \pi_k$ be hash functions from k independent min-wise permutations, such that each π maps trigrams to integers. Then $\pi(S)$ is the set of integers thus mapped, and let $min\{\pi(S)\}$ be the smallest integer in this set. The hash value, or signature, of the string is a concatenation of these smallest integers from each permutation:

$$min(\pi_1(S)) \oplus min(\pi_2(S)) \oplus \cdots \oplus min(\pi_k(S))$$

The concatenation operator is implemented as simple arithmetic summation. [19] further shows that, for each permutation π:

$$P\Big(min\big\{\pi(S_1)\big\} = min\big\{\pi(S_2)\big\}\Big) = Jaccard(S_1, S_2)$$

4.3.3 Dictionary Lookup via String Similarity

Precomputing MinHash tables. Given two highly similar strings and one MinHash scheme or one MinHash table, there is still a probability of $1 - Jaccard(S_1, S_2)$ that the two strings are not hashed to the same bucket. Therefore to increase recall, we employ multiple MinHash tables for the same collection of entity names in the constructed dictionary. When looking up text mentions at runtime, we take the union of matching dictionary entity names from all MinHash tables.

The bulk of computation is in the setting up of the MinHash tables, namely, the selection of random permutations in the form of trigram-to-integer mappings, as well as the hashing of signatures. This computation, however, is invariant to the text mentions to be looked up, and can be precomputed ahead of time.

False positive pruning. Due to the probabilistic nature of MinHash, collisions of different entity names being hashed to the same bucket are inevitable, leading to false positives. We detect false positives by comparing the Jaccard similarity of character trigram sets between a text mention and a dictionary entity name. When the Jaccard similarity scores below a threshold, the entity name is discarded as a false positive.

4.3.4 Selecting Text Mentions for Lookup

Fast and robust dictionary lookups contribute to only half of the success. The other half comes from the module that selects text mentions for lookups. We present two strategies that adhere to the theme of minimizing NLP machinery.

Consecutive words. This strategy is nearly NLP-free: Simply take consecutive words as text mentions, and use heuristics to trim the text mentions or discard undesirable ones. We start with the word length of 1, or every single word. Discard the word when it is a stop word, or when the word has only 3 or fewer characters. When looking at word lengths of 2 or more, remove any leading stop words. Discard the text mention when there is a punctuation dividing the words, as this indicates the text mention would not represent a coherent entity. Since the dictionary only contains entity names up to 5 words long, we also limit the length of text mentions to 5 words. Despite the large number of text mentions generated, this strategy is viable because it is fast, and the accuracy of mapped entities is taken care of by the dictionary lookup module.

Noun phrases. This strategy is NLP-light: Take only noun phrases as text mentions. We use the Stanford CoreNLP tool [207] to assign part-of-speech tags, and then use OpenNLP[7] to perform noun phrase chunking. We further identify complex noun phrases in the form of:

noun group – preposition – noun group

where a noun group is in the form of:

[[optional adverb] – optional adjective] – noun

Complex noun phrases allow us to capture text mentions like *shortness of breath* and *left lower lobe of lung*, as well as *lobe of lung* when the optional adjectives *left lower* are omitted. Notice that this strategy generates strictly a subset of those text mentions from the consecutive words strategy. This distinction addresses the question regarding the balance between precision and recall, as we want to investigate whether using more selective text mentions improves precision.

4.4 Evaluation

4.4.1 Data and Software Setup

Parameter tuning. To tune the MinHash parameters, we performed preliminary experiments. Besides reviewing precision and recall, we also wanted to minimize lookup time and bucket collisions, or false positives. We applied MinHash to UMLS dictionary entities, and looked up 500 random entity names. The optimal parameters that yielded the best results were as follows: choose k=30 from 14k permutations, project the dataset into 12m dimensions, use 2 MinHash tables to increase recall, and set the Jaccard similarity threshold for pruning false positive at 0.8.

Software implementation and hardware. We programmed the aforementioned MinHash method in Java, and ran the experiments in standard Linux machines with 8 Intel Xeon CPUs at 2.4GHz and 48Gb of main memory. The entire precomputation, including dictionary construction and the building of MinHash tables, took 30 minutes. To maximize lookup speed, the program loads the precomputed MinHash tables in main memory, at a one-off cost of 20 seconds when the program starts up.

Test documents. The collection of test documents are randomly selected from both biomedical scientific literature and layman-oriented health portals on the Web (see Table 4.2).

[7] opennlp.apache.org

Corpus	**Genre**	**Number of documents selected**	**Average number of words in one document**
PubMed MEDLINE abstracts from 2011	Scientific literature	5,000	181.47
PubMed Central full-length articles from 2011		500	3,038.01
MayoClinic	Health portal on the Web	500	1,793.47
UpToDate		500	2,570.94
Wikipedia Health Portal		500	510.05

Table 4.2: Composition of test documents

MetaMap. MetaMap is the baseline system against which we compared performances, and we took care to ensure that MetaMap achieved the best possible performance. Specifically, we loaded the MetaMap program and all of its associated data files into shared memory (shm). We cut MetaMap's runtime by half by issuing one request per document rather than one per sentence. In addition, MetaMap used the UMLS 2012AB base dataset, which corresponded to the same portion of UMLS we constructed our dictionary from. Finally, MetaMap provides scored entities for each text mention. We only used the top-scoring entity in our evaluation; where multiple entities shared the same top score, all of those entities were taken into consideration.

4.4.2 Precision

UMLS entity names sharing the same lexical form often represent semantically different entities. The text mention *medicine* can be mapped to entity C0013227 the pharmacological substance, and to entity C0025118 the occupational discipline. While our string-similarity-based method lacks the power to discern between these semantic differences, MetaMap has a WSD module that removes incorrect entities. In order to assess how much WSD contributes to the final mappings, we evaluated precision using two rating schemes. In the lenient rating scheme, as long as a text mention is mapped to at least one correct entity, we rated this text mention as correct. In the strict rating scheme, the presence of any incorrect entity would rate the text mention as incorrect. In other words, the strict rating scheme penalizes failure in WSD.

Table 4.3 shows the precision results of our program under various combinations of experimental setups. In the table's headers, *UMLS* denotes the UMLS-subset dictionary; $+P$ and $+V$ denote augmenting it with the plural nouns and verb conjugations, respectively. *sci* and *web* denote the genres of the test documents, namely scientific literature and health portals on the Web, respectively. Each cell in the table presents the precision evaluated from 100 randomly sampled text mentions. Overall, the MinHash method trails behind MetaMap in precision, though in a few settings their precision results are comparable (between 4% worse to 2% better). A few trends are observed and elaborated below.

	UMLS		**UMLS+P**		**UMLS+V**		**UMLS+PV**		**MetaMap**	
	sci	**web**	**sci**	**web**	**sci**	**web**	**sci**	**web**	**sci**	**web**
Consecutive words	0.94	0.98	0.97	0.96	0.94	0.99	0.96	0.99	0.94	0.96
Noun phrases	0.91	0.96	0.94	0.99	0.94	0.97	0.92	0.96		

(a) Lenient rating scheme

	UMLS		**UMLS+P**		**UMLS+V**		**UMLS+PV**		**MetaMap**	
	sci	**web**	**sci**	**web**	**sci**	**web**	**sci**	**web**	**sci**	**web**
Consecutive words	0.71	0.81	0.74	0.71	0.67	0.80	0.75	0.83	0.79	0.81
Noun phrases	0.73	0.78	0.78	0.83	0.64	0.79	0.74	0.81		

(b) Strict rating scheme

Table 4.3: Precision of the MinHash method and MetaMap

Lenient vs. strict rating scheme. Under the lenient rating scheme, the MinHash method scores consistently over 90% in precision. This result is expected, as MinHash finds, for a text mention, all entity names in the dictionary spelled similarly. Almost all the time, at least one of these entity names would be the entity expressed in the text mention. In fact, MinHash fails when the text mention is spelled similarly to unrelated entity names *and* when the correct entity does not offer a lexical variation similar to the text mention. For instance, the word *architecture* in the phrase *interfere[nce] with sleep architecture* is mapped to entity C0003737 the occupation, the only entity name in the dictionary with that spelling.

Naturally, both the MinHash method and MetaMap lose precision under the strict rating scheme. While the MinHash method loses between 15% to 30%, MetaMap only loses 15% across all settings. MinHash'es heavier loss can be attributed to its lack of WSD. Although the WSD machinery of MetaMap has never been explicitly published, an oblique reference [78] hints at a disambiguation method based on Journal Descriptor Indexing [73]. After years of using MetaMap, our empirical experience suggests that MetaMap has a preference for certain semantic types. The word *medicine*, for instance, is always disambiguated to C0013227 of semantic type *Pharmaceutical Preparations*, thus discarding C0025118 of semantic type *Biomedical Occupation or Discipline*. As the prior distribution of such ambiguous entity names coincide with the semantic type preferences, it is reasonable that MetaMap makes the right choices more often than not.

Consecutive words vs. noun phrases. The precision of text mentions selected via the noun phrases strategy is generally lower than that via the consecutive words strategy. We find this result surprising, as one would expect noun phrases to be text mentions whose lexical variations are likely to be found in the dictionary. Upon closer examination, it turns out that the root of the problem lies in the noun chunks identified by the chunking tool. Many such chunks are acronyms, numbers spelled out as English words, and single words high in ambiguity such as *form* and *system* –

precisely the types of text mentions a simple string-similarity-based approach does not handle well. Compared to the consecutive words strategy, the noun phrases strategy uses a higher proportion of such problematic text mentions, dragging the precision down.

Corpus genre. Precision observed in layman-oriented documents generally outperforms that in scientific ones. One contributing factor is the lack of acronym detection across multiple sentences, as scientific literature features acronyms more frequently. More importantly, we observe that sentences in scientific documents – especially abstracts – are often long with convoluted sentence structures. As soon as a text mention does not adequately express the full nature of the corresponding entity, a simple string-similarity-based lookup would fail. A common example is the listing of multiple items, as in *cell proliferation, differentiation and migration*, where *differentiation* and *migration* are incomplete text mentions, and only *cell proliferation* fully expresses the entity.

One key observation here is that noun phrase chunking does not rectify this situation. We conjecture that a better solution lies in leveraging the sentences' dependency parse trees, such that text mentions may be properly constructed before looking up the dictionary.

4.4.3 Coverage

To the best of our knowledge, although there are corpora annotated for highly focused sub-domains such as proteins and their interactions, there is none annotated with all types of biomedical entities. To provide an indication of recall, then, we rated every text mention in 30 random PubMed abstracts from the test documents. We took the union of all correct text mentions mapped by either the MinHash method or MetaMap, and let this larger set of text mentions be an estimation of complete coverage. Using the lenient and strict rating schemes, the 30 abstracts covered a total of 2,481 and 2,401 text mentions, respectively.

	MinHash consecutive words	**MinHash noun phrases**	**MetaMap**
Lenient	0.8420	0.2048	0.9105
Strict	0.7839	0.1930	0.9138

Table 4.4: Coverage of the MinHash method and MetaMap

Table 4.4 shows the coverage of the MinHash method and MetaMap for these 30 abstracts. Since augmenting the UMLS subset dictionary with verb and noun variations yielded the best precision, here we used only this UMLS + PV dictionary for the MinHash method. Again, some trends are observed and elaborated below.

Low coverage of noun phrases strategy. The most glaring observation in Table 4.4 is the low coverage of the noun phrases strategy. Although disappointing, the numbers are not surprising. In the 30 abstracts, only 13% of all words are chunked as noun phrases, and thus further taken as text mentions for lookups by the MinHash method. Compare this with the consecutive words strategy, where all words regardless of parts-of-speech are considered, and verbs in particular contribute to many text mentions. Under the consecutive words strategy, 65% of all words are eventually included in text mentions with (both correctly and incorrectly) mapped entity names.

MetaMap is capable of analyzing text mentions syntactically thanks to its heavy NLP machinery, such that “less important” words within a text mention may be skipped. Consider the text mentions *aerobic anoxygenic phototrophic bacteria* and *drug-endogenous substance interaction*. They are mapped to *aerobic bacteria* and *drug interactions* respectively, which are indeed correct entities despite losing some specificity. The MinHash method only considers a sequence of words in its entirety, and would never have found such coarser-grained entity names, further contributing to a loss in coverage.

MinHash and MetaMap complement each other. Regardless of the lenient or strict rating scheme, MetaMap achieves a stable coverage at 91%. As with precision, the MinHash method with consecutive words strategy also trails behind MetaMap in coverage, but is no rival here due to a gap of up to 13%. Notice that neither the MinHash method nor MetaMap finds every text mention in our estimated complete coverage. Let us visit some notable patterns that allude to the strengths and weaknesses of both programs, which may shed some light on why both programs pick up entities the other does not.

As mentioned, MetaMap is capable of analyzing text mentions syntactically in order to skip “less important” words. This syntactic analysis is not accurate all the time, however. Where it makes a mistake is where the MinHash method may prove complementary. Perhaps due to chunking errors, text mentions like *shortness of breath* and *pain breakthrough* do not always remain intact; MetaMap may split the text mention into shorter text mentions of single words. Consequently, single words are mapped to their own, separate entities, causing the original, longer text mention as a whole to miss out on getting mapped to more applicable entities. (Using the “term processing” option, one can force MetaMap to take a text mention as-is without splitting it, but this requires the user to provide the text mentions, which in turn requires the user to precompute some linguistic analysis.) The MinHash method’s consecutive words strategy, being blind to syntactic analysis, would always attempt to lookup all text mentions consisting of a sequence of words.

Finally, MetaMap has built-in support for acronym detection, a feature that the MinHash method does not provide. An acronym such as BAC represents four different entities, and the correct entity can only be inferred from the spelled out entity name usually appearing prior to the acronym in the same document. Similar cases contribute to text mentions the MinHash method misses but are picked up by MetaMap.

4.4.4 Throughput

We recorded the time required to apply the MinHash method and MetaMap to all the test documents. Both text mention selection strategies, consecutive words and noun phrases, were performed using the UMLS + PV dictionary. We report the average processing time from 5 repeated runs in Table 4.5.

	Number of PubMed abstracts per minute	**Number of other documents per minute**	**Number of words per minute**
MinHash consecutive words	1,720.19	175.15	339,508
MinHash noun phrases	863.22	85.42	166,533
MetaMap	15.26	1.41	2,786

Table 4.5: Throughput of the MinHash method and MetaMap

Employing almost no NLP machinery, the consecutive words strategy is the fastest. The noun phrases strategy, which uses light NLP machinery, takes twice as long as the consecutive words strategy. Given that the consecutive words strategy performs better in precision *and* coverage, let us use this setting to revisit the scenario presented in this Chapter's Motivation (Sub-section 4.1.1). Instead of 26 days, processing 600k PubMed abstracts now with the MinHash method will take less than 6 hours.

4.5 Summary

MetaMap is the de facto standard biomedical entity recognition (ER) software tool that uses much NLP machinery. At the cost of higher quality, however, is its lower throughput. In this chapter, we present an alternative, fast biomedical ER method, with the aim of achieving near-MetaMap quality at a fraction of the time. This alternative is a string-similarity-based method built upon the MinHash algorithm, operating over a carefully constructed dictionary of entity names based on UMLS. Our method's precision is comparable to that of MetaMap (between 4% worse to 2% better), though our coverage trails behind MetaMap by 13%. It appears that while heavy NLP machinery does boost precision and coverage, only a minority of text mentions benefit from it; the majority of text mentions can be accurately mapped to entities using the MinHash method alone. With running speed two magnitudes faster than MetaMap, our method makes it possible for other biomedical text mining tasks to analyze PubMed-scale corpora.

Chapter 5

Semantic Type Classification of Common Words

Complex noun phrases are pervasive in biomedical texts, but are largely under-explored in entity discovery and information extraction. Such expressions often contain a mix of highly specific names and common words. These words can have different semantic types depending on their context in noun phrases. In this Chapter, we address the task of classifying these common words onto fine-grained semantic types. For information extraction tasks, it is crucial to consider common nouns only when they really carry biomedical meaning; hence the classifier must also detect the negative case when nouns are merely used in a generic, uninformative sense. Our solution harnesses a small number of labeled seeds and employs label propagation, a semi-supervised learning method on graphs. Experiments on 50 frequent nouns show that our method computes semantic labels with a micro-averaged accuracy of 91.34%.

5.1 Introduction

5.1.1 Motivation

In biomedical texts, entities are written as natural language phrases. Previous works on information extraction (IE) in the biomedical domain have focused on short phrases that work well, for instance, with dictionary-based approaches. A typical scenario is to use the MetaMap tool [6], the most notable method, to recognize text mentions that match dictionary entity names, disambiguate them to canonical entities, and then apply further processing upon the resulting entities to extract entity-entity relations. A limitation of this line of approach is that long phrases, which are not cataloged in a dictionary, are neglected; an IE task would therefore miss out on information expressed in these phrases.

We observe that long phrases are actually ubiquitous in biomedical texts. These long phrases, however, have to date remain largely under-explored. Noun phrases that are long are inherently more complex, and, in the biomedical domain, they are often a mixture of domain-specific names (of diseases, symptoms, drugs, etc.) with common nouns such as *condition*, *degree* or *process.*

In Table 5.1 are two examples of such complex phrases. In the first example, *process* is a vital part of the phrase and carries biomedical meaning, namely, denoting a body function. In the second example, *processes* is used in the generic sense of the common noun and is relatively uninformative for the purpose of detecting biomedical entities in text. Therefore, the first challenge in addressing long noun phrases is in

Example phrase
1. monitoring of the carcinogenic process
2. development of processes for the prognosis of malaria

Table 5.1: Sample phrases containing *process*

determining whether a noun carries critical biomedical information or is used in a generic, uninformative way. For information extraction tasks like entity discovery, relation mining and knowledge base population, it is crucial to distinguish between these two situations.

Moreover, in the case of information-bearing nouns, we would like to further annotate that noun with a semantic type. Since the semantic type captures the usage of the word within the surrounding noun phrase, it is an invaluable asset for further analysis of that phrase, for example in entity disambiguation. One possibility is to adopt the UMLS (Unified Medical Language System) semantic types as the typing system. For instance, the word *reaction* in the three example phrases in Table 5.2 could be annotated with UMLS semantic types. Since *Chemical Reaction* is not a UMLS semantic type, the third example phrase can only use the broad *Phenomenon and Process* type (which is the semantic type for the entity C0596319 *chemical reaction*). In fact, with only a total of 133 semantic types to classify over 3.4 million entities, UMLS has rather coarse-grained and sometimes fuzzy types. A second possibility is to adopt WordNet senses, called synsets [43], as the typing system using techniques for word sense disambiguation [136]. Although WordNet synsets are more fine-grained, they have limited coverage of the biomedical domain and cannot adequately represent many essential biomedical semantic types. A second challenge, therefore, is to identify semantic types with a suitable level of granularity as well as adequate coverage.

Example phrase	**UMLS semantic type of *reaction***
3. a hybrid material for oxygen reduction reaction	Phenomenon and Process
4. asocial reaction related to a first-episode psychosis	Mental Process
5. hypersensitivity reaction in cancer patients receiving carboplatin	Pathological Function

Table 5.2: Sample phrases containing *reaction* and their UMLS semantic types

Text genre directly impacts the constitution of semantic types of nouns. PubMed MEDLINE abstracts, being scientific in nature and written in terse prose, have a sharp focus on biomedical content. As a result, words with non-biomedical meanings are sparse; in other words, semantic types of nouns found in abstracts are generally of a biomedical nature. When we turn to PubMed Central full-length articles, their contents still focus on biomedicine but the relaxation in article length allows the inclusion of verbose prose, English idioms, and other discussion that bring in non-biomedical usage of nouns. As we go beyond scientific literature and tap into Web content, this phenomenon of mixing non-biomedical content in a predominantly biomedical-themed document becomes very common. Patient discussion forums are at the extreme end of the spectrum regarding biomedical focus; discussion participants often ramble on about their personal life and personal problems before focusing on medical issues. Since we aim to address all these aforementioned text genres, it is mandatory that we address non-biomedical semantic types of common nouns as well. A third challenge, therefore, is to distinguish the semantic type of a noun from a mixture of both biomedical and non-biomedical types.

Addressing all three challenges simultaneously, our goal is to label common words in complex noun phrases with the most appropriate semantic type (biomedical or otherwise), or inferring that the word is merely used in a generic sense without specific biomedical meaning. We focus on a judiciously chosen list of 50 common nouns, referred to as *target nouns*, that frequently appear within long noun phrases in biomedical texts. The resulting annotations – for example, labeling *process* in *monitoring of the carcinogenic process* as body function – can in turn enhance the coverage and quality of information extraction tasks.

5.1.2 Contribution

We devise a semi-supervised method for labeling a target noun within a given noun phrase with its most suitable semantic type or tagging it as biomedically unspecific and uninformative. Our method is based on label propagation over a graph that connects noun phrases and has a small number of manually labeled seed nodes. Each distinct noun phrase becomes a node, and an edge connects two nodes that share a target noun with a weight reflecting the similarity between the contexts of the respective target noun occurrences. We then apply the MAD label propagation algorithm [200] to infer the best type labels for the target nouns in the graph's nodes.

Experiments show that our method achieves 91.34% micro-averaged and 83.57% macro-averaged accuracy over 50 frequently appearing target nouns. Moreover, our method is capable of classifying both target nouns with and without an uninformative semantic type. To the best of our knowledge, this contribution is the first work that explicitly addresses general-domain semantic types mixed in biomedical text. The 50 commons words, their fine-grained custom semantic types, and their seed phrases are released as an open dataset.

5.2 Related work

Long Noun Phrases

General domain. The semantic interpretation of complex phrases is a long-studied problem in computational linguistics, and widely viewed as a very demanding task [134, 170]. A solution to disambiguate entire noun phrases in the general domain is the KODA system [129]. It implements a knowledge-driven method with a strong focus on named entities. Entities in a knowledge base are first leveraged as RDF

(Resource Description Framework) resources. A co-occurrence matrix is built via integer linear programming so as to maximize an objective function that reflects co-occurrence amongst resource-resource pairs. The method further classifies a noun phrase as highly ambiguous, ambiguous, or non-ambiguous based on the number of RDF resources retrieved as candidate entities, and finally disambiguates the highly ambiguous noun phrases using its context.

Bendersky and Croft [10] study long phrases of a different nature; instead of grammatically correct noun phrases, they focus on long queries sent to search engines. Their goal is to identify key concepts in such long queries, so that these key concepts can be given more weight when executing the query. The crux of their method is a probabilistic model that incorporates query-dependent features (such as capitalization of query words), corpus-dependent features (such as tf-idf (term frequency and inverted document frequency) in a corpus of queries), and corpus-independent features (term frequencies derived from external statistics like Google n-grams counts).

Biomedical domain. For biomedical texts, complex phrases are an infrequently studied problem. Golik et al. [53] propose to handcraft rules based on linguistic cues to identify longer noun phrases beyond dictionary entries. Similar to our method, they are also motivated by the needs of a knowledge acquisition application. Their work makes a point in analyzing “semantically poor” terms, some of which essentially entail the uninformative semantic type we employ.

SimConcept [223] is a method that disambiguates composite biomedical named entities, which is one kind of longer noun phrases. Such a noun phrase consists of more than one entity; for example, *BRAC1/2* refers to two genes, BRAC1 and BRAC2. An interesting type of this composite phenomenon is exemplified in the example *COUP (chicken oval-bumin upstream promoter) transcription factor*, where the two individual entities refer to the same canonical entity. The authors of this system propose to address 6 types of composite entities via a CRF model using patterns based on orthographic features.

Entity Typing

General domain. ClusType [164] is a method that classifies fine-grained entity types, but that is only one of three goals of the elaborate machinery. The other two goals are to recognize entities, and to mine and cluster entity-entity relational phrases. To achieve the first goal, entity types for the clusters are determined, and the types are propagated back to the individual text mentions. The entire problem is modeled as a single graph-based optimization problem to be solved via block coordinate descent, where the objective function incorporates all three goals.

Biomedical domain. Jimeno-Yepes et al. [81] propose a method to disambiguate single words to 6 UMLS semantic groups. This work takes a major departure from mainstream approaches, in that noise is introduced into a corpus of biomedical texts by mixing in texts from bioinformatics and computer science domains. The 6 UMLS semantic groups are, however, very coarse-grained; they are CHED (chemicals and drugs), CONC (concepts and ideas), DISO (disorders), LIVB (living beings), PHYS (physiology), and PRGE (proteins and genes). To the best of our knowledge, this work is the only other existing work besides our own that actively combats non-biomedical content regarding entity type classification.

DIEL [12] targets entities in lists; an example list is contained in *get medical help if you have chest pain, shortness of breath, slurred speech, or problems with vision.* The entities are disambiguated to 4 types, namely diseases, symptoms, drugs, and drug ingredients. This method uses MultiRankWalk, a variant of label propagation different from the one employed by our method. The underlying graph is bipartite; candidate entities form one part, and the lists and text mentions form the other part. DIEBOLDS [13] is a refinement from DIEL by the same research team, this time for the disambiguation of two biomedical entity types, namely diseases and drugs. Again the method targets entities in lists. The method retains label propagation as the underlying algorithm, and uses document structure as an additional ingredient.

Word Sense Disambiguation

The problem setting closest to word usage detection is undoubtedly word sense disambiguation (WSD) of free text. For the general domain, the vast body of work has been surveyed by Navigli [136], and mature software tools such as It Makes Sense [237] covers most words. For the biomedical domain, the majority of previous works center around two WSD datasets [83, 219] that together contain 253 ambiguous words, multi-word terms, and abbreviations. In addition, multiple existing works [24, 42, 196] have proposed methods to generate labeled data. As for methodologies, vector space models [117, 172] are a common choice. Another common approach is to exploit the rich knowledge embedded in UMLS. For instance, Agirre et al. [1] and Humphrey et al. [73] leverage entity-entity relations and semantic type information in UMLS, respectively.

Semantic Relatedness Metrics

One ingredient in the method we shall present shortly, the soft variant of the context similarity between entity types, is calculated based on metrics designed for taxonomies. There are a number of metrics proposed in the general domain. In a series of works, McInnes and colleagues [118, 119, 120, 121, 122] review how these metrics describe entity-entity similarity and semantic relatedness for the biomedical domain. Specifically, they arrange entities in UMLS as a taxonomy via the parent-child relations in UMLS. These works further use WSD to elucidate how the similarity and relatedness measures for different UMLS semantic types as well as UMLS semantic groups impact the difficulty of the WSD task.

5.3 Methodology

5.3.1 Outline

Our method operates on one target noun at a time, such that the methodology described in this section is to be repeated for each target noun.

On the one hand comes the manual preparation of the custom semantic types and their seed phrases. On the other hand comes the automatic computation of similarities of noun phrase pairs. This similarity is based on *context* – a window of k words before and after the target noun in a noun phrase (for clarity purposes, we denote by *context words* those words in the window surrounding the target). This context, in turn, is captured by three features, namely word occurrences, part-of-speech tags, and entity types (again for clarity purposes, we distinguish *context entity types* that are precomputed, from custom semantic types that we want to classify). Using the seed phrases and context similarities, we cast the the noun phrases into a graph and apply the MAD label propagation algorithm [200].

5.3.2 Manual Preparation

Custom semantic types. As mentioned, neither UMLS semantic types nor WordNet synsets are satisfactory candidates for our purposes. UMLS semantic types cover biomedical concepts, but are too coarse-grained and do not contain any non-biomedical types. WordNet synsets are fine-grained, but are lacking in biomedical coverage. Therefore, we devised a small collection of fine-grained *custom semantic types* ourselves. The novelty of our custom types lies in the explicit provision for non-biomedical types, as well as the uninformative type where applicable; Table 5.3 shows both of these elements in play for the target nouns *culture* and *degree*.

Target noun	Custom semantic types
culture	medical sample social construct
degree	metric for temperature metric for bending stage in progression (e.g. second degree burn) academic degree degree of freedom in statistics edges out of a node in a graph generic, uninformative

Table 5.3: Custom semantic types for the target nouns *culture* and *degree*

For each target noun, we manually specify its applicable custom semantic types based on two criteria. First, a custom semantic type should have a discernible presence in the corpus. Second, the contexts of custom semantic types should be amenable to a learning algorithm, i.e. they should be sufficiently distinct from each other. Recall that we would also like to identify the case when the target noun is used in a generic, uninformative way. We facilitate this by adding a uninformative semantic type. We observe, however, that not all target nouns require this uninformative type. For instance, *culture* has two overwhelmingly dominant types (medical sample and social construct) such that the rest is negligible and does not need an explicit representation. This specification of custom semantic types is based on manual observation, over both the corpus noun phrases and UMLS entries relevant to the target noun. Appendix B contains the complete list of target nouns and their corresponding custom semantic types. In our compilation, one target noun has on average 3.78 custom semantic types.

Seed phrases. Once the custom semantic types are set, we nominate a few representative phrases as seed phrases. This process is again manual, where we aim for phrases which are sufficiently prevalent, and which convey the custom semantic type with high certainty. Table 5.4 shows all custom semantic types and all the seed phrases for the target noun *activity*, and the complete list is in Appendix B. In our compilation, one custom semantic type has on average 2.68 seed phrases.

Custom semantic type	**Seed phrases**
physical activity	fetal activity physical activity
body & protein process	catalytic activity disease activity inflammatory activity kinase activity
generic, uninformative	of activity of of activity in

Table 5.4: Custom semantic types and seed phrases for the target noun *activity*

5.3.3 Node Construction

Noun phrase selection. From a text corpus comprising articles from PubMed and encyclopedic Web portals, as well as discussions in patient forums on the Web, we collect noun phrases that contain the target noun. Specifically, for each sentence, we first use the Stanford CoreNLP [116] tool to determine part-of-speech tags and dependency parse tree. Then we find, based on the parse tree, the largest noun phrase sub-tree; in terms of CoreNLP data representation, such a sub-tree has either NP (noun phase) or PP (prepositional phrase) as its root.

Noun phrases as nodes. We cast each selected noun phrase as a node. We further say that the words in the node are either the target noun or the *context* words. In other words, the target noun is surrounded by context words to its left and/or right. When a noun phrase begins with the target noun, there is no left context; similarly, when a noun phrase ends with the target noun, there is no right context.

Node labels. A node featuring a seed phrase inherits the corresponding custom semantic type as the node label. We call such a node as *seed node*. All other nodes do not have labels at this point.

Context entity type estimation. We would like to assign an entity type to each context word. However, since a comprehensive entity disambiguation tool is not available, we estimate the entity types by a popularity-based approach that exploits the repetitiveness of dictionary entries and semantic assets in UMLS. First, we take note of UMLS entity names that contain a single word. Next, for each distinct entity name, we take note of the entities (distinct CUI's), as well as the number of occurrences (MRCONSO entries) represented. A few count-based heuristics determine which entity is the most popular, and the corresponding CUI's UMLS semantic type[1] becomes the word's entity type. Taking *cat* as an example, it appears 16 times as a mammal, 3 times as the abbreviation for CAT scan, and 1 time as an enzyme. Therefore *cat*'s entity type is *Mammal*, the UMLS semantic type for CUI 0007450 the mammal entity. In essence, this approach approximates the entity type with the largest prior distribution probability. Since biomedical word senses are often highly skewed [83], we believe this approach is a reasonable interim substitute to a full-fledged entity disambiguation tool.

In addition to the 133[2] UMLS semantic types, we introduce an extra type to represent measurement units such as mg/kg and μmol.

Similarity between context entity types. We investigate two variants of context entity type similarity. Under the hard variant, only the same entity type occurrences contribute towards context similarity (e.g. *Cell* and *Cell Component* would therefore be considered completely dissimilar). Under the soft variant, similar entity types also contribute (*Cell* and *Cell Component* now have a similarity of 0.9375). The similarity between two entity types A and B is:

$$0.5 \times group(A, B) + 0.5 \times lch(A, B)$$

where $group()$ returns 1 if A and B belong to the same UMLS semantic group, and 0 otherwise. $lch(A, B)$ is the similarity score between A and B in the UMLS semantic type hierarchy according to Leacock and Chodorow's method [99], normalized to range between 0 and 1. The use of $group()$ is necessary because some semantic type pairs are highly similar but far apart in the hierarchy (e.g. *Body System* and *Tissue*).

5.3.4 Edge Construction

We connect every <node, node> pair with a weighted edge, whose weight is determined by context similarity, which is in turn computed via weighted Jaccard similarities.

Weighted Jaccard similarity. Intuitively, $J_w(S_1, S_2)$ captures not only the overlap between two sets of items S_1 and S_2, but also the significance or weight of the items. In our setting, an item is a word, a part-of-speech (POS) tag, or a context entity type, and the sets are multisets, i.e. sets where repeated elements are allowed. The underlying computation operates on two vectors representing the two sets via

[1] Not to be confused with the custom semantic types in Sub-section 5.3.2. They are used independently in this work.

[2] At the time of investigation, there were indeed 133 UMLS semantic types. UMLS has since then retired 6 of them.

element-to-element comparisons. We illustrate this computation using a running example.

Suppose we have the following two noun phrases:

$$\begin{array}{c} a \;\; e \;\; b \;\; \text{target-noun} \;\; d \\ c \;\; \text{target-noun} \;\; a \;\; b \;\; e \end{array}$$

where a letter represents a context item. The corresponding sets containing the context items are:

$$\begin{array}{l} S_1 = \{a, b, \quad d, e\} \\ S_2 = \{a, b, c, \quad e\} \end{array}$$

and the distances between the items and the target noun (i.e. the offset in number of words) are:

$$\begin{array}{llllllll} \text{for } S_1 & - & a:3 & b:1 & & d:1 & e:2 \\ \text{for } S_2 & - & a:1 & b:2 & c:1 & & e:3 \end{array}$$

Notice that it is possible to have two items sharing the same distance in the same set; this happens when one item is to the left of the target noun and the other item to the right. We make no distinction between left and right context items; in other words, we let both contexts impact the similarity computation equally.

We compile vectors covering the union of all items in both sets; in our example, such a vector has 5 elements from a to e. The value of each element is a weight that depends on the corresponding item's distance to the target noun – the smaller the distance, the higher the weight. We choose the following weighting scheme:

$$\text{weight} = \begin{cases} \frac{1}{\text{distance to target noun}} & \text{if item exists in set} \\ 0 & \text{otherwise} \end{cases}$$

since, based on preliminary experiments, the inverse of distance is found to be the best weighting scheme.

We therefore compile the following two vectors, with their elements corresponding to a, b, c, d, e in that order:

$$\begin{array}{lllllllll} \text{for } S_1 & - & \overrightarrow{v_1} & = & < {}^1/_3 & {}^1/_1 & 0 & {}^1/_1 & {}^1/_2 > \\ \text{for } S_2 & - & \overrightarrow{v_2} & = & < {}^1/_1 & {}^1/_2 & {}^1/_1 & 0 & {}^1/_3 > \end{array}$$

Finally, the weighted Jaccard similarity between S_1 and S_2 is:

$$\begin{aligned} J_w(S_1, S_2) &= \frac{\sum_i \; min(v_{1_i}, v_{2_i})}{\sum_i \; max(v_{1_i}, v_{2_i})} \\ &= \frac{{}^1/_3 + {}^1/_2 + 0 + 0 + {}^1/_3}{{}^1/_1 + {}^1/_1 + {}^1/_1 + {}^1/_1 + {}^1/_2} \\ &= 0.25926 \end{aligned}$$

For word, POS, and the hard variant of context entity type, only exact matches count towards $J_w()$ item overlap (singular/plural and American/British spellings of the same word qualify as exact matches). For the soft variant of context entity type, the 1/distance weight is further scaled by the similarity score between context entity types.

Context similarity. We model the similarity between two phrases by calculating a similarity score between their contexts. Specifically, the similarity score is a linear combination of the contributions from the contexts' words, POS tags, and entity types (either the hard or the soft variant):

$$\begin{aligned} sim(\text{context}_1,\ \text{context}_2) \quad = & \quad \alpha_1 \times J_w(\text{words}_1,\ \text{words}_2) \\ + & \quad \alpha_2 \times J_w(\text{POS tags}_1,\ \text{POS tags}_2) \\ + & \quad \alpha_3 \times J_w(\text{entity types}_1,\ \text{entity types}_2) \end{aligned}$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$.

5.3.5 Label Propagation

Intuition behind label propagation. Label propagation, also known as belief propagation, has been independently proposed in two different research areas, namely community structure discovery [152] and machine learning [239]. It is a semi-supervised, iterative learning method on graphs. Some nodes, i.e. the seed nodes, in the graph are initially labeled. Informally, over the iterations, the seed nodes exert influence on their neighbors, whom in turn influence their neighbors, such that eventually all nodes become labeled. The intensity of influence is determined by the edge weight between two connected nodes.

Adsorption algorithm [8]. Formally, we start with an undirected graph $G = (V, E, W)$ where V, E, and W are nodes, edges, and real-valued weights, respectively. The weight w_{ab} between two nodes a and b reflects the similarity between them. Let $\mathcal{L}$ be the set of possible m labels; each node either has a known label based on prior knowledge, or is unlabeled prior to running the algorithm. For each node v, this initialization is formally denoted as a row vector $\mathbf{Y}_v \in \mathbb{R}^m_+$, where the l-th element encodes prior knowledge about label $l \in \mathcal{L}$. The higher the l-th value in $\mathbf{Y}_v$, the stronger the prior knowledge points to l being the correct label; a value of zero denotes no prior knowledge. An unlabeled node therefore has only zeroes in this vector; such an all-zero vector with m elements is denoted by $\mathbf{0}_m$. The goal of the algorithm is to populate an analogous row vector $\hat{\mathbf{Y}}_v$ with non-zero values.

The algorithm can be viewed as performing a controlled random walk over the graph G. This control is dictated by three possible actions: inject, continue, and abandon. For each node v, with probability p_v^{inj} the random walk stops and returns the initialization vector $\mathbf{Y}_v$; with probability p_v^{cont} the walk continues to some neighbor v' with a probability proportional to $w_{vv'}$; with probability p_v^{abnd} the walk is abandoned and $\mathbf{0}_m$ is returned, i.e. no nodes, even those with prior knowledge, are labeled. In addition, $p_v^{inj} + p_v^{cont} + p_v^{abnd} = 1$ and $p_v^{inj}, p_v^{cont}, p_v^{abnd} \geq 0$. The output vector $\hat{\mathbf{Y}}_v$ is then given by:

$$\hat{\mathbf{Y}}_v = p_v^{inj} \times \mathbf{Y}_v + p_v^{cont} \times \sum_{v':(v',v)\in E} p(v'|v)\hat{\mathbf{Y}}_{v'} + p_v^{abnd} \times \mathbf{0}_m$$

MAD label propagation. MAD [200], for Modified Adsorption, is a state-of-the-art variant of the Adsorption algorithm. The key improvement is that MAD guarantees convergence via regularization.

Specifically, MAD algorithm employs the following additional components:

- A dummy label to explicitly model ignorance, bringing the number of labels to $m+1$
- Three hyperparameters μ_1, μ_2, μ_3
- A matrix $\mathbf{M}$

such that, at initialization time, $\mathbf{Y}_v \in \mathbb{R}_+^{m+1}$ now includes an extra value for the dummy label. For nodes with prior knowledge, this extra value is 0. For unlabeled nodes, however, this extra value is 1; its initialization vector is denoted by $\mathbf{r}_{m+1}$ (all values are zero except for the dummy label, where it is 1). The diagonal of matrix $\mathbf{M}$ is defined as:

$$\mathbf{M}_{vv} = \mu_1 \times p_v^{inj} + \mu_2 \sum_{u \neq v} (p_v^{cont} w_{vu} + p_v^{cont} w_{uv}) + \mu_3$$

and is used in scaling the output vector $\hat{\mathbf{Y}}_v$:

$$\hat{\mathbf{Y}}_v = \frac{1}{\mathbf{M}_{vv}} \Big(\mu_1 \times p_v^{inj} \times \mathbf{Y}_v + \mu_2 \times p_v^{cont} \times \sum_{v':(v',v) \in E} p(v'|v) \hat{\mathbf{Y}}_{v'} + \mu_3 \times p_v^{abnd} \times \mathbf{r}_{m+1} \Big)$$

[200] shows that these modifications guarantee convergence of the algorithm.

Applying label propagation. Using the already constructed nodes and weighted edges, we build a graph. To make this graph ready for label propagation, a small number of nodes containing seed phrases become the seed nodes, and the seed phrase's custom semantic type is the label. We apply the MAD label propagation algorithm to label all the nodes, effectively classifying each node with the best custom semantic type. Recall that each target noun requires its own graph and hence separate application of MAD.

5.4 Evaluation

5.4.1 Dataset and Parameter Tuning

Corpus. Our corpus consists of documents from a diverse set of biomedical free texts: PubMed abstracts and full-length articles, encyclopedic Web pages from health portals, and online discussion forums. As a preprocessing step, each document is segmented into sentences by the LingPipe tool[3], and further tagged with POS and parsed into dependency graphs by the Stanford CoreNLP tool [116].

Selection of target nouns. In our corpus, we observe that 90% of all noun occurrences come from 5000+ distinct nouns. Since it is infeasible to study so many of them, we pick 50 highly common but semantically ambiguous ones to be our target nouns. Specifically, we first compile a list of distinct nouns that have been observed as head words in noun phrases, and rank them according to occurrence frequencies. We then go through this ranked list, starting from the most frequent nouns, and select suitable nouns until we have 50 of them. To be suitable, a noun must fulfill two

[3] `alias-i.com/lingpipe`

conditions. The first condition is that the noun must be ambiguous in nature. For instance, *virus* and *mRNA* are not selected since they rarely occur in an ambiguous way. The second condition is that the noun must not be used in a non-informative way most of the time. For instance, *kind* and *use* as nouns appear virtually all the time along the lines of *a kind of* and *the use of.* Not only it is very difficult to find meaningful custom semantic types for them, it is even more difficult to come up with sufficient seed phrases. Finally, in the selection process, we also take care to curate a mixture of nouns exhibiting different characteristics, such as different levels of ambiguity, nouns with predominantly biomedical usage as well as those with significant non-biomedical usage, and nouns sometimes used in a non-informative way as well as those that are practically never so used.

Selection of custom semantic types and seed phrases. As mentioned in Subsection 5.3.2, we manually nominate custom semantic types and their seed phrases. This process is mostly a trial-and-error exercise. Having the 50 target nouns, we sample the sentences in the corpus in which they occur. For the custom semantic types, we use UMLS entities relevant to the target nouns as a guide, and keep track of the types we encounter in the sample sentences. For the seed phrases, we select them so that, for each custom semantic type, they embody just enough of a trend for a machine learning method to generalize upon, while leaving enough non-seed phrases to genuinely test the performance of the method.

Selection of complex noun phrases. Using the same corpus, we extract the longest compound noun phrases that contain the target nouns. For each target noun, we make one collection by randomly selecting noun phrases containing that noun. The average noun phrase length across collections are relatively uniform from 13 to 17 words.

Label annotations. Two annotators annotated a random sample of selected noun phrases with the correct custom semantic types. The value of Fleiss' Kappa was 0.76, which indicates substantial inter-annotator agreement.

MAD label propagation software. We use the junto software tool[4], a java implementation released by the authors of the MAD algorithm.

Parameter tuning via development dataset. Based on preliminary experiments, $\mu_1 = 10 \times \mu_2 = 100 \times \mu_3$ were found to be the best hyperparameters for MAD.

We tuned the method's parameters using a development dataset of 1,000 randomly selected noun phrases for each target noun. Keeping the proportion of seed nodes at 5%, we obtained the best parameter setting (the α's in context similarity and window size k) for each individual noun.

[4] `github.com/parthatalukdar/junto`

Test dataset. In the test dataset, for each target noun, we further select another 10,000 random noun phrases not used in the development dataset. Here we also keep the proportion of seed nodes at 5%, and apply the best parameters per target noun. On average, 1428 and 437 noun phrases were evaluated for each target noun and for each custom semantic type, respectively. Since a graph with n nodes contains $O(n^2)$ edges, we prune low-weight edges to avoid excessively time consuming computations.

5.4.2 Accuracy

Table 5.5 lists the micro- and macro-averaged accuracy, as well as the best context settings. Table 5.6 further showcases some sample classified noun phrases for the target noun *activity*.

Target word	**#Types**	**Micro**	**Macro**	**Best context**
activity	3	0.91	0.91	WPH
administration	2	0.93	0.84	WPS
area	6	0.92	0.89	WP
body	4	0.96	0.94	WPH
case	5	0.83	0.88	WPS
concentration	4	0.95	0.98	WPH
condition	2	0.95	0.96	WPH
control	4	0.98	0.97	WPS
culture	2	0.99	0.79	WP
degree	7	0.76	0.72	WP
development	5	0.88	0.86	WP
distribution	2	0.96	0.96	WPS
effect	2	0.93	0.75	WPS
expression	4	0.96	0.81	WPH
factor	6	0.96	0.72	WP
flow	5	0.83	0.90	WPH
form	4	0.92	0.63	WPS
function	3	0.94	0.94	WPS
group	3	0.92	0.74	WPS
information	4	0.95	0.95	WPH
line	5	0.89	0.85	WPS
measure	2	0.90	0.80	WPS
mechanism	2	0.85	0.76	WPS
model	3	0.96	0.63	WPS
pattern	6	0.77	0.81	WP
period	3	0.91	0.92	WPS
point	8	0.92	0.76	WP
pressure	6	0.79	0.89	WP
problem	4	0.89	0.67	WP
process	4	0.85	0.91	WPH
product	6	0.95	0.91	WP
profile	3	0.98	0.84	WP
program	5	0.92	0.85	WPH
rate	3	0.95	0.78	WP
reaction	5	0.97	0.94	WP
reduction	3	0.72	0.75	WPS
region	4	0.90	0.50	WPS
report	2	0.99	0.97	WPH
resistance	3	0.98	0.66	WPS
response	5	0.89	0.73	WPS
result	4	0.91	0.89	WPH
role	3	0.98	0.99	WPH
sequence	2	0.97	0.95	WPS
set	2	0.98	0.97	WPS
site	4	0.96	0.85	WPH
solution	2	0.99	0.94	WPS
state	4	0.98	0.82	WP
strain	3	0.66	0.59	WPS
system	4	0.92	0.85	WPS
technique	2	0.91	0.92	WPS

Table 5.5: Number of custom semantic types, micro- and macro-averaged accuracy, and the best context setting of 50 target nouns. W, P, H, S denote word, POS, hard and soft context entity types, respectively.

Micro-averaged accuracy. Micro-averaged accuracy reflects accuracy per target noun without any further aggregation. The results are very encouraging, overall at 91.34% with only 5 lower-performing target nouns below 80%. Recalling that the average number of custom semantic types is 3.78, and only 5% of nodes need to be seeds, the results demonstrate that it is indeed feasible to use label propagation to solve our classification problem while minimizing the use of labeled nodes.

Three of lower-performing target nouns (*degree*, *pattern*, and *pressure*) have higher numbers (6 or 7) of custom semantic types. As the number of custom semantic

Custom semantic type	**Sample classified noun phrases**
physical activity	instruction in self-directed exercises and activity diaries day-to-day household activities that create the backbone of healthy environments
body & protein process	lower insulin-stimulated GS activity in PCOS patients compared with controls plasma anti-pneumococcal polysaccharide antibody activity (serotypes 3, 6a and 23)
generic, uninformative	dual activity of exploring karanjin isolation for medicinal purposes the orchestration of a set of activities that should be executed in order to deliver an output

Table 5.6: Custom semantic types and sample classified noun phrases for the target noun *activity*

types increases for one target noun, it becomes harder for the types' contexts to be sufficiently distinct from each other. This phenomenon leads to noisy edge weights in the graph, which in turn leads to poorer classification results.

The remaining two lower-performing target nouns (*reduction* and *strain*) also have weak micro-averaged accuracy despite having fewer (3) custom semantic types. In both cases here, the dominant custom semantic type is used in such a broad way that a few seed phrases are not sufficient to describe the context. Specifically, a reduction of quantity can be about just anything; an organism strain can be described at the population, experiment, organism, gene, or molecular level, or can be described via the characteristic effect the strain causes.

Macro-averaged accuracy. This accuracy is averaged across custom semantic types, i.e. it reflects the performance per type. Here, the overall accuracy of 83.68% is 8% lower than its micro-averaged counterpart. Moreover, performance is less consistent and varies across target nouns, ranging from 50% to 99%. These discrepancies imply that some target nouns have custom semantic types that are difficult to classify while occupying only a small proportion of all noun phrases.

Examination of our annotations reveals that the skew of the custom semantic types' distribution is indeed the overriding contributing factor for the inconsistent performance. The most frequent label of one target noun constitutes from 23% to 91% of occurrences. When a sparse type is represented by few labeled examples in the graph, naturally there is less generalization power to classify correctly.

5.4.3 Impact of Context Types

In terms of how much context words, POS, and context entity types contribute towards the solution, we are surprised that the use of words and POS alone are sufficient for 28% of the target nouns to achieve the best experimental setting. While the rest of the target nouns benefit to varying degrees the hard and soft variants of context entity types, it is worth noting that even a rudimentary estimation of context entity types based on UMLS empowers better context comparisons for the other 72% of target nouns.

5.4.4 Other Sources of Error

Other errors in the classification stem from two main sources. The first source is the omission of a critical cue, be it a word or a context entity type, in the context window. For instance, when the window size k is 5, there are at most 10 words that lie within the context window. Together with the target noun, there are at most 11 words actively participating in our method. Recalling that the selected noun phrases are on average between 13 to 17 words long, there are up to 6 words routinely disregarded which would otherwise provide valuable cues. In other cases, a cue may lie much earlier in the document, such that even considering the entire noun phrase would not be of help. Consider the following example:

same way as a control protein with no retrovirus cytoplasmic domain

The *control* here could be the protein's function to regulate some cellular process, or it could refer a specific choice of protein used in a controlled experiment. Only the detailed description in the document prior to this noun phrase can indicate the correct choice.

The second source of error is that significant expert knowledge is needed to put the puzzle together. Take the following noun phase as an example:

differences from subjects in PRL concentrations
related to stress, food or sleep

With mentions of *stress, food or sleep*, and not knowing what *PRL* is, one might be inclined to conclude that *concentration* here means the mental facility. As it turns out, *PRL* refers to prolactin, a hormone. An expert would undoubtedly say *concentration* here refers to density of matter. In this particular example, context entity types would even influence the label towards to the wrong custom semantic type since there are three items (*stress*, *food*, *sleep*) semantically more related to mental facility and only one item (*PRL*) to density of matter.

5.5 Summary

In this chapter, we present a semi-supervised method that classifies a word's semantic type in complex noun phrases. With 50 common words, we demonstrate that a small number of labeled seeds can enable a label propagation algorithm to assign both conventional semantic type labels as well as the negative case of uninformative label. The overall micro-averaged accuracy of 91.34% indicates that our method is suitable for the problem setting. On the other hand, the overall macro-averaged accuracy of 83.68% indicates that some semantic types remain too difficult for our method to overcome. To the best of our knowledge, this is the first work that explicitly includes general-domain semantic types as part of the classification problem.

Chapter 6

Fast Entity Disambiguation in Topically Annotated Texts

Many existing works in biomedical entity disambiguation share a limitation, in that their model training prerequisite and/or runtime computation are too expensive to be applied to all ambiguous entities in real-time. We devise an automatic, light-weight method that processes MEDLINE abstracts at large-scale and with high-quality output. Our method exploits MeSH terms and knowledge in UMLS to first identify unambiguous anchor entities, and then disambiguate remaining entities via heuristics. Experiments showed that our method is 79.6% and 87.7% accurate under strict and relaxed rating schemes, respectively. When compared to MetaMap's disambiguation, our method is one order of magnitude faster with a slight advantage in accuracy.

6.1 Introduction

6.1.1 Motivation

The ever-growing volume of biomedical literature is published at a phenomenal pace. PubMed, the de facto corpus for biomedical text mining, is currently growing at more than one million new citations per year. Text mining can unearth the rich information buried in this corpus, but in light of this phenomenal growth of corpora size, different modules in a text mining pipeline face additional challenges. The entity disambiguation (ED) module faces in particular two challenges.

The first challenge is scalability – the ideal ED solution must cope with the sheer volume of textual input. Most existing biomedical entity disambiguation methods that do address all entities cannot be applied to a large corpus in practice for several reasons. Methods based on machine learning [23, 77, 172, 197] must identify in advance an exhaustive list of all ambiguous entity names. Where the training is supervised, labeled examples must also be obtained, either by expensive manual annotation or by automatic curation [78]. Furthermore, the models must be trained prior to disambiguation at runtime, and in general one individual model is required for each ambiguous entity name. All these setup costs render the methods impractical when all ambiguous entity names must be addressed. An alternative line of methods [1, 236] generates at runtime an entire instance of the problem customized per input text. This style of massive setup cannot cope with, for instance, real-time feeds of new documents; one million new PubMed citations per year equates to one new citation every 30 seconds, not to mention any additional new Web content. Finally, although MetaMap [6], the de facto standard software tool for biomedical ED, does disambiguate amongst all entity types, its disambiguation functionality is limited or, as the authors put it, "arguably [MetaMap's] greatest weakness." Besides, MetaMap relies heavily on time-consuming natural language processing (NLP) analysis such that it is too slow for large-scale usage.

The second challenge is coverage of entity types – the ideal ED solution should be able to tackle the full spectrum of entities without limiting its scope to narrow specializations such as genes, chemicals, and diseases. For information extraction tasks such as relation mining and knowledge base construction, it is crucial to go beyond the few staple specializations so as to fully leverage all the knowledge expressed in the text. Knowing that there are many existing biomedical ED methods that address individual sub-domains only, one possible approach is to gather these methods into an ensemble. For instance, there are ample methods available for disambiguating entities such as genes, proteins, chemicals, and cells. However, one must continue to ask which sub-domains are not yet addressed in the ensemble, and what new disambiguation methods should be developed for them. Clearly, this line of approach is feasible only when most if not all sub-domains have mature disambiguation methods already at hand. That is unfortunately not the case currently, where specialties not at the molecular level, for example drug side effects and lifestyle risk factors, are particularly under-explored with regards to entity disambiguation.

In this work, we aim to tackle the the biomedical ED task with two overarching directives that address the aforementioned two challenges. To ensure that the method is fast, we aim for a light-weight approach. To ensure that the method addresses all entity types, we aim for an entity-type-agnostic approach. In other words, we intentionally investigate an approach that steers away from mainstream practices that are computationally intensive and/or applicable to single sub-domains.

We choose PubMed MEDLINE abstracts[1] and UMLS (Unified Medical Language System) [15] as our corpus and knowledge base for this work, respectively, because our method can then leverage the following unique characteristics of these biomedical resources:

- MEDLINE abstracts are a large corpus indexed with rich, manually assigned MeSH (Medical Subject Heading)[2] terms; we safely consider all MeSH terms to be accurate. In addition, since abstracts are very compactly written, their content rarely strays away from the biomedical domain. In other words, non-biomedical entities occur only rarely.
- UMLS is the authoritative and comprehensive knowledge base of the biomedical domain covering all aspects of the domain, with a vast collection of entities plus their lexical variations, semantic types, and inter-relationships.
- MeSH terms are themselves a crisp ontology that is already part of UMLS.

Putting these together: All the entities found in a MEDLINE abstract are of a biomedical nature, and all of them can be disambiguated to some canonical entity in

[1] PubMed indexes multiple sources of citations, of which MEDLINE is the primary component.

[2] `www.nlm.nih.gov/mesh`

UMLS. These observations lead us to devise an entity disambiguation approach effective on MEDLINE abstracts, keeping ever in mind that the approach must balance quality with high throughput while addressing all entities.

6.1.2 Contribution

We devise an automatic and light-weight method that disambiguates all entities in an indexed document by exploiting its indexing as well as domain knowledge. Specifically, the indexed documents are MEDLINE abstracts, which are the bulk of scientific literature in the biomedical domain. As for domain knowledge, the method draws upon UMLS. Given an abstract, its MeSH terms as ground truth, and all the text mentions in the abstract, the method first identifies unambiguous entities that we shall call *anchors*. The remaining text mentions are then disambiguated using the anchors as well as heuristics based on linguistic-semantic patterns and knowledge base assets.

Under the best setting, our method achieves an average of 79.6% and 87.7% accuracy using the strict and relaxed rating schemes, respectively. In terms of throughput, our method processes 240k abstracts containing 24.5m text mentions in 400 minutes. We also present evaluations against established gold standards via a comparison to MetaMap. To the best of our knowledge, this is the first work in the biomedical domain that evaluates all text mentions found in an abstract. The resulting code has been released as an open source software.

6.2 Related Work

Sub-domain-agnostic Entity Disambiguation

In the biomedical domain, the terms entity disambiguation (ED) and word sense disambiguation (WSD) are often used interchangeably, since the distinction between entity and sense is not always clear-cut. ED refers mostly to named entities. WSD refers mostly to single words that may be a spelled out term (such as *expression* and *Astragalus*) or an abbreviation (such as *TMJ*); clearly, *Astragalus* is a named entity, too. Existing methods may emphasize their ED or WSD aspect based on their choice of gold standard.

Gold standards. Two gold standards, NLM WSD [219] and MSH WSD [83], are widely used. As their names suggest, both emphasize the WSD aspect although their annotations do include named entities, especially in MSH WSD. Both gold standards feature ambiguous text mentions taken from MEDLINE abstracts.

NLM WSD contains 50 ambiguous terms. Each ambiguous text mention is annotated by at least four human experts, upon which a final annotation is reviewed and harmonized manually. The distribution of senses has not been adjusted, so that the relative proportions of annotated senses reflect the actual distribution in a large corpus.

MSH WSD contains 203 ambiguous terms. In contrary to NLM WSD, which is the culmination of substantial expert effort, the MSH WSD gold standard is constructed automatically. Specifically, the 203 terms are chosen so that each has multiple, corresponding MeSH terms. MEDLINE abstracts containing both a text mention of the ambiguous term and one of the predefined MeSH terms are chosen, with the text mention inheriting the MeSH term as its disambiguated entity. As another difference to NLM WSD, the distribution of senses in MSH WSD are adjusted so that each sense has the same number of annotations.

Fan and Friedman [42] also similarly generated a less often used gold standard via co-occurring ambiguous text mentions and MeSH terms in MEDLINE abstracts. This gold standard contains 59 ambiguous terms, and the authors divide them into 3 categories that reflect levels of disambiguation difficulties. The three categories are (1) two senses related to two different UMLS semantic types, hence easy to disambiguate, (2) two senses from two types that are known difficult cases, such as a pathogen and the disease it causes, and (3) two senses with the same type, hence the most difficult to disambiguate.

Knowledge-based methods. Existing works leveraging knowledge bases (KB's) generally select certain KB's or certain elements in a KB amenable to the respective methodology. The construction of a custom knowledge graph is the backbone of a collective inference approach by Zheng et al. [236]. This graph is built using the 300 biomedical ontologies of the BioPortal. The approach ranks candidate entities based on an entropy metric derived from the knowledge graph, and then disambiguates multiple entities simultaneously via collective inference.

Agirre et al. [1] also use a graph-based approach. Recalling that UMLS is a collection of many heterogeneous resources, they opt to use the co-occurrence information (encoded in the MRCOC table) and relations information (MRREL table) of UMLS to build their custom knowledge graph. Then they cast the entity disambiguation problem as an instance of the Personalized PageRank algorithm, and study the impact of using different subsets of UMLS knowledge.

Stevenson and Guo [195] leverage UMLS assets to generate labeled examples for a memory-based learning algorithm that operates over a vector space model. Specifically, they define two types of entity-to-related-entity relations as features of the model. The first type is co-occurrence relations based on UMLS co-occurrence information. The second type is monosemous relations based on entity names that occur only once in the entire UMLS.

Exploiting MeSH terms. Since MeSH terms in MEDLINE abstracts are expert-assigned, they provide high quality clues when disambiguating entities in the abstracts. Several existing methods exploit this arrangement. In a series of works, Stevenson and colleagues [193, 195, 197] use MeSH terms as features in their various machine-learning-based methods. Recall also that two the aforementioned gold standards (MSH WSD and the one by Fan and Friedman) are also constructed by exploiting MeSH terms as ground truth.

Machine-learning-based methods. Machine-learning-based methods, both supervised and unsupervised, dominate existing works that address all entity types. Apart from the three works exploiting MeSH terms just mentioned, the most recent work by Jimeno-Yepes [77] uses long short-term memory in a recurrent neural network model. This method employs an extensive battery of features, including orthographic and word-level features such as stemming, unigrams and bigrams, concept-level features such as UMLS identifiers and UMLS semantic types extracted from MEDLINE abstracts, and global features such as word embeddings. Chen et al. [23] apply active learning to support vector machine (SVM). They consider three algorithms – least confidence, margin, and entropy – when selecting the next round of instances for the active learning. Finally, four further methods are compared by Jimeno-Yepes and Aronson [78].

Sub-domain-specific Entity Disambiguation

When it comes to disambiguating only highly specialized entity types, a large body of works exists. The most representative specializations are genes and proteins, the major works for which have been compared in a survey [161]. Other specializations include species of genes [64, 218], chemicals [9, 103], and diseases [36, 75, 88]. A common theme amongst these works regardless of sub-domain is that their methods address primarily named entities.

Rule-based Entity Disambiguation

Two existing methods [75, 88] use a rule-based approach to disambiguate disease names. In [75], the rules are arranged in a decision tree manner, so that as soon as an entity name matches a known synonym, or as soon as a search engine retrieval score is beyond a threshold, the algorithm stops and declares a candidate entity correct. In [88], however, the 5 rule modules function independently, and any number of them can be employed simultaneously in a plug-and-play manner. The rules involved are not about making the final yes/no decision as in [75], but are rather rules implementing various natural language processing tasks, such as adjusting noun phrase boundaries, performing coordination resolution, and filtering out false positives via a word overlap measure.

To the best of our knowledge, the method we present in this Chapter is the only other rule-based entity disambiguation method; as a corollary, it is also the only rule-based method that addresses all entity types.

6.3 Methodology

6.3.1 Outline

The input to our method is a MEDLINE abstract and its MeSH terms. We use the fast dictionary-based entity recognition presented in Chapter 4 to identify all longest text mentions that match UMLS entity names. Then the method proceeds in two phases. Phase 1 applies two heuristics to identify unambiguous anchors amongst the text mentions. Phase 2 leverages the anchors as well as five further heuristics to disambiguate the remaining text mentions.

6.3.2 Phase 1: Identifying Unambiguous Anchors

A text mention becomes an *anchor* when the method determines that it has one UMLS entity that underlies this text mention unambiguously. A text mention may become an anchor via one or both of the following heuristics.

Heuristic 1 – MeSH term (MESH). Medical Subject Headings (MeSH) are a crisp and concise ontology developed by the United States National Library of Medicine, the same institution that maintains PubMed citations. In the 2016 version, there are 27,883 entries called Descriptors in the ontology. They are arranged in a taxonomy containing 16 trees such as *Anatomy*, *Diseases*, *Phenomena and Processes*, and *Analytical, Diagnostic and Therapeutic Techniques, and Equipment.*

When a new scientific publication joins the PubMed collection, human experts manually assign MeSH terms to that publication as a means of indexing it. This allows us to assume that these MeSH terms are accurate ground truth, which is the crux of this heuristic. When a text mention is also a MeSH term for the abstract, this heuristic declares the text mention as an anchor. This strategy is similar to the one used in MSH WSD [83], which selects abstracts via co-occurrence of MeSH term and text mention to curate the gold standard.

We note that there are usually more MeSH terms than our method can leverage, but the phenomenon does not affect our heuristic. As an indexing service, most of the time only a publication's abstract is retained as the PubMed citation. The publication, however, is a full-length article, and the experts assign MeSH terms that fully characterize the entire article. As a result, experts may and often do assign MeSH terms that do not appear in or cannot be inferred from the abstract. A typical example is the term D006801 Human; a scientific study about some human genes generally do not explicitly mention the species in the abstract. Since this heuristic is driven by text mentions, having extra MeSH terms that do not match any text mentions do not cause any error.

Heuristic 2: Only one UMLS match (ONE). UMLS is the largest and most authoritative metathesaurus of the biomedical domain. It contains 3.4 million entities, each complete with its synonyms and lexical variations. We therefore assume, for the purposes of this heuristic, that UMLS has complete coverage of all biomedical entities. Under this heuristic, a text mention that matches only a single UMLS entity is considered unambiguous.

In this work, due to license restrictions imposed by UMLS, we use only the license level 0 subset of UMLS since it offers unrestricted usage. However, this heuristic works the same way for larger subsets. For instance, researchers based in the USA can incorporate the level 4 subset, which offers unrestricted use within USA.

Fixing anchors. At the end of phase 1, we consider all anchors already correctly disambiguated. Their underlying canonical entities are final for the rest of the method.

6.3.3 Phase 2: Disambiguating Non-anchor Entities

In phase 2, the method disambiguates any remaining, non-anchor text mentions. Recall that the entity recognition tool already provides multiple matching UMLS entities to such a text mention. Taking these UMLS entities as candidates, the method selects one candidate using one or more of the following heuristics.

Heuristic 3: Singular/plural (SP). This heuristic embodies the famous one-sense-per-discourse assertion [232]. Specifically in our problem setting, since abstracts are very short documents, we assume that, within one abstract, text mentions sharing the same surface string also share the same entity. Therefore, singular and plural forms of the same word should refer to the same entity. In UMLS, when the plural form is a unique entity, that same entity is extended to the singular form, and vice versa.

For example, when *diets* is a unique entity (C0012155 food and drink consumption), *diet* is disambiguated as the same entity, thus casting other candidates (such as C0012159 food-based therapy, C1549512 food supply, among others) aside. This heuristic exploits the popularity of an entity implicitly encoded in UMLS – the more popular an entity is, the more dictionary sources will contain it, hence the more likely the singular and plural forms are both captured by these dictionaries.

Heuristic 4: Linguistic-semantic pattern (PAT). This heuristic combines three ingredients in a bigram – an anchor, the anchor's UMLS semantic type, and part-of-speech tag of the non-anchor word – into one pattern. Figure 6.1 depicts all these ingredients in action.

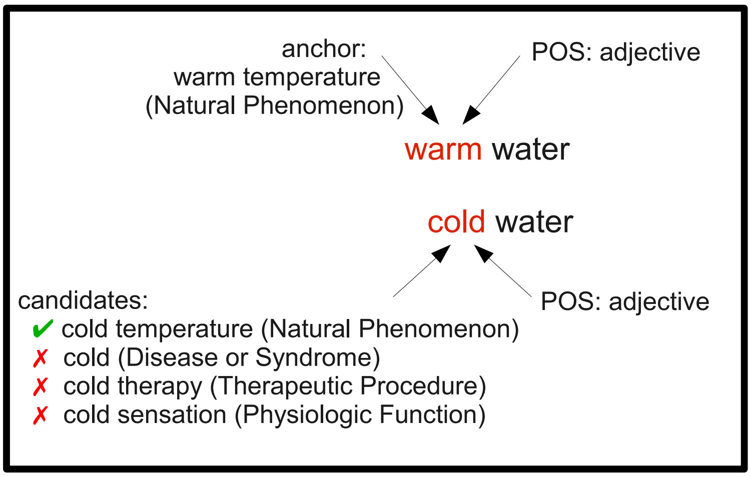
anchor:
warm temperature
(Natural Phenomenon)
POS: adjective
warm water
cold water
candidates:
✔ cold temperature (Natural Phenomenon)
✗ cold (Disease or Syndrome)
✗ cold therapy (Therapeutic Procedure)
✗ cold sensation (Physiologic Function)
POS: adjective

Figure 6.1: The linguistic-semantic pattern heuristic

For example, the two bigrams are *warm water* and *cold water*. When one word (*water*) appears in both bigrams in the same position, and when the other words (*warm* and *cold*) have the same part-of-speech, *warm* and *cold* ought to share the same linguistic function and some analogous meaning. Since *warm* is an anchor, this heuristic takes its UMLS semantic type (Natural Phenomenon), and pick for *cold* a candidate with the same type (cold temperature the Natural Phenomenon).

Heuristic 5: Co-occurring semantic types (CO). The intuition behind this heuristic is that objects of the same semantic type often co-occur in the same abstract. This heuristic identifies the most prevalent UMLS semantic type in the abstract, and selects, whenever possible, a candidate of the same type. As a corollary, when none of an ambiguous text mention's candidates has a matching, most-prevalent UMLS semantic type, this heuristic cannot be applied.

For example, an abstract mentioning different fish species naturally also mentions the word *fish*. However, the candidates for *fish* belong to different UMLS semantic types (Fish, Gene or Genome, Organic Chemical (for fish extract), and Molecular Biology Research Technique (for Fluorescence in situ Hybridization)). When the entities in the abstract exhibit a predominant semantic type (such as Fish), this heuristic picks the candidate (fish the animal) with the same type.

Heuristic 6: Ranked preferences of dictionary sources (RANK). UMLS comes with a predefined preference list of dictionary sources. When a text mention matches multiple candidates, each candidate's dictionary source leads to a corresponding rank number. This heuristic picks the candidate with the best rank.

More specifically in terms of how UMLS organizes its data, a single entity name is listed separately for each dictionary's contribution, corresponding to multiple entries or rows in the MRCONSO table. In each row, we note the values in the SAB column (for source name abbreviated, i.e. the dictionary) and the TTY column (for term type). From the MRRANK table, using this <SAB, TTY> value pair, one can look up the corresponding RANK column, which is an integer. The higher this rank integer, the better the rank.

For example, *HIV* has 15 matching entries or rows in MRCONSO table. The best ranked entry comes from the MTH (Metathesaurus) dictionary and the PN (preferred name) term type, preferring *HIV* the virus over *HIV* the infection and *HIV* the vaccine.

Heuristic 7: Prior probability (PRIOR). Thanks to the heterogeneous nature of UMLS, the listing of entity names contains much redundancy. Recall that a single entity name is listed separately for each dictionary's contribution. A more popular meaning of the word exists in more dictionaries, and hence appears in more rows of the MRCONSO table than a less popular meaning. The prior probability distribution of candidates is thus estimated based on counts of entity name occurrences in the MRCONSO table. The results in Chapter 5 show that even when rudimentarily estimated in this way, prior probabilities contribute to enriching disambiguation contexts 72% of the time. Here, the heuristic picks the candidate with the highest estimated prior probability.

For example, *cat* the animal appears in 16 rows of the MRCONSO table, more so than other less popular meanings (4 rows for the taxonomic family of the animal, 3 rows for the gene, 3 rows for the scan procedure, and 1 row for the enzyme). Therefore, *cat* the animal is estimated to have the highest prior probability.

6.4 Evaluation

6.4.1 Data and Software Setup

Development and test datasets. We used disjoint sets of MEDLINE abstracts published in 2014 as the development and test datasets. The test dataset, in particular, consists of 20 randomly selected abstracts. We use the fast dictionary-based entity recognition presented in Chapter 4 to identify all longest text mentions that match UMLS entity names; we used the UMLS 2015AB version as the underlying dictionary. In total, 2,549 text mentions were recognized in the 20 test abstracts.

Entity annotations. For the test dataset abstracts, two annotators evaluated all the recognized text mentions, including the anchors, rating the candidates as “completely correct”, “partially correct”, or “completely wrong”. The inner-annotator agreement, calculated as Cohen’s kappa, was 0.64, which indicates mostly substantial agreement. The presence of fine shades of the same underlying entity in UMLS prompted the “partially correct” annotation choice. For instance, *children* exists as two separate entities with the semantic types Age Group and Family Group, and the exact distinction is difficult even for human judges. We therefore present results in two rating schemes: Under the strict rating scheme, only “completely correct” annotations count as correct; under the relaxed scheme, both “completely correct” and “partially correct” annotations count as correct.

Software implementation and hardware. We implemented our method in java, and ran the experiments in a standard linux machine with 8 Intel Xeon CPUs at 2.4GHz and 48Gb of main memory.

MetaMap When running experiments with MetaMap, we ensure that it is using the same UMLS 2015AB dictionary, as well as the same linux machine. We further ensure that MetaMap is set up to achieve the best possible performance. This is achieved, firstly, by using the latest 2016 software release. Secondly, we issue a single query to MetaMap per abstract rather than per sentence. This arrangement alone cut MetaMap’s processing time by half. Thirdly, we enable the relevant MetaMap settings so as to maximize its overall disambiguation performance; these settings enable the use of a strict model, inclusion of all derivational variants, inclusion of all acronym and abbreviation variants, and the use of the word sense disambiguation module. The overall effect is that for each text mention, only one disambiguated entity is returned.

6.4.2 Ablation Study of Heuristics

In order to investigate the accuracy and contribution of each heuristic, we performed an ablation study. Table 6.1 shows the results. An aggregated accuracy number is generally not the sum of its contributing numbers because multiple heuristics may be applicable to the same text mention.

	Anchors		Non-anchors		All text mentions	
Heuristic(s)	**Strict**	**Relaxed**	**Strict**	**Relaxed**	**Strict**	**Relaxed**
MESH	0.160	0.169	not applicable		0.089	0.094
ONE	0.833	0.850			0.465	0.475
MESH + ONE	0.903	0.930			0.504	0.519
MESH + ONE + CO	remains at 0.903	0.930	0.472	0.723	0.713	0.839
MESH + ONE + PAT			0.078	0.099	0.539	0.563
MESH + ONE + PRIOR			0.539	0.689	0.742	0.824
MESH + ONE + RANK			0.632	0.779	0.785	0.863
MESH + ONE + SP			0.186	0.221	0.586	0.617
Best successive filtering	remains at		0.662	0.790	**0.796**	0.868
Best majority voting	0.903	0.930	0.643	0.811	0.788	**0.877**

Table 6.1: Contribution of different heuristics to accuracy

Heuristics for anchors. While MESH correctly identifies 16% to 17% of the anchors, ONE is by far the primary contributor identifying over 80% of the anchors. Together, both heuristics identify over 90% of the anchors.

We note that over half of all text mentions are actually anchors; on average, 56% of all text mentions in an abstract are anchors. In other words, while entity disambiguation is always described as a difficult problem as a whole, the *proportion* of difficult cases is also an important aspect of the problem. Taking the distribution of our anchors as a guide, roughly 56% of the time the disambiguation task is relatively easy, which of course does not detract from the level of difficulty for the remaining 44%.

Since MESH is the only heuristic that relies on expert annotated MeSH terms, and since its contribution is relatively small, one can consider applying the method to texts that do not have indexing terms. Indeed, in the open source software we released, the user can use any text as input, such that when there are no MeSH terms available, the MESH heuristic remains inactive.

Heuristics for non-anchors. Recalling that these heuristics are applicable only after the anchors are identified, we present therefore accuracy for MESH, ONE, plus a single non-anchor heuristic. The non-anchor heuristics have a wide range of contribution, from less than 10% for PAT to over 70% for RANK.

PAT and SP contribute relatively little because they are often not applicable. For the PAT heuristic to fire, the text mention must fulfill all three requirements – matching anchor, matching part-of-speech, and matching UMLS semantic type. Therefore it is not surprising that PAT is the least used heuristic, which in turn leads to being the least contributing heuristic. The SP heuristic exploits either incompleteness in UMLS or linguistic usage of specific words; for instance, *blacks* in plural can only refer to the population group and not the visual color. Since both conditions are infrequent, naturally the SP heuristic is less used.

Of the remaining three heuristics, CO, PRIOR, and RANK, it is only possible for CO to be not applicable once in a while; this happens when the most frequent UMLS semantic type in the abstract is not represented in any candidate of a given text mention. PRIOR and RANK, being reliant only on UMLS, always suggests some best candidate. Not only do these three heuristics fire often, they also contribution much to accuracy.

Strict vs. relaxed rating schemes. Using the two different rating schemes have little effect on anchor accuracy. This is not surprising, as anchors are relatively uncontroversial or "easier" as noted earlier. Non-anchors, on the other hand, get rather different accuracy to varying degrees. While going from strict to relaxed scheme boosts the accuracy of PAT and SP by only 2% to 3%, the boost to other non-anchor heuristics can be 15% for PRIOR and RANK, and even as high as 25% for CO. The "partially correct" entities are often the very fine-grained ones such as the *children* as Age Group and *children* as Family Group example mentioned above. When the annotators withhold from assigning "completely correct" gold labels, naturally the corresponding entities will only be considered correct under the relaxed rating scheme, hence the big increase in accuracy.

Putting all the heuristics together. We experimented with two types of ensembles, namely majority voting and applying heuristics as successive filters similar to D'Souza and Ng [36]. Under the relaxed rating scheme, majority voting consistently performed better. The best ensemble used, as expected, all heuristics to reach 87.7% accuracy. Under the strict rating scheme, on the other hand, successive heuristic filters consistently performed better. The best ensemble scored 79.6% accuracy using the following order of heuristics: MESH, ONE, SP, RANK, PRIOR, PAT, CO.

Figure 6.2, showing some sample disambiguation outcomes taken from MEDLINE abstract 24188907, illustrates our method at work under majority voting. In the Figure, underlined and overlined text mentions are anchors and non-anchors, respectively. A green tick and a red cross indicate that our method selects the correct and incorrect entity, respectively. We highlight in particular the following observations.

- One or both of the anchor heuristics, MESH and UMLS, may be applicable to the same text mention, as indicated by the heuristic abbreviations in the gray boxes. As evidenced by these sample outcomes, anchors are in general highly accurate.

- For the non-anchor text mentions, the ratio in the gray box indicates the number of heuristics voting for the correct entity to that voting for an incorrect entity. Not all 5 heuristics are always applicable, since only PRIOR and RANK rely purely on information in UMLS and that information is available to all entities. Most of the time, however, 3 to 4 heuristics are applicable, and occasionally all 5 heuristics are applicable (such as the case for *morbidity* in the sample).

- The non-anchor heuristics sometimes vote unanimously for the correct entity, and sometimes vote for different entities. The strength of our method is the combined effect that the 81% of the time, more heuristics vote for the correct entity than otherwise.

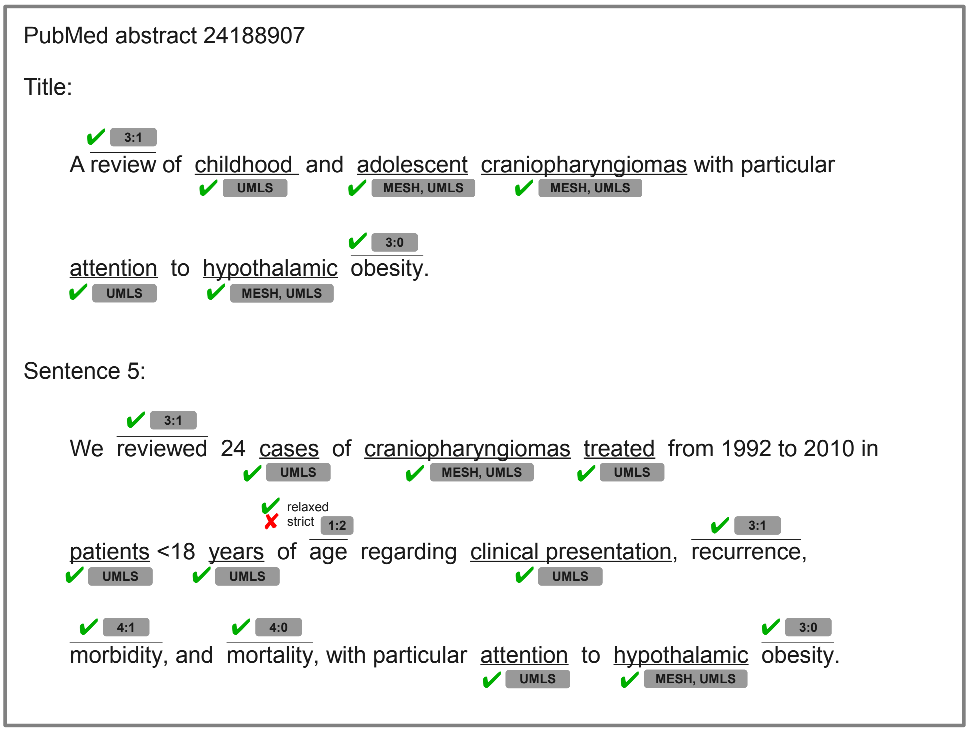
PubMed abstract 24188907
Title:
3:1
A review of childhood and adolescent craniopharyngiomas with particular
UMLS
MESH, UMLS
MESH, UMLS
3:0
attention to hypothalamic obesity.
UMLS
MESH, UMLS
Sentence 5:
3:1
We reviewed 24 cases of craniopharyngiomas treated from 1992 to 2010 in
UMLS
MESH, UMLS
UMLS
relaxed
strict
1:2
3:1
patients <18 years of age regarding clinical presentation, recurrence,
UMLS
UMLS
UMLS
4:1
4:0
3:0
morbidity, and mortality, with particular attention to hypothalamic obesity.
UMLS
MESH, UMLS

Figure 6.2: Sample disambiguation outcomes taken from MEDLINE abstract 24188907

- For the text mention *age*, RANK votes for entity C0001179 for number of years elapsed since birth; this entity has the UMLS semantic type of Organism Attribute. On the other hand, PRIOR votes for entity C1114365 for a specific point in time; this entity has the UMLS semantic type of Clinical Attribute. CO becomes the tie breaker – since the most prevalent UMLS semantic type of this abstract is Clinical Attribute, CO also votes for the specific point in time. The 2-to-1 vote declares that the Clinical Attribute entity is the final selection, which is still considered correct under the relaxed rating scheme. This outcome would be considered incorrect under the strict rating scheme, further highlighting how fine-grained UMLS entities can be, and how determining the exact disambiguated entity can be difficult. Finally, notice that none of the three applicable heuristics votes for a completely incorrect entity, such as C0162574 the protein product.

- Since our method does not treat text mentions differently according to their relative positions in an abstract, occurrences of the same text mention within the same abstract always share the same outcome. As an implementation detail, for each abstract, the method caches disambiguation results (e.g. for *craniopharyngiomas* and *obesity* in the title) so that a text mention occurring for the second time (the same two mentions in sentence 5) requires only a fast cache lookup. This arrangement further improves our method's processing time.

6.4.3 Comparison with MetaMap and Other Datasets

We compared the best setting of our method with MetaMap using the aforementioned custom test dataset as well as 3 other datasets. Table 6.2 shows the accuracy for both systems.

	Custom strict	**Custom relaxed**	**NLM WSD**	**EBI disease**	**CRAFT subset**
Our method	0.796	0.877	0.399	0.873	0.388
MetaMap	0.681	0.761	0.337	0.784	0.330

Table 6.2: Comparison of accuracy between our method and MetaMap

Selection of gold standards. In the selection of existing gold standards, we face some constraints. First, the text must be MEDLINE abstracts, since our method relies on MeSH terms. Second, the annotated gold labels must go beyond identifying an entity type (such as those proliferated by the GENIA corpus, namely genes, proteins, chemicals, DNA's, RNA's, cells, and cell lines) to pinpoint an exact entity. Third, this exact entity must be from the UMLS, so that our UMLS-specific heuristics can be applicable. Finally, as much as possible, the gold standard should address all entities and not only those from certain sub-domains.

Under these constraints, we identified three gold standards.

- **NLM WSD** [219]: Although this dataset is geared towards word sense disambiguation, there are 203 ambiguous abbreviations and terms representing a good range of entities.
- **EBI disease corpus** [80]: Although this corpus features only diseases, it complements the other two gold standards' coverage of entity types.
- **Subset of the CRAFT corpus** [7, 29]: The complete corpus contains text mentions annotated with entities from 7 different biomedical ontologies focusing on entity types such as chemicals, genes, proteins, cells, and species. Of the 7 ontologies, only the Gene Ontology and NCBI Taxonomy provide mappings to equivalent UMLS entities. Therefore we used the subset of the annotations that can be traced to UMLS entities. This subset features 4 sub-domains: biological processes, cellular components, and molecular functions based on Gene Ontology, as well as species based on NCBI Taxonomy.

Notice that the MSH WSD dataset [83] was not used here because it was essentially constructed with the MESH heuristic; using this dataset would not offer further insight.

Comparison in accuracy. In terms of accuracy, both systems showed analogous trends for each dataset, though our method outperformed MetaMap by 5% to 11%. Both systems performed poorly over the NLM WSD and CRAFT datasets due to their wide variety of highly ambiguous entity names. The disambiguation module in MetaMap is known to be a weaker module in the system [6], while our method's heuristics are too simplistic for sophisticated cases. For example, in the excerpt *the hypothesis is that the protein is localized to these vesicle membranes*, the word *localized* is referring to a specific cellular process that fixes a protein at a cellular location. The correct entity, C1744691 establishment and maintenance of cellular component localization, has its entity name much more detailed than the single word *localized*. While a human expert can easily infer from the rest of the excerpt that *vesicle membranes* implies a cellular location, our method and MetaMap both disregarded this information, and simply gravitated towards the only but incorrect UMLS entity with an exact string match (C0392752 localized, in the generic sense that an object is localized in some area). The same rationale explains why accuracy in EBI disease corpus was high, because disease names are much less ambiguous in general.

Comparison in throughput. We recorded the time required to process the same datasets for our method and for MetaMap, disregarding any time spent on preprocessing dictionary (for our method) or loading data into main memory (for both methods). For both methods, the reported time reflect the processing from text input to disambiguated entities; in other words, determining text mentions and entity recognition are included in this processing time.

In terms of throughput, our system and MetaMap processed 600 and 11 abstracts per minute, respectively. MetaMap is known to employ much NLP machinery that is time consuming. On the other hand, not only does our method perform entity recognition quickly via our fast dictionary-based method, entity disambiguation is also fast since the heuristics employs zero or minimal NLP machinery. Consequently, as MetaMap remains unsuitable for large-scale text mining, our method makes handling PubMed-scale corpus and other real-time processing tasks possible.

6.4.4 Distribution of Entity Types

Now that we are armed with an entity disambiguation method that applies to all entity types, we can analyze the distribution of entity types in different genres of input text. Specifically, we can apply the best setting of our method (majority voting using all heuristics) to three genres, namely scientific literature, encyclopedic Web portals, and online discussion forums. As mentioned earlier, when no MeSH terms are available, the MESH heuristic remains inactive; such is the case for Web portals and discussion forums. Table 6.3 shows the summary by grouping entity types into coarse-grained UMLS semantic groups. The detailed breakdown per corpus and per finer-grained UMLS semantic types is provided in Appendix C.

UMLS semantic group	**Scientific literature**	**Encyclopedic Web portals**	**Online discussion forums**
Activities & Behaviors	4.46%	3.27%	5.13%
Anatomy	4.31%	3.42%	4.89%
Chemicals & Drugs	13.19%	22.81%	7.54%
Concepts & Ideas	47.26%	38.92%	44.62%
Devices	0.77%	0.61%	0.93%
Disorders	5.69%	12.67%	11.44%
Genes & Molecular Sequences	3.46%	0.59%	1.88%
Geographic Areas	0.59%	0.29%	0.51%
Living Beings	5.01%	5.07%	4.71%
Objects	2.70%	1.94%	3.72%
Occupations	0.42%	0.34%	0.20%
Organizations	0.42%	0.53%	0.39%
Phenomena	1.28%	0.89%	0.96%
Physiology	4.81%	3.59%	9.21%
Procedures	5.62%	5.06%	3.89%

Table 6.3: Distribution of entity types in different genres

Different genres feature different mixes of entity types. Not surprisingly, scientific literature, being oriented towards science at the molecular level, features more genes and molecular sequences than the other two genres. In contrast, encyclopedic Web portals, being oriented towards disease and drug information for consumers and laymen, feature more of those entities instead.

We note that, according to our method, online discussion forums feature genes and molecular sequences 3 times as often as encyclopedic Web portals. This result exposes a limitation of our method, namely that it assumes all textual content is of a biomedical nature. Text mentions that are both common English words as well as gene names are often incorrectly identified as gene entities, since our method disregards the possibility that non-biomedical content may be present.

Prominence of *Concepts & Ideas*. In each of the three genres, the UMLS semantic group *Concepts & Ideas* is by far the most prevalent. This phenomenon can be attributed to the UMLS semantic types belonging to this group, especially the three types *Qualitative Concept*, *Quantitative Concept*, and *Temporal Concept*. Qualitative Concept includes expressions such as *high risk*, *minor*, *severe*, and *stabilized*. Quantitative Concept includes numbers, measurement units such as *mm* and *count/minute*, as well as items that can be measured such *dose*, *score*, and *volume*. Finally, Temporal Concept include expressions such as *late stage*, *yearly*, and even *often*. Since these expressions occur very frequently, they make up a large part of the overall entity type distribution.

6.5 Summary

We present a large-scale, high-quality, and automatic method that disambiguates entities in MEDLINE abstracts by exploiting MeSH terms as well as applying heuristics based on linguistic cues and knowledge assets in UMLS. The method first identifies unambiguous text mentions as anchors, and then disambiguates the remaining text mentions by further leveraging the anchors as additional cues. Not only is the method one order of magnitude faster than MetaMap, the overall accuracy is also slightly superior to that of MetaMap. These improvements position our method as a viable alternative for processing PubMed-scale corpus and for tasks that require real-time responses.

Chapter 7

Corpus-driven Entity Discovery and Disambiguation

Named entity disambiguation (NED) is the task of mapping ambiguous text mentions to entities in a knowledge base (KB). Out-of-KB (OOKB) entities are either disregarded, or become description-less *NIL* placeholders. We address this limitation of prior works by devising a corpus-driven approach. By computing latent embeddings with Latent Dirichlet Allocation (LDA) or word2vec, our method first discovers the latent topics an ambiguous name expresses in a corpus; each topic then describes an entity, whether it exists in the KB or not. This way, an ambiguous mention is mapped to the best fitting entity in a KB, or to a latent entity with an embedding-based description. We demonstrate the viability of our corpus-driven NED method by experiments in the biomedical domain, where micro-averaged accuracy reaches 81.6%. We further demonstrate that our method can generalize to the politics domain with a micro-averaged accuracy of 77.1%.

7.1 Introduction

7.1.1 Motivation

Named entity disambiguation (NED) is the task of taking an ambiguous mention in a text document, and selecting the correct underlying entity from multiple candidates. To date, NED has been predominantly driven by knowledge bases (KB's). Some KB is first hailed as ground truth, the candidates are chosen from this KB, and then the ambiguous mention is mapped to one of the candidates. This approach works well due to two implicit assumptions. First, the KB has good coverage of entities, especially those that are widely known and occur frequently. Second, the vast majority of ambiguous text mentions refer to such entities. Therefore the overall disambiguation outcome is satisfactory, since most of the time the correct entity is present as one of the candidates. In the general domain, a popular choice of KB is Wikipedia, especially the English version thereof, where each Wikipedia article is taken as an entity. As an encyclopedia collaboratively curated by the public, it coalesces knowledge from numerous individuals to reach expansive coverage. In the biomedical domain, UMLS (Unified Medical Language System) is the default KB since it is the largest and most authoritative metathesaurus built by expert contributors.

Two limitations immediately arise concerning out-of-KB (OOKB) entities, which is when a text mention does not have a corresponding entity in the KB. First, no KB can be complete. New or emerging entities are constantly being created but they will be incorporated into the KB only after a time lapse. As an example, the hashtag *#brexit* was first tweeted in 2012 but the Wikipedia page for *Brexit* was established in October 2015. As another example, the Bourbon virus was discovered in 2014 but was introduced into UMLS as entity C4005701 in 2015.

The second limitation is that state-of-the-art NED methods handle OOKB cases by mapping them to a placeholder *NIL*. This is typically based on a threshold for the confidence that a text mention corresponds to any entity in the KB. Such a *NIL* label is description-less; there is no further information about the underlying entity. In addition, when one entity name refers to multiple OOKB entities, these entities are lumped together to share the same *NIL* label, losing any clue that they could be further distinguished.

Thus far we have considered the NED problem from a KB-driven perspective. While KB's are repositories that systematically catalog entities (among other knowledge), we observe that a corpus of unstructured text is also a repository, only that the entities are obscured in a disorganized manner. The crucial difference between these two kinds of repositories is that a corpus also contains OOKB entities.

We therefore devise a new approach by making NED *corpus-driven* to overcome the aforementioned limitations. Given an ambiguous entity name, we first observe its mentions and their contexts in a corpus, and then determine how many underlying entities are in play, and finally map the mentions to entities in the KB or to OOKB entities with descriptions derived from their corpus contexts. The goal of this work is to improve the quality of NED while also making the output more informative – higher recall for OOKB entities with descriptions, while maintaining high precision for KB entities.

The main challenge of a corpus-driven approach lies in the discovery of entities. We do not know in advance how many entities are represented by the same entity name in the corpus. To meet this challenge, we consider applying dimensionality reduction techniques to the corpus. These techniques condense the entire corpus into a vector space model with a small, predefined number of dimensions, where each dimension can be viewed as embodying a latent topic. Each latent topic can then be further viewed as an entity in latent form.

While there has been prior works that follow this line of approach, they stop short at bridging the gap between the latent topic and KB entities. That is the next challenge – to marry up latent entities with KB entities, keeping in mind to cater for the case when the latent entity is OOKB. This way, each OOKB entity has its own latent description, thus overcoming the limitations regarding the *NIL* label.

Finally, when mapping latent entities to KB entities, there is one more party involved beyond these entities: the text snippet that contains the ambiguous text mention, i.e. the context. How to leverage this context is a further issue to be investigated.

In this work, we restrict our scope to studying ambiguous entity names that have a few different meanings, for example, 5 to 10 different entities in the KB or OOKB. We do not consider cases where a name can be mapped to hundreds of different entities (e.g., a common person name like *Smith*). For this reason, we choose to work with domain-specific texts.

7.1.2 Contribution

We devise an NED method that uses latent topic models as the crux, and context enrichment as a further aid. Specifically, to discover entities in a corpus, regardless of whether they are in-KB or OOKB, we harness methods for computing latent topic models, i.e., embeddings. We use Latent Dirichlet Allocation (LDA) [14] and skip-gram word embeddings, popularly known as word2vec [124]. As opposed to typical LDA or word2vec models, in our case the number of dimensions (i.e., latent topics) is low and should reflect the number of different entities with the same surface name (i.e., the degree of ambiguity of a mention). Our method computes a vector signature for each ambiguous name and candidate entity. We further enrich the signatures of each name with contexts of highly similar mentions. These are precomputations on a given corpus. The output is a set of <name, entity> pairs where some of the entities correspond to KB entities and others are OOKB entities along with embedding-based descriptions. Later, when we are presented with a new (i.e., previously unseen) text document, our method applies the precomputed signatures to compute similarity scores between a mention and the candidate entities. The highest similarity score determines the disambiguated entity, either in the KB or a latent OOKB entity.

We present experimental results that demonstrate the viability of our corpus-driven NED method with a biomedical corpus. Our method achieves micro-averaged accuracy of 81.6%. To the best of our knowledge, this work is the first in the biomedical domain that applies a corpus-driven approach in NED as well as the first that provides latent descriptions to OOKB entities.

We also demonstrate that our method is applicable beyond the biomedical domain. Using a news corpus focused on political organizations and politicians, the same method achieves micro-averaged accuracy 77.1%, with significant improvements over a baseline that does not consider the underlying corpora.

7.2 Related Work

Biomedical Entity Disambiguation

Related works addressing biomedical entity disambiguation has already been discussed at length in the previous Chapter (see Section 6.2). Here we only summarize that the large body of existing works divides into two main flavors (sub-domain-agnostic and sub-domain-specific), and that a wide range of methodologies have been proposed.

Corpus-driven Approaches for Dictionary Construction

Although we are not aware of other corpus-driven entity disambiguation methods in the biomedical domain, there are several works that do take a corpus-driven approach for constructing a dictionary. These works aim at deriving a complete collection of all the domain- or sub-domain-specific terms expressed in the corpus. However, further enrichment of the collected terms, such as reconciliation between terms and KB entities, or categorization of the terms, are out of scope in these works.

The C-value / NC-value method [47] is a seminal work for deriving multi-word terms from a corpus. It first uses part-of-speech tags to filter for noun phrases in the corpus, and then leverages relative frequencies of these noun phrases and their n-gram substrings to determine which n-grams are terms. The output is a list of terms ranked in their C-values, which indicate likelihoods of being a term.

In other works, Li et al. [107] propose the i-SWB topic model. It is a generative model inspired by LDA, and it captures three kinds of topics, namely background, general, and document-specific topics. Besides applying the model to molecular biology, other domains studied include electrical engineering and metallurgy. Two further works [38, 131] focus on texts from the consumer health sub-domain. Both works derive equivalent pairs of specialist and layman terms, and the latter work additionally extracts definitions of the specialist terms.

Distributed Word Representations and Topic Models

Various models of a distributional nature are applicable in a number of biomedical text mining tasks. Prior to the rise of word embeddings, Cohen and Widdows [30] provide an overview of the effectiveness of Latent Dirichlet Allocation (LDA) and other topic models in capturing semantic relatedness. Later, with the advent of word embeddings [124], Muneed et al. [130] demonstrate that word2vec (the skip-gram variant of word embedding models) is a better model for capturing biomedical concepts than than GloVe [145], which is another vector space model that captures word contexts. Two works [25, 143] further consider word2vec as the state-of-the-art model, and investigate the quality of models derived from different corpora and corpus sizes via both intrinsic and extrinsic evaluations.

As word embeddings gain popularity, a number of methods use them as a building block for various biomedical-entity-centric tasks. For entity recognition, Segura-Bedmar et al. [175] enhance standard features of a Conditional Random Field (CRF) model with embeddings for recognizing drug names. For word sense disambiguation (WSD), Tulkens et al. [211] incorporate embeddings in an adaptation of the 2-MRD (second order Machine Readable Dictionary) approach. Sabbir et al. [169] derive embeddings for UMLS entities using the entities' definitions and related entities; their approach is the closest to our construction of candidate signatures. For disambiguating abbreviations in clinical text, Liu et al. [114] incorporate embeddings in a ranking algorithm for prioritizing the abbreviation expansions. Wu et al. [229] take a more conventional approach and incorporate embeddings in an Support Vector Machine (SVM) classifier.

As for using LDA as the modeling approach, two domain-oriented works, though outside of the biomedical domain, leverage LDA to perform WSD. Boyd-Graber et al. [17] propose to extend LDA with the WordNet hierarchy to build the LDAWN model. Their algorithm performs probabilistic posterior inference to determine simultaneously word senses and their respective domains. Preiss and Stevenson [149] also propose to extend LDA with WordNet, but this time with WordNet senses. Their model captures not only word-topic distributions but also sense-topic distributions. This model is domain-specific, and has been studied for the finance and sports domains.

Sense Discovery and Entity Discovery

To be able to discover new word senses or new entities, one must first have a collection of known senses and entities. Moreover, one must be able to compare the new item against the existing collection in order to claim that a discovery has indeed been made. Seen in this light, the new items are often described as *emergent*.

In the general domain, Lau et al. [98] propose an LDA-based method to detect emerging senses. Their method first learns the number of topics before building either a regular or a hierarchical LDA model. In order to identify the emerging senses, an older corpus and a newer one (in terms of publication timestamp) are compared against the LDA model, such that the new senses can be picked up via the new corpus.

Also in the general domain, state-of-the-art general-purpose NED methods such as AIDA [70], Spotlight [123], TagMe [46], and Wikifier [156] are all KB-driven. In other words, the collection of entities come from some KB, and that emerging entities are OOKB entities. The case of OOKB entities is handled by setting a threshold for the confidence for when a given mention corresponds to any of the existing entities in the KB (see, e.g., [68]). When the confidence is below the threshold, NED methods merely declare mentions as unknown entities, using a generic placeholder *NIL* and without giving any further cues about the different entities. A recent work on named entity recognition (NER) and NED for emerging entities [69] has proposed to put the human KB curator in the loop, with interactive support from the system by showing candidate contexts to the curator. In a multi-lingual setting, the Entity Discovery and Linking task at the TAC-KBP (Text Analysis Conference Knowledge Base Population) competition requires participating methods to not only identify *NIL*'s, but also cluster them as a step towards fully disambiguating OOKB entities.

Context Enrichment

Improving disambiguation results by exploiting the context of similar documents is proposed by Li et al. [108, 109], whose method incrementally enriches the context with the most similar documents before proceeding to less similar ones. On the other hand, Wang et al. [217] perform context enrichment in two ways. With a fixed number of instances in their LDA-inspired model, word embeddings are incorporated to inject more context. Alternatively, by incorporating external corpora into the same model, the number of instances can be expanded so as to enrich contexts captured by the overall model.

7.3 Methodology

Recall that we assume an ambiguous entity name (e.g. *Brexit*) expresses only a few entities (the referendum in the United Kingdom to decide whether to withdraw from the European Union, the ensuing withdrawal, and the movie). Under this assumption, the method operates on a per-entity-name basis in two stages. In the preparation stage, entities are discovered from a background corpus by means of latent topic models. Signature vectors of ambiguous text mentions as well as candidates are derived from the resulting model. In the application stage, the similarities of these signatures are compared to select the best entity candidate (in the KB or OOKB).

7.3.1 Preparation Stage

Compiling background corpus. This stage takes as input a set of ambiguous entity names and a large collection of text documents. From this collection, we retrieve all text snippets within a window of s sentences surrounding the same entity name. Here, a dictionary of entity names is used to locate their occurrences by string matching. The resulting snippets are the *background corpus*, and there is one background corpus for each entity name. We compute a latent topic model for each background corpus using either LDA or word2vec.

LDA. Latent Dirichlet Allocation (LDA) is a generative probabilistic model independently proposed by Blei et al. [14] and Pritchard et al. [150]. When applied to natural language processing, it is a designed to model the documents, or multiple series of words, in a text corpus. Intuitively, LDA posits that there are K latent topics, and each topic has a probability distribution of generating words. A document, in turn, is a mixture of topics with different probabilities. Using a combination of these probabilities, the words of the document are generated, thus forming the document. Using the actual words in the corpus as observed outcomes, LDA computes a model that generates those outcomes.

Formally, the generative process with smoothing is as follows.

First, for each topic:

1. Choose $\phi_k \sim \mathrm{Dir}(\beta)$, where $k \in \{1, \ldots, K\}$ for the distribution of words

Then, for each of the M documents in the corpus:

2. Choose $N \sim \mathrm{Poisson}(\xi)$ for the number of words in the document
3. Choose $\theta \sim \mathrm{Dir}(\alpha)$ for the distribution of topics
4. For each word in the document:
   (a) Choose a topic $z_k \sim \mathrm{Multinomial}(\theta)$
   (b) Choose a word $w_n \sim \mathrm{Multinomial}(\phi_{z_k})$

Using plate notation, the above generative process is depicted in Figure 7.1.

[14] details the derivations in full. We note here especially the two hyperparameters, α and β, which are parameters for the two Dirichlet distributions for per-document topics and per-topic words, respectively. We note also that the number of topics K is a number predetermined in some other manner; the model itself does not attempt to derive one K or the best K. In other words, in order to apply LDA, it is our responsibility to come up with a judicious combination of α, β, and K.

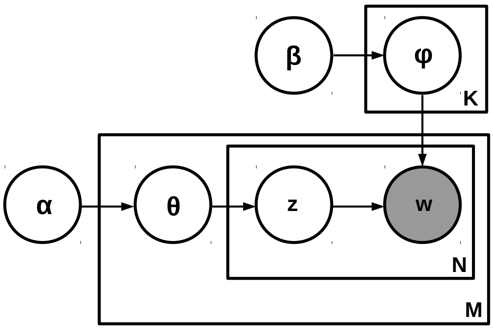

Figure 7.1: Plate notation of smoothed LDA

Constructing latent topic models with LDA. Under the assumption that one entity name can represent at most a few entities, we aim to construct models with low numbers of latent topics. Therefore, the goal is to find an LDA model with K topics such that $2<K<10$ and the model has the best discerning power. This is achieved by picking the model with K topics and the lowest perplexity while maximizing the increase in perplexity in analogous models with K–1 and K+1 topics. The hyperparameters α and β are determined by preliminary experiments.

word2vec. The continuous skip-gram model [124], also known as word2vec, captures the context of individual words. Intuitively, given one word in the middle of a sentence, the model predicts its context, namely the words before and after it. A corpus of sentences is required to construct the model, so that the word sequences in the corpus can be leveraged to form such predictions. Under this model, similar words cluster close together in the resulting latent topic space.

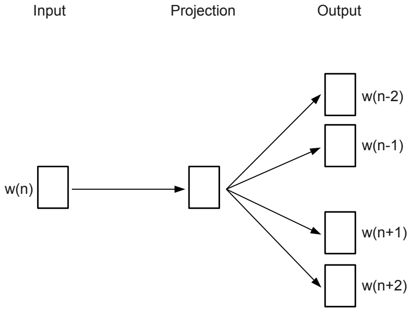

Figure 7.2: Model architecture of word2vec

Formally, word2vec is model using two kinds of classifiers, and Figure 7.2 depicts its architecture. The current word w_n in a sentence is the input to a log-linear classifier, which projects not to the current word position n, but to a predefined window of words before and after n. During model construction, K log-linear classifiers are employed in the projection layer, and softmax classifiers are employed in the output layer. At the end of model construction, the weight matrix recorded in the log-linear classifiers encodes the latent topic space.

Both the original creators of word2vec [124, 125] and other researchers [52] have presented details of the model and its parameters in full. Here, we note that K is a predetermined number; similar to LDA, the word2vec model itself does not attempt to derive one K or the best K. It is again our responsibility to come up with a judicious K.

Constructing latent topic models with word2vec. For word2vec, the notion of inherent perplexity as in LDA does not apply. Therefore we set the number of topics K to 5, 10, 15, or 20. We consider all four word2vec models (for each ambiguous entity name), and later report on the one that gives the best NED performance in experiments. Since multiple latent topics may map to the same entity later, having as many as 20 topics does not violate the assumption that there are only a few entities for the same name.

Constructing latent entity signatures. Regardless of the dimensionality reduction algorithm used, the result is a W-by-K matrix, where W is the number of unique words in the corpus. The columns of matrix are the signature vectors of the K latent entities that are associated with the same ambiguous name.

Linking latent entities to KB entities. Entities in the KB have textual descriptions, which we cast into the same latent topic space as the latent entities. This yields signature vectors for K latent entities and, say, K' in-KB entities where $K' \leq K$ (assuming our choice of K was reasonable).

We now try to link each of the latent entities to one of the in-KB entities, using the cosine similarity between signature vectors. For latent entity j, we choose the in-KB entity e with the highest similarity if it is above a specified threshold. If none of the in-KB entities has a similarity above the threshold, we keep j as a latent OOKB entity, but with a latent-model description and clearly distinguishable from other OOKB entities with the same ambiguous name. The output of this step is a partial linking of latent entities to in-KB entities.

Constructing mention signatures. Given a text mention in a previously unseen document, we fold it into the latent topic space. We again take a snippet of text within a window of s sentences surrounding a text mention. Suppose a sentence in this snippet is *Elderly voters decry shocking Brexit poll results*. Intuitively, we want to encapsulate this sentence's orientation in the latent topic space by combining the information embodied in all the words in the sentence.

Formally, we construct a *base signature* by taking all the words' corresponding vectors in the latent topic space, and averaging them. Although averaging vectors is a simplistic approach, it has been shown to be competitive against other more complicated alternatives [74].

This signature can be further enriched or *contextualized* by incorporating a weighted version of signatures from other highly related snippets containing the same mention.

Specifically, we consider the most related snippets from the background corpus, and set their relatedness to their word overlap. The final signature of a mention is:

$$\mathrm{sig}_{\mathrm{men}} = a_0 \times \mathrm{sig}_{\mathrm{base}} + a_1 \times \mathrm{sig}_1 + a_2 \times \mathrm{sig}_2 + \ldots$$

where sig_i's are signatures of the other snippets from the most related to least related, $\sum a_i = 1$, and $a_0{>}a_1{>}a_2 \ldots$, implicitly constraining $a_0{>}0$ to ensure that the base signature is always present and always the dominant contributor. The values for the a_i parameters are tuned based on a withheld development set. We use up to 10 related snippets under the condition that each snippet must be reasonably related to the base snippet; this is achieved by discarding "related" snippets whose word overlap with the base snippet is below a threshold.

Constructing candidate signatures. Each entity candidate has a description consisting of words (from the KB or from the latent topic model). In addition, each in-KB candidate has semantic relations with further entities in the KB. By adding the names of these related entities to the description, we construct an informative pseudo-document for each entity. This document can then be transformed into a *candidate signature* the same way as the text mention signature (without contextualization). In our Brexit example, three candidate signatures are computed for the referendum, the withdrawal, and the movie.

7.3.2 Application Stage

Mapping mentions to candidates. When presented with a new document containing an ambiguous name, our method compares the signature of the ambiguous mention to each candidate signature and computes the following similarity score:

$$\mathrm{similarity} = (1 - b) \times cos(\mathrm{sig}_{\mathrm{men}}, \mathrm{sig}_{\mathrm{cand}}) + b \times \mathrm{prior}_{\mathrm{cand}}$$

where $cos()$ is the cosine similarity function, $\mathrm{prior}_{\mathrm{cand}}$ is the candidate's prior probability according to AIDA [70], and $0 \leq b < 1$. The candidate with the highest similarity score becomes the disambiguated winner.

7.4 Evaluation

7.4.1 Dataset and Parameter Tuning

Dataset. Table 7.1 summarizes the dataset that we used for experimental evaluation.

For the background corpora, we extracted text snippets from a wide range of genres from scientific literature to encyclopedic Web pages to online discussion threads. The background corpora have on average ca. 31,000 snippets per ambiguous name. A sentence window size s of 5 was used.

As knowledge base, we tap into UMLS, the authoritative KB of the biomedical domain, as it provides both candidate descriptions as well as entity-entity relations.

Background corpus	Collected from PubMed scientific literature, encyclopedic Web portals, online discussion forums
Knowledge base	UMLS
Test dataset	MSH WSD Distribution of 203 ambiguous entity names: 106 abbreviations, 88 terms, 9 mixture of both

Table 7.1: Summary of biomedical dataset

For the test dataset, we use an established gold standard, the MSH WSD dataset [83] which features a mixture of abbreviations and terms. The average number of candidates per ambiguous entity name is 2.08.

Disambiguation priors. Since entity prior information is not available in the biomedical domain, we set the weight b to zero in the mention-vs.-candidate signature similarity comparison; in other words, no disambiguation priors were applied.

LDA software and parameter tuning. We use arbylon[1], a java implementation of the LDA algorithm. The specific method used to perform inference is Gibbs sampling; the implementation's accompanying technical report [65] contains the full derivation. We set aside a randomly chosen 20% subset of the test dataset for parameter tuning. Recalling that since each ambiguous entity name requires its own latent topic model, the parameter combinations of α, β, K are per-entity-name.

word2vec software and parameter tuning. We use the java implementation of word2vec provided by deeplearning4j[2]. Since [25] shows that quality of the resulting model is higher when the sentences in the background corpus are shuffled randomly, we took care to do the same. Recalling that we have, for each ambiguous entity name, 4 word2vec models with $K = 5$, 10, 15, and 20, we report the best choice of the dimensionality K, as determined post-hoc by the experimental evaluation.

We used preliminary experiments to determine other model parameters, and found that the best combination does not vary from one ambiguous entity name to another. Therefore we reached the following combination: negative sampling is always enabled; window size and minimum word frequency are both set to 10.

Parameter tuning for thresholds. Our method uses two thresholds, namely the threshold that determines whether a latent entity is OOKB, and the threshold that determines whether a related text snippet is similar enough for the contextualization of text mention signature. Both thresholds were tuned using the same 20% subset of the test dataset, similarly on a per-entity-name basis.

[1] `www.arbylon.net`

[2] `deeplearning4j.org/word2vec`

7.4.2 Disambiguation of In-KB Entities

Table 7.2 shows the micro-averaged accuracy (on the 80% subset of the test dataset, not used in parameter tuning). The best performing setting is the one based on word2vec with contextualization, reaching 81.6% accuracy. These results are very encouraging, especially since the biomedical domain is notoriously difficult in entity disambiguation. To the best of our knowledge, there is no software tool that provides genuine disambiguation for comparison as baseline here.

	Abbrev.	**Terms**	**Both**	**Overall**
LDA	0.742	0.672	0.739	0.711
LDA + context	0.748	0.674	0.741	0.714
word2vec	0.869	0.722	0.885	0.805
word2vec + context	**0.882**	**0.730**	**0.905**	**0.816**

Table 7.2: Micro-averaged accuracy results for biomedical dataset

Impact of dimensionality reduction technique. We observe that using word2vec as the dimensionality reduction technique consistently and significantly outperforms using LDA. We believe this performance boost comes from the additional information of word order captured by word2vec models; in contrast, LDA is a bag-of-words model that disregards word order.

Impact of context enrichment. We observe that context enrichment also consistently improves accuracy, though not by as big a margin as using word2vec over LDA. This result confirms the intuition that more context means more information, which in turn means more discriminative power. On average, 3.60 similar contexts were used in constructing text mention signatures.

Impact of ambiguity type. Ambiguous entity names in the test dataset come in three types: abbreviations, terms that have their words completely spelled out, and names that exhibit both aspects (e.g. *STEM* is an abbreviation for science, technology, engineering, and mathematics, while *stem* is an anatomical part of a plant). Entities sharing the same abbreviation are generally starkly different; for instance, *PVC* stands for both polyvinyl chloride the common plastic, and for premature ventricular contractions the medical condition. On the other hand, terms tend to reflect multiple shades of the same underlying entity; for instance, *Pneumocystis* refers to both the genus of yeast, as well as the specific form of pneumonia the yeast causes. In other words, the differences in context between abbreviation candidates are larger than those between term candidates. This observation explains the significantly higher accuracy in abbreviations than in terms.

Sample outcomes. Table 7.3 shows some sample outcomes of correctly disambiguated ambiguous entity names using our method. Although in this Table we show only the sentences that contain the text mentions, we note that each sentence is surrounded by the rest of the text snippet bearing more context.

Abbreviation: EMS

Entity	**Sample disambiguated sentences**
Emergency medical service	Of EMS, by EMF, for EMS. Let's stop the abuse & improve relations between field and in-hospital providers.
Ethyl methanesulfonate	In vivo genotoxicity of EMS: statistical assessment of the dose response curves.

Term: Hybridization

Entity	**Sample disambiguated sentences**
Crossbreeding of species	We also tested the hypothesis that hybridization may contribute to PA diversity within plants, by comparing PA expression in parental species to that in artificially generated F(1) hybrids, and also in later generation natural hybrids between S. jacobaea and S. aquaticus.
Hybridization of nucleic acids	Samples were collected from infected leaves before treatment, 7 and 15 days after treatment for DNA and molecular hybridization analysis.

Both abbreviation and term: Ice

Entity	**Sample disambiguated sentences**
Interleukin-1 converting enzyme	LPS increases the expression levels of IL-18, ICE and IL-18 R in mouse testes.
Water in solid state	Molecular cloning and expression analysis of a cytosolic Hsp70 gene from Antarctic ice algae Chlamydomonas sp.
Methamphetamine	Smokable ("ice", "crystal meth") and non smokable amphetamine-type stimulants: clinical pharmacological and epidemiological issues, with special reference to the UK.

Table 7.3: Sample outcomes of ambiguous biomedical entity names

7.5 Applying the Methodology in the Politics Domain

Although the method described in this Chapter is designed for domain-specific usage, it is not designed only for the biomedical domain. In order to demonstrate the applicability of the method to other domains, in this Section we present an evaluation of the method against the politics domain. We begin with a brief discussion of relevant existing works.

7.5.1 Related Work

To the best of our knowledge, there is no existing work that addresses NED specifically in the politics domain. We can, however, consider the news domain. News cover current events, many of which are of a political nature, and so the news domain can be said to encompass the politics domain. For NED in the news domain, however, there are still few previous works. Fernández et al. [45] and Redondo García et al. [163] both propose ranking-based approaches that leverage meta-data that come with news. The method in [45] leverages new articles' temporal tags to detect trends such as a certain sports event, so that news articles deemed to share the same trend can provide extra contextual coherence to each other. On the other hand, the method in [163] leverages Google relevance scores and even news video's subtitles to provide extra cues for the NED task.

7.5.2 Methodology

As the methodology is not domain-specific, the same procedures for the biomedical domain were applied to data in the politics domain.

7.5.3 Evaluation

Dataset and parameter tuning Table 7.4 summarizes the dataset that we used for experimental evaluation.

Background corpus	Collected from gigaword5
Knowledge base	Freebase
Test dataset	Custom constructed from Wikipedia Distribution of 100 ambiguous entity names: 25 abbreviations, 25 organizations, 25 persons, 25 places

Table 7.4: Summary of political dataset

For the background corpora, we extracted text snippets from gigaword5 [144]. It is a collection of English news articles published between 1994 and 2010 by 7 news outlets from 4 countries. A sentence window size s of 5 was used.

As knowledge base, we use Freebase [16] so as to leverage its entity-entity relations.

For the test dataset, there is no established gold standard for the political domain to the best of our knowledge. We therefore constructed a custom test dataset from Wikipedia's Politics Portal. From a Wikipedia disambiguation page, persons and organizations of a political nature, geographical places, as well as their abbreviations are selected as candidates. Hyperlinks in a content page are taken as annotated ambiguous mentions, where the link destinations are the disambiguated entities. The opening section of a candidate's Wikipedia article is used as its description. Since the background corpora covered news only up to 2010, we took care to select only candidates that were established entities by that year. The full list of ambiguous entity names and their candidates is in Appendix D.

The average number of candidates per ambiguous entity name is 5.32. The background corpora have on average ca. 17,000 snippets per ambiguous name; the number of snippets is much lower than that for biomedical domain (ca. 31,000) since gigaword5 is smaller than the biomedical text collection. Disambiguation priors from AIDA [70] were applied.

We performed parameter tuning analogous to that for the biomedical domain. Here, we note a key difference from the biomedical domain: In word2vec, the minimum word frequency is 5 (instead of 10) to adjust for the smaller background corpora.

Disambiguation of in-KB entities. Table 7.5 shows the micro-averaged accuracy (on the 80% subset of the test dataset, not used in parameter tuning). Similar to the biomedical domain, the best performing setting is the one based on word2vec with contextualization, reaching 77.1% accuracy.

	Abbrev.	**Org.**	**Persons**	**Places**	**Overall**
LDA	0.614	0.668	0.603	0.699	0.648
LDA + context	0.620	0.676	0.623	0.701	0.655
word2vec	0.762	0.702	0.660	0.794	0.742
word2vec + context	**0.788**	**0.735**	**0.696**	**0.818**	**0.771**
AIDA	0.619	0.648	0.679	0.695	0.656

Table 7.5: Micro-averaged accuracy results for political dataset

For the political domain, achieving an accuracy of 77.1% is quite good given that the average number of candidates is 5.32. For instance, the candidates for *Bush* include both ex-presidents of USA as well as Jeb Bush. Under this difficult setting, the best setting of our method achieved an accuracy of 75%.

The two trends we observed in the biomedical domain are repeated in the political domain. Specifically, results derived from word2vec models outperform those from LDA models, and using context enrichment boosts performance. These repeated observations confirm that additional information, captured as word order in word2vec models and as extra contexts, translates to additional cues for the NED task.

As a baseline for the political domain, we used a state-of-the-art NED tool out of the box, namely AIDA [70] which uses YAGO as its KB. This comparison is a bit unfair, as AIDA is not customized to the domain and its KB contains a large number of entity candidates that are irrelevant for the dataset. The observation that our method drastically outperforms AIDA shows that corpus-aware NED has inherent advantages in such domain-specific situations.

Table 7.6 show some sample outcomes of correctly disambiguated ambiguous entity names using our method. Although in this Table we show only the sentences that contain the text mentions, we note that each sentence is surrounded by the rest of the text snippet bearing more context.

Abbreviation: DNC

Entity	**Sample disambiguated sentences**
Democratic National Committee	Labor Party founded the National Democratic Policy Committee (NDPC), a political action committee whose name drew complaints from the DNC, who saw these efforts as infiltration.
Democratic National Convention	He returned to his law practice and in 1868 served as a delegate to the DNC.

Organization: Democratic Progressive Party

Entity	**Sample disambiguated sentences**
Political party in Malawi	Chimombo is a member of the DPP and a former member of the UDF.
Political party in Taiwan	The first national election to be held after Chen Shui-bian's victory in the 2000 presidential election, the election resulted for the first time in the Kuomintang (KMT) losing its majority and President Chen's DPP emerging as the largest party in the legislature.

Person: Gallup

Entity	**Sample disambiguated sentences**
David Gallup	Mr. Gallup, delegate to the Republican National Convention 1860, Connecticut State Senator 1869, Lieutenant Governor of Connecticut 1879-1881.
George Gallup	Ogilvy cites Gallup as one of the major influences on his thinking, emphasizing meticulous research methods and adherence to reality.

Place: Bombay

Entity	**Sample disambiguated sentences**
City in India	The cave temples of Elephanta Island (near Mumbai or Bombay, as it was known formerly), Ajanta, and Ellora (in Maharashtra), and structural temples of Pattadakal, Aihole, Badami in Karnataka and Mahaballipuram and Kanchipuram in Tamil Nadu are enduring legacies of otherwise warring regional rulers.
State in India	The Deccan States Agency, also known as the Deccan States Agency and Kolhapur Residency, was a political agency of British India, managing the relations of the British government of the Bombay Presidency with a collection of princely states.
Bombay in New York State in USA	The Mohawk Tribe views the reservation as a "sovereign nation," but shares jurisdiction with the State of New York, the United States, and the Town of Bombay, in which it is located.

Table 7.6: Sample outcomes of ambiguous political entity names

7.6 Summary

We present a corpus-driven approach to named entity disambiguation, as an alternative to the knowledge-base-driven (KB-driven) mainstream line of prior works. Using dimensionality reduction techniques such as LDA and word2vec, we model a background corpus in a low-dimensional topic space, such that each topic becomes a latent entity. The key advantage of our method is that it handles the case of out-of-KB (OOKB) entities in a more informative way, by distinguishing different OOKB entities and by providing them with entity descriptions derived from latent embeddings. Experiments using ambiguous entity names from the biomedical and political domains demonstrate that our method is an alternative to mainstream, KB-based NED approaches. To the best of our knowledge, out method is the first work to address OOKB entities in the biomedical domain by transforming them into latent representations.

Chapter 8

Conclusion

Entity recognition and disambiguation (ERD) for the biomedical domain are difficult research problems due to many and diverse challenges. The variety of sub-domains and their nomenclature, the mixture of proper names and compound noun phrases, the heterogeneous text genres, only to name the main challenges – all contribute to the complexity and difficulty. In this thesis, we devise solutions that address some aspects of the overall problems.

8.1 Contributions

The first contribution of this thesis is a fast dictionary lookup method for entity recognition. Our method balances the trade-off between small losses in precision and coverage in exchange for huge gains in throughput. The second contribution is a semantic type classification method for common nouns in long noun phrases. By incorporating the generic, non-informative type and non-biomedical types, our method distinguishes crucial words with biomedical meanings as a precursor towards better information extraction. The third contribution is a fast entity disambiguation method applicable across all entity types in MEDLINE abstracts. The method leverages expert-assigned indexing MeSH (Medical Subject Headings) terms, UMLS (Unified Medical Language System) knowledge, and fast heuristics to produce high quality results at high throughput. The fourth contribution is a corpus-driven entity disambiguation method that addresses out-of-knowledge base (OOKB) entities. The method first captures entities in a corpus as latent representations before mapping text mentions to entities either in-knowledge base or OOKB.

8.2 Outlook

Despite the contributions of this thesis, there are still further aspects of the biomedical ERD problem that remain to be tackled.

Entities with attributes. The methods presented in this thesis all leverage UMLS as the underlying knowledge base. However, when only 13% of UMLS entities appear in MEDLINE abstracts [91], we must question how much of that rich knowledge – entities and their accompanying information – our and other UMLS-based methods have truly harnessed. Viewing from a different angle, the entities not found in

MEDLINE are long, composite entity names with multiple attributes (for example, *acquired deformity of left forearm excluding fingers*). These two perspectives are two sides of the same coin, in that entities with highly specific attributes are in demand by the nature of the profession. How to bring these perspectives together and forge a solution capable of handling the myriad entities and their myriad details will be a daunting task.

Text genres such social media and patents. To date, biomedical text mining, and hence biomedical ERD, have focused heavily on scientific literature. In this thesis, we make some inroads into Web content such as encyclopedic health portals and patient discussion forums. Dissemination of information, however, has evolved beyond these relatively static and formal forms of communication. Microblogs such as tweets offer a continuous stream of texts, but they are short and therefore lacking in context, not to mention the heavy use of layman terms cluttered in non-biomedical content. Instead of having most of the text focusing on biomedicine and a small portion of the text on out-of-domain content, the reverse distribution will be the norm. How to winnow out the irrelevant content, separate the cases when the same term is used in a biomedical way and otherwise, and compensate for the lack of context will be a major theme.

Patents are yet another neglected text genre. They are especially important to pharmaceutical companies, where combing through existing patents to identify conflicts and opportunities are very much part of the drug development process. Here, besides the need to adapt to the genre's specific language style, high recall will be a key requirement.

Conjunctions in noun phrases. Conjunctions are a common linguistic phenomenon not addressed in this thesis. Specifically, resolving the scope of conjunction reveals the entity implied by the writer. Consider, for example, *non-activated Factors II, IX, and X*, and one sees immediately with high confidence that *IX* and X are also non-activated Factors. But in another example, *side effects from chlorothiazide and methyldopa therapy*, without prior knowledge one cannot tell if chlorothiazide is a drug, or if chlorothiazide therapy is a standard procedure and hence the intended entity, or even if both cases apply. Despite being a fundamental problem, to the best of our knowledge there has been no prior study on conjunctions in biomedical text.

One text mention, multiple entities. Both entity disambiguation methods proposed in this thesis offer black-and-white answers: A text mention is mapped to exactly this entity or that entity. However, natural language is inherently imprecise, leading to different readings of the same text by different readers. Using the same chlorothiazide example, indeed both readings (being a drug, and being a specific drug therapy) are valid. Consider also a text mention as simple as *children*, for which there are very fine-grained UMLS entities referring to the age group and the familial function. In such cases, it can be beneficial to allow multiple answers with different levels of confidence or with different reasons for the decision. There is no question that the "one text mention, multiple entities" phenomenon exists; the question is how to incorporate multiple answers in a disambiguation method.

Bibliography

[1] E. Agirre, A. Soroa, and M. Stevenson. Graph-based word sense disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896, 2010.

[2] N. Alnazzawi, P. Thompson, R. Batista-Navarro, and S. Ananiadou. Using text mining techniques to extract phenotypic information from the PhenoCHF corpus. *BMC Medical Informatics and Decision Making*, 15(2):1–10, 2015.

[3] S. Ananiadou, D. Sullivan, W. Black, G.-A. Levow, J. J. Gillespie, C. Mao, S. Pyysalo, B. Kolluru, J. Tsujii, and B. Sobral. Named entity recognition for bacterial type IV secretion systems. *PLoS ONE*, 6(3):e14780, 2011.

[4] C. N. Arighi, A. Siu, C. O. Tudor, J. A. Nchoutmboube, C. H. Wu, and V. K. Shanker. *eFIP: A tool for mining functional impact of phosphorylation from Literature*, in volume 694 of *Methods in Molecular Biology – Bioinformatics for Comparative Proteomics*, pages 63–75, 2011.

[5] A. Aronson. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. In *American Medical Informatics Association Annual Symposium Proceedings (AMIA)*, pages 17–21, 2001.

[6] A. R. Aronson and F.-M. Lang. An overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association (JAMIA)*, 17(3):229–236, 2010.

[7] M. Bada, M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, W. A. Baumgartner, K. B. Cohen, K. Verspoor, J. A. Blake, and L. E. Hunter. Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(1):1–20, 2012.

[8] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. Video suggestion and discovery for YouTube: Taking random walks through the view graph. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 895–904, 2008.

[9] R. T. Batista-Navarro, R. Rak, and S. Ananiadou. Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics. *Journal of Cheminformatics*, 7(S-1):S6, 2015.

[10] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In *Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 491–498, 2008.

[11] D. Benikova, C. Biemann, M. Kisselew, and S. Pado. GermEval 2014 named entity recognition: Companion paper. In *Proceedings of the Konferenz zur Verarbeitung Natürlicher Sprache (KONVENS) GermEval Shared Task on Named Entity Recognition*, pages 104–112, 2014.

[12] L. Bing, S. Chaudhari, R. C. Wang, and W. W. Cohen. Improving distant supervision for information extraction using label propagation through lists. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 524–529, 2015.

[13] L. Bing, M. Ling, R. Wang, and W. Cohen. Distant IE by bootstrapping using lists and document structure. *arXiv preprint arXiv:1601.00620*, 2016.

[14] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[15] O. Bodenreider. The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267, 2004.

[16] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the International Conference on Management of Data (SIGMOD)*, pages 1247–1250, 2008.

[17] J. Boyd-Graber, D. Blei, and X. Zhu. A topic model for word sense disambiguation. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1024–1033, 2007.

[18] L. Boytsov, D. Novak, Y. Malkov, and E. Nyberg. Off the beaten path: Let's replace term-based retrieval with k-NN search. In *Proceedings of the International on Conference on Information and Knowledge Management (CIKM)*, pages 1099–1108, 2016.

[19] A. Z. Broder, M. Charikar, A. M. Frieze, and M. Mitzenmacher. Min-wise independent permutations (extended abstract). In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, pages 327–336, 1998.

[20] D. Campos, S. Matos, and J. Oliveira. Gimli: Open source and high-performance biomedical name recognition. *BMC Bioinformatics*, 14(1):54, 2013.

[21] E. Chaix, B. Dubreucq, A. Fatihi, D. Valsamou, R. Bossy, M. Ba, L. Deléger, P. Zweigenbaum, P. Bessieres, L. Lepiniec, and C. Nédellec. Overview of the regulatory network of plant seed development (SeeDev) task at the BioNLP shared task 2016. In *Proceedings of the Workshop on Biomedical Natural Language Processing Shared Task (BioNLP-ST)*, pages 1–11, 2016.

[22] M. S. Charikar. Similarity estimation techniques from rounding algorithms. In *Proceedings of the Annual ACM Symposium on Theory of Computing (STOC)*, pages 380–388, 2002.

[23] Y. Chen, H. Cao, Q. Mei, K. Zheng, and H. Xu. Applying active learning to supervised word sense disambiguation in MEDLINE. *Journal of the American Medical Informatics Association (JAMIA)*, 20(5):1001–1006, 2013.

[24] W. Cheng, J. Preiss, and M. Stevenson. Scaling up WSD with automatically generated examples. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 231–239, 2012.

[25] B. Chiu, G. Crichton, A. Korhonen, and S. Pyysalo. How to train good word embeddings for biomedical NLP. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 166–174, 2016.

[26] O. Chum, J. Philbin, and A. Zisserman. Near duplicate image detection: min-hash and tf-idf weighting. In *Proceedings of the British Machine Vision Conference (BMVC)*, volume 3, pages 4–13, 2008.

[27] J. J. Cimino. Desiderata for controlled medical vocabularies in the twenty-first century. *Methods of information in medicine*, 37(4-5):394, 1998.

[28] A. Cohen and W. Hersh. A survey of current work in biomedical text mining. *Briefings in bioinformatics*, 6(1):57–71, 2005.

[29] K. B. Cohen, K. Verspoor, K. Fort, C. Funk, M. Bada, M. Palmer, and L. Hunter. The Colorado richly annotated full text (CRAFT) corpus: Multi-model annotation in the biomedical domain. *Handbook of Linguistic Annotation*, 2015.

[30] T. Cohen and D. Widdows. Empirical distributional semantics: Methods and biomedical applications. *Journal of Biomedical Informatics*, 42(2):390–405, 2009.

[31] N. Collier, T. Groza, D. Smedley, P. N. Robinson, A. Oellrich, and D. Rebholz-Schuhmann. PhenoMiner: From text to a database of phenotypes associated with omim diseases. *Database: The Journal of Biological Databases and Curation*, 2015:bav104, 2015.

[32] M. Collins and Y. Singer. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 189–196, 1999.

[33] A. Coulet, N. H. Shah, Y. Garten, M. Musen, and R. B. Altman. Using text to build semantic networks for pharmacogenomics. *Journal of Biomedical Informatics*, 43:1009–1019, 2010.

[34] L. Deléger, R. Bossy, E. Chaix, M. Ba, A. Ferré, P. Bessieres, and C. Nédellec. Overview of the bacteria biotope task at BioNLP Shared Task. In *Proceedings of the Workshop on Biomedical Natural Language Processing Shared Task (BioNLP-ST)*, pages 12–22, 2016.

[35] M. Douyère, L. F. Soualmia, A. Névéol, A. Rogozan, B. Dahamna, J.-P. Leroy, B. Thirion, and S. J. Darmoni. Enhancing the MeSH thesaurus to retrieve french online health resources in a quality-controlled gateway. *Health Information & Libraries Journal*, 21(4):253–261, 2004.

[36] J. D'Souza and V. Ng. Sieve-based entity linking for the biomedical domain. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, volume 2: Short Papers, pages 297–302, 2015.

[37] G. Durrett and D. Klein. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490, 2014.

[38] N. Elhadad and K. Sutaria. Mining a lexicon of technical terms and lay equivalents. In *Proceedings of the Workshop on BioNLP: Biological, Translational, and Clinical Language Processing (BioNLP)*, pages 49–56, 2007.

[39] P. Ernst, C. Meng, A. Siu, and G. Weikum. KnowLife: A knowledge graph for health and life sciences. In *Proceedings of the International Conference on Data Engineering (ICDE)*, pages 1254–1257, 2014.

[40] P. Ernst, A. Siu, D. Milchevski, J. Hoffart, and G. Weikum. DeepLife: An entity-aware search, analytics and exploration platform for health and life sciences. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL) System Demonstrations*, pages 19–24, 2016.

[41] P. Ernst, A. Siu, and G. Weikum. KnowLife: A versatile approach to constructing a knowledge graph for biomedical sciences. *BMC Bioinformatics*, 16:157, 2015.

[42] J.-W. Fan and C. Friedman. Generating quality word sense disambiguation test sets based on MeSH indexing. In *American Medical Informatics Association Annual Symposium Proceedings (AMIA)*, pages 183–187, 2009.

[43] C. Fellbaum. *WordNet: An Electronic Lexical Database.* MIT Press, 1998.

[44] Y. Feng, Z. Han, and K. Zhang. *Overview of the NLPCC 2015 Shared Task: Entity recognition and linking in search queries*, pages 550–556. Springer International Publishing, Cham, 2015.

[45] N. Fernández, J. A. Fisteus, L. Sánchez, and G. López. IdentityRank: Named entity disambiguation in the news domain. *Expert Systems with Applications*, 39(10):9207 – 9221, 2012.

[46] P. Ferragina and U. Scaiella. TAGME: On-the-fly annotation of short text fragments (by Wikipedia entities). In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1625–1628, 2010.

[47] K. Frantzi, S. Ananiadou, and H. Mima. Automatic recognition of multi-word terms: the C-value/NC-value method. *International Journal on Digital Libraries*, 3(2):115–130, 2000.

[48] C. Funk, W. Baumgartner, B. Garcia, C. Roeder, M. Bada, K. Cohen, L. Hunter, and K. Verspoor. Large-scale biomedical concept recognition: An evaluation of current automatic annotators and their parameters. *BMC Bioinformatics*, 15(1):59, 2014.

[49] W. Gatens, B. Konev, and F. Wolter. Lower and upper approximations for depleting modules of description logic ontologies. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 345–350, 2014.

[50] S. Gaudan, A. Jimeno-Yepes, V. Lee, and D. Rebholz-Schuhmann. Combining evidence, specificity, and proximity towards the normalization of gene ontology terms in text. *EURASIP Journal on Bioinformatics and Systems Biology*, 2008(1):342746, 2008.

[51] M. Gerner, G. Nenadic, and C. M. Bergman. LINNAEUS: A species name identification system for biomedical literature. *BMC bioinformatics*, 11(1):1, 2010.

[52] Y. Goldberg and O. Levy. word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[53] W. Golik, R. Bossy, Z. Ratkovic, and C. Nédellec. Improving term extraction with linguistic analysis in the biomedical domain. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing), Special Issue of the Journal Research in Computing Science*, pages 24–30, 2013.

[54] G. H. Gonzalez, T. Tahsin, B. C. Goodale, A. C. Greene, and C. S. Greene. Recent advances and emerging applications in text and data mining for biomedical discovery. *Briefings in Bioinformatics*, 17(1):33, 2016.

[55] N. Grabar, T. Hamon, and O. Bodenreider. Ontologies and terminologies: Continuum or dichotomy? *Applied ontology*, 7(4):375–386, 2012.

[56] T. Grego, P. Pęzik, F. M. Couto, and D. Rebholz-Schuhmann. Identification of chemical entities in patent documents. In *Proceedings of the International Work-Conference on Artificial Neural Networks (IWANN)*, Part II, pages 942–949, 2009.

[57] A. Grycner, P. Ernst, A. Siu, and G. Weikum. Knowledge discovery on incompatibility of medical concepts. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 114–125, 2013.

[58] N. Guarino and P. Giaretta. Ontologies and knowledge bases towards a terminological clarification. *Towards Very Large Knowledge Bases: Knowledge Building & Knowledge Sharing*, 25:32, 1995.

[59] M. Habibi, D. L. Wiegandt, F. Schmedding, and U. Leser. Performance of gene name recognition tools on patents. In *Proceedings of the International Symposium on Semantic Mining in Biomedicine (SMBM)*, pages 3–10, 2016.

[60] M. Habibi, D. L. Wiegandt, F. Schmedding, and U. Leser. Recognizing chemicals in patents: S comparative analysis. *Journal of Cheminformatics*, 8(1):59, 2016.

[61] M. A. Haendel, F. Neuhaus, D. Osumi-Sutherland, P. M. Mabee, J. L. Mejino, C. J. Mungall, and B. Smith. CARO – The Common Anatomy Reference Ontology. In *Anatomy Ontologies for Bioinformatics*, volume 6 of *Computational Biology (A. Burger, D. Davidson, and R. Baldock, editors)*, pages 327–349, 2008.

[62] K. Hakala, S. Kaewphan, T. Salakoski, and F. Ginter. Syntactic analyses and named entity recognition for PubMed and PubMed central – up-to-the-minute. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 102–107, 2016.

[63] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(suppl 1):D514, 2005.

[64] N. Harmston, W. Filsell, and M. Stumpf. Which species is it? Species-driven gene name disambiguation using random walks over a mixture of adjacency matrices. *Bioinformatics*, 28(2):254–260, 2012.

[65] G. Heinrich. Parameter estimation for text analysis. Technical report, Fraunhofer IGD, 2009.

[66] K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. M. Hendriksen, B. J. A. Schijvenaars, E. M. v. Mulligen, J. Kleinjans, and J. A. Kors. A dictionary to identify small molecules and drugs in free text. *Bioinformatics*, 25(22):2983–2991, 2009.

[67] L. Hirschman, M. Colosimo, A. Morgan, and A. Yeh. Overview of BioCreAtIvE task 1B: Normalized gene lists. *BMC Bioinformatics*, 6(1):S11, 2005.

[68] J. Hoffart, Y. Altun, and G. Weikum. Discovering emerging entities with ambiguous names. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 385–396, 2014.

[69] J. Hoffart, D. Milchevski, G. Weikum, A. Anand, and J. Singh. The knowledge awakens: Keeping knowledge bases fresh with emerging entities. In *Proceedings of the International Conference on World Wide Web (WWW)*, volume: Companion, pages 203–206, 2016.

[70] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 782–792, 2011.

[71] A. Holzinger, J. Schantl, M. Schroettner, C. Seifert, and K. Verspoor. Interactive knowledge discovery and data mining in biomedical informatics: State-of-the-art and future challenges. *Biomedical Text Mining: State-of-the-Art, Open Problems and Future Challenges*, pages 271–300, 2014.

[72] C.-C. Huang and Z. Lu. Community challenges in biomedical text mining over 10 years: Success, failure and the future. *Briefings in Bioinformatics*, 17(1):132–144, 2016.

[73] S. M. Humphrey, W. J. Rogers, H. Kilicoglu, D. Demner-Fushman, and T. C. Rindflesch. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology*, 57(1):96–113, 2006.

[74] I. Iacobacci, M. Taher Pilehvar, and R. Navigli. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, volume 1, pages 897–907, 2016.

[75] R. Islamaj Doğan and Z. Lu. An inference method for disease name normalization. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Fall Symposium Series*, pages 8–13, 2012.

[76] M. Ivanović and Z. Budimac. An overview of ontologies and data resources in medical domains. *Expert Systems with Applications*, 41(11):5158 – 5166, 2014.

[77] A. Jimeno-Yepes. Higher order features and recurrent neural networks based on long-short term memory nodes in supervised biomedical word sense disambiguation. *arXiv preprint arXiv:1604.02506*, 2016.

[78] A. Jimeno-Yepes and A. R. Aronson. Knowledge-based biomedical word sense disambiguation: Comparison of approaches. *BMC Bioinformatics*, 11(1):1–12, 2010.

[79] A. Jimeno-Yepes and A. R. Aronson. Knowledge-based and knowledge-lean methods combined in unsupervised word sense disambiguation. In *Proceedings of the ACM SIGHIT International Health Informatics Symposium (IHI)*, pages 733–736, 2012.

[80] A. Jimeno-Yepes, E. Jimenez-Ruiz, V. Lee, S. Gaudan, R. Berlanga, and D. Rebholz-Schuhmann. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(3):1–10, 2008.

[81] A. Jimeno-Yepes, R. B. Llavori, and M. P. Catalán. Disambiguating automatically-generated semantic annotations for life science open registries.

In *Proceedings of the International Workshop on Exploiting Large Knowledge Repositories (E-LKR)*, 2012.

[82] A. Jimeno-Yepes, A. MacKinlay, B. Han, and Q. Chen. Identifying diseases, drugs, and symptoms in Twitter. *Studies in health technology and informatics*, 216:643—647, 2015.

[83] A. Jimeno-Yepes, B. McInnes, and A. R. Aronson. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. *BMC Bioinformatics*, 12:223, 2011.

[84] A. Jimeno-Yepes, B. T. Mclnnes, and A. R. Aronson. Collocation analysis for UMLS knowledge-based word sense disambiguation. *BMC Bioinformatics*, 12(3):1–12, 2011.

[85] A. Jimeno Yepes, A. Névéol, M. Neves, K. Verspoor, O. Bojar, A. Boyer, C. Grozea, B. Haddow, M. Kittner, Y. Lichtblau, P. Pecina, R. Roller, R. Rosa, A. Siu, P. Thomas, and S. Trescher. Findings of the WMT 2017 biomedical translation shared task. In *Proceedings of the Conference on Machine Translation (WMT)*, volume 2: Shared Task Papers, pages 234–247, 2017.

[86] Y. Jin, R. McDonald, K. Lerman, M. Mandel, S. Carroll, M. Liberman, F. Pereira, R. Winters, and P. White. Automated recognition of malignancy mentions in biomedical literature. *BMC Bioinformatics*, 7(1):492, 2006.

[87] S. Kaewphan, S. Van Landeghem, T. Ohta, Y. Van de Peer, F. Ginter, and S. Pyysalo. Cell line name recognition in support of the identification of synthetic lethality in cancer from text. *Bioinformatics*, 32(2):276–282, 2016.

[88] N. Kang, B. Singh, Z. Afzal, E. M. van Mulligen, and J. A. Kors. Using rule-based natural language processing to improve disease normalization in biomedical text. *Journal of the American Medical Informatics Association (JAMIA)*, 20(5):876-881, 2012.

[89] N. Kang, B. Singh, C. Bui, Z. Afzal, E. van Mulligen, and J. Kors. Knowledge-based extraction of adverse drug events from biomedical text. *BMC Bioinformatics*, 15(1):64, 2014.

[90] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. GENIA corpus – a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(Suppl 1):i180–i182, 2003.

[91] S. Kim, Z. Lu, and W. J. Wilbur. Identifying named entities from PubMed for enriching semantic categories. *BMC Bioinformatics*, 16(1):1–10, 2015.

[92] S. Kim and J. Yoon. Link-topic model for biomedical abbreviation disambiguation. *Journal of Biomedical Informatics*, 53:367–380, 2015.

[93] S. Kocbek and T. Groza. Building a dictionary of lexical variants for phenotype descriptors. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 186–190, 2016.

[94] S. Köhler, N. A. Vasilevsky, M. Engelstad, E. Foster, J. McMurry, S. Aymé, G. Baynam, S. M. Bello, C. F. Boerkoel, K. M. Boycott, et al. The Human Phenotype Ontology in 2017. *Nucleic Acids Research*, 45(D1):D865–D876, 2016.

[95] B. Kolluru, L. Hawizy, P. Murray-Rust, J. Tsujii, and S. Ananiadou. Using workflows to explore and optimise named entity recognition for chemistry. *PLoS ONE*, 6(5):e20181, 05 2011.

[96] I. Korkontzelos, D. Piliouras, A. W. Dowsey, and S. Ananiadou. Boosting drug named entity recognition using an aggregate classifier. *Artificial Intelligence in Medicine*, 65(2):145 – 153, 2015.

[97] M. Kreuzthaler and S. Schulz. Disambiguation of period characters in clinical narratives. In *Proceedings of the International Workshop on Health Text Mining and Information Analysis (Louhi) at the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 96–100, 2014.

[98] J. H. Lau, P. Cook, D. McCarthy, D. Newman, and T. Baldwin. Word sense induction for novel sense detection. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 591–601, 2012.

[99] C. Leacock and M. Chodorow. Combining local context and WordNet similarity for word sense identification. *WordNet: An Electronic Lexical Database*, 49(2):265–283, 1998.

[100] R. Leaman and G. Gonzalez. BANNER: An executable survey of advances in biomedical named entity recognition. In *Proceedings of Pacific Symposium on Biocomputing*, pages 652–663, 2008.

[101] R. Leaman, R. Islamaj Doğan, and Z. Lu. DNorm: Disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909, 2013.

[102] R. Leaman, R. Khare, and Z. Lu. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics*, 57:28–37, 2015.

[103] R. Leaman, C.-H. Wei, and Z. Lu. tmChem: A high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*, 7(supplement 1), 2015.

[104] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. DBpedia – A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, 6(2):167–195, 2012.

[105] C. Li, M. Liakata, and D. Rebholz-Schuhmann. Biological network extraction from scientific literature: State of the art and challenges. *Briefings in Bioinformatics*, 15(5):856–877, 2013.

[106] D. Li, K. Kipper-Schuler, and G. Savova. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing (BioNLP)*, pages 94–95, 2008.

[107] S. Li, J. Li, T. Song, W. Li, and B. Chang. A novel topic model for automatic term extraction. In *Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 885–888, 2013.

[108] Y. Li, S. Tan, H. Sun, J. Han, D. Roth, and X. Yan. Entity disambiguation with linkless knowledge bases. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 1261–1270, 2016.

[109] Y. Li, C. Wang, F. Han, J. Han, D. Roth, and X. Yan. Mining evidences for named entity disambiguation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1070–1078, 2013.

[110] N. Limsopatham and N. Collier. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, pages 1014–1023, 2016.

[111] S.-d. Lin and K. Verspoor. A semantics-enhanced language model for unsupervised word sense disambiguation. In *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 287–298, 2008.

[112] X. Ling, S. Singh, and D. Weld. Design challenges for entity linking. *Transactions of the Association for Computational Linguistics (TACL)*, 3:315–328, 2015.

[113] X. Ling and D. S. Weld. Fine-grained entity recognition. In *Proceedings of Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence*, 2012.

[114] Y. Liu, T. Ge, K. Mathews, H. Ji, and D. McGuinness. Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 92–97, 2015.

[115] Z. Lu, H.-Y. Kao, C.-H. Wei, M. Huang, J. Liu, C.-J. Kuo, C.-N. Hsu, R. T.-H. Tsai, H.-J. Dai, N. Okazaki, H.-C. Cho, M. Gerner, I. Solt, S. Agarwal, F. Liu, D. Vishnyakova, P. Ruch, M. Romacker, F. Rinaldi, S. Bhattacharya, P. Srinivasan, H. Liu, M. Torii, S. Matos, D. Campos, K. Verspoor, K. M. Livingston, and W. J. Wilbur. The gene normalization task in BioCreative III. *BMC Bioinformatics*, 12(8):S2, 2011.

[116] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014.

[117] B. T. McInnes. An unsupervised vector approach to biomedical term disambiguation: Integrating umls and medline. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) on Human Language Technologies: Student Research Workshop (HLT-SRWS)*, pages 49–54, 2008.

[118] B. T. McInnes, Y. Liu, T. Pedersen, G. B. Melton, and S. V. Pakhomov. Umls::similarity: Measuring the relatedness and similarity of biomedical concepts. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technology (HLT)*, pages 28–31, 2013.

[119] B. T. McInnes and T. Pedersen. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of Biomedical Informatics*, 46(6):1116–1124, 2013.

[120] B. T. McInnes and T. Pedersen. Evaluating semantic similarity and relatedness over the semantic grouping of clinical term pairs. *Journal of Biomedical Informatics*, 54:329–336, 2015.

[121] B. T. McInnes, T. Pedersen, Y. Liu, G. B. Melton, and S. V. Pakhomov. Knowledge-based method for determining the meaning of ambiguous biomedical terms using information content measures of similarity. In *American Medical Informatics Association Annual Symposium Proceedings (AMIA)*, pages 895–904, 2011.

[122] B. T. McInnes and M. Stevenson. Determining the difficulty of word sense disambiguation. *Journal of Biomedical Informatics*, 47:83–90, 2013.

[123] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia Spotlight: Shedding light on the Web of documents. In *Proceedings of the International Conference on Semantic Systems (I-SEMANTICS)*, pages 1–8, 2011.

[124] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[125] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[126] A. Morgan, Z. Lu, X. Wang, A. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, C. Sun, H.-H. Liu, R. Torres, M. Krauthammer, W. Lau, H. Liu, C.-N. Hsu, M. Schuemie, K. B. Cohen, and L. Hirschman. Overview of BioCreative II gene normalization. *Genome Biology*, 9(Suppl 2):S3, 2008.

[127] A. Moro, A. Raganato, and R. Navigli. Entity linking meets word sense disambiguation: A unified approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2, 2014.

[128] D. L. Mowery, B. R. South, L. Christensen, L.-M. Murtola, S. Salanterä, H. Suominen, D. Martinez, N. Elhadad, S. Pradhan, G. Savova, and W. W. Chapman. Task 2: ShAre/CLEF eHealth evaluation lab. In *Proceedings of Conference and Labs of the Evaluation Forum (CLEF) Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis*, 2013.

[129] Y. Mrabet, C. Gardent, M. Foulonneau, E. Simperl, and E. Ras. Towards knowledge-driven annotation. In *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence*, pages 2425–2431, 2015.

[130] T. Muneeb, S. K. Sahu, and A. Anand. Evaluating distributed word representations for capturing semantics of biomedical concepts. *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing (ACL/IJCNLP)*, pages 158–163, 2015.

[131] S. Muresan and J. L. Klavans. Inducing terminologies from text: A case study for the consumer health domain. *Journal of the American Society for Information Science and Technology*, 64(4):727–744, 2013.

[132] N. Naderi, T. Kappler, C. J. O. Baker, and R. Witte. OrganismTagger: Detection, normalization and grounding of organism entities in biomedical documents. *Bioinformatics*, 27(19):2721–2729, 2011.

[133] N. Nakashole, T. Tylenda, and G. Weikum. Fine-grained semantic typing of emerging entities. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, pages 1488–1497, 2013.

[134] P. I. Nakov and M. A. Hearst. Semantic interpretation of noun compounds using verbal and other paraphrases. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(3):13:1–13:51, 2013.

[135] M. Narayanaswamy, K. Ravikumar, and K. Vijay-Shanker. A biological named entity recognizer. In *Proceedings of Pacific Symposium on Biocomputing*, pages 427–438, 2003.

[136] R. Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(2):10:1–10:69, 2009.

[137] V. Nebot, M. Ye, M. Albrecht, J.-H. Eom, and G. Weikum. DIDO: a disease-determinants ontology from web sources. In *Proceedings of the International Conference on World Wide Web (WWW)*, pages 237–240, 2011.

[138] A. Névéol, W. Kim, W. Wilbur, and Z. Lu. Exploring two biomedical text genres for disease recognition. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 144–152, 2009.

[139] M. Neves and U. Leser. A survey on annotation tools for the biomedical literature. *Briefings in Bioinformatics*, 15(2):327, 2012.

[140] D. Nguyen, M. Theobald, and G. Weikum. J-NERD: Joint named entity recognition and disambiguation with rich linguistic features. *Transactions of the Association for Computational Linguistics (TACL)*, 4:215–229, 2016.

[141] N. Okazaki, S. Ananiadou, and J. Tsujii. Building a high-quality sense inventory for improved abbreviation disambiguation. *Bioinformatics*, 26(9):1246–1253, 2010.

[142] D. Osumi-Sutherland, S. J. Marygold, G. H. Millburn, P. A. McQuilton, L. Ponting, R. Stefancsik, K. Falls, N. H. Brown, and G. V. Gkoutos. The drosophila phenotype ontology. *Journal of Biomedical Semantics*, 4(1):30, 2013.

[143] S. V. Pakhomov, G. Finley, R. McEwan, Y. Wang, and G. B. Melton. Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics*, 32(23):3635, 2016.

[144] R. Parker, D. Graff, J. Kong, K. Chen, and K. Maeda. English gigaword fifth edition, June 2011. LDC2011T07.

[145] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[146] S. Pradhan, N. Elhadad, W. Chapman, S. Manandhar, and G. Savova. SemEval-2014 task 7: Analysis of clinical text. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval)*, pages 54–62, 2014.

[147] S. Pradhan, N. Elhadad, B. R. South, D. Martinez, A. Vogel, H. Suominen, W. W. Chapman, and G. Savova. Task 1: ShARe/CLEF eHealth evaluation lab. In *Proceedings of Conference and Labs of the Evaluation Forum (CLEF) Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis*, 2013.

[148] J. Preiss and M. Stevenson. DALE: A word sense disambiguation system for biomedical documents trained using automatically labeled examples. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technology (HLT)*, pages 1–4, 2013.

[149] J. Preiss and M. Stevenson. Unsupervised domain tuning to improve word sense disambiguation. In *Proceedings of the Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL): Human Language Technology (HLT)*, pages 680–684, 2013.

[150] J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

[151] S. Pyysalo and S. Ananiadou. Anatomical entity mention recognition at literature scale. *Bioinformatics*, 30(6):868, 2013.

[152] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E: Statistical, Nonlinear, Biological, and Soft Matter Physics*, 76:036106, 2007.

[153] A. Rahman and V. Ng. Inducing fine-grained semantic classes via hierarchical and collective classification. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 931–939, 2010.

[154] B. Rance, E. Doughty, D. Demner-Fushman, M. G. Kann, and O. Bodenreider. A mutation-centric approach to identifying pharmacogenomic relations in text. *Journal of Biomedical Informatics*, 45(5):835–841, 2012.

[155] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 147–155, 2009.

[156] L. Ratinov, D. Roth, D. Downey, and M. Anderson. Local and global algorithms for disambiguation to Wikipedia. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL): Human Language Technologies (HLT)*, volume 1, pages 1375–1384, 2011.

[157] D. Ravichandran, P. Pantel, and E. Hovy. Randomized algorithms and NLP: Using locality sensitive hash function for high speed noun clustering. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, pages 622–629, 2005.

[158] D. Rebholz-Schuhmann, S. Clematide, F. Rinaldi, S. Kafkas, E. M. van Mulligen, C. Bui, J. Hellrich, I. Lewin, D. Milward, M. Poprat, A. Jimeno-Yepes, U. Hahn, and J. A. Kors. Entity recognition in parallel multi-lingual biomedical corpora: The CLEF-ER laboratory overview. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*, volume 8138 of *Lecture Notes in Computer Science (P. Former et al., editors)*, pages 353–367, 2013.

[159] D. Rebholz-Schuhmann, A. J. Jimeno-Yepes, E. M. van Mulligen, N. Kang, J. A. Kors, D. Milward, P. T. Corbett, E. Buyko, K. Tomanek, E. Beisswanger, and U. Hahn. The CALBC silver standard corpus for biomedical named entities – a study in harmonizing the contributions from four independent named entity taggers. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 568–573, 2010.

[160] D. Rebholz-Schuhmann, S. Kafkas, J.-H. Kim, A. Jimeno-Yepes, and I. Lewin. Monitoring named entity recognition: The league table. *J. Biomedical Semantics*, 4:19, 2013.

[161] D. Rebholz-Schuhmann, S. Kafkas, J.-H. Kim, C. Li, A. Jimeno-Yepes, R. Hoehndorf, R. Backofen, and I. Lewin. Evaluating gold standard corpora against gene/protein tagging solutions and lexical resources. *J. Biomedical Semantics*, 4:28, 2013.

[162] D. Rebholz-Schuhmann, A. J. Yepes, C. Li, S. Kafkas, I. Lewin, N. Kang, P. Corbett, D. Milward, E. Buyko, E. Beisswanger, K. Hornbostel, A. Kouznetsov, R. Witte, J. B. Laurila, C. J. Baker, C.-J. Kuo, S. Clematide, F. Rinaldi, R. Farkas, G. Móra, K. Hara, L. I. Furlong, M. Rautschka, M. L. Neves, A. Pascual-Montano, Q. Wei, N. Collier, M. F. M. Chowdhury, A. Lavelli, R. Berlanga, R. Morante, V. Van Asch, W. Daelemans, J. L. Marina, E. van Mulligen, J. Kors, and U. Hahn. Assessment of NER solutions against the first and second CALBC silver standard corpus. *Journal of Biomedical Semantics*, 2(5):S11, 2011.

[163] J. L. Redondo García, G. Rizzo, and R. Troncy. The concentric nature of news semantic snapshots: Knowledge extraction for semantic annotation of news items. In *Proceedings of the International Conference on Knowledge Capture (K-CAP)*, pages 16:1–16:8, 2015.

[164] X. Ren, A. El-Kishky, C. Wang, F. Tao, C. R. Voss, H. Ji, and J. Han. ClusType: Effective entity recognition and typing by relation phrase-based clustering. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 995–1004, 2015.

[165] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1524–1534, 2011.

[166] T. Rocktäschel, M. Weidlich, and U. Leser. ChemSpot: A hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640, 2012.

[167] R. Rodriguez-Esteban and M. Bundschus. Text mining patents for biomedical knowledge. *Drug Discovery Today*, 21(6):997 – 1002, 2016.

[168] C. Rosse and J. L. Mejino Jr. A reference ontology for biomedical informatics: The Foundational Model of Anatomy. *Journal of Biomedical Informatics*, 36(6):478–500, 2003.

[169] A. Sabbir, A. J. Yepes, and R. Kavuluru. Knowledge-based biomedical word sense disambiguation with neural concept embeddings and distant supervision. *arXiv preprint arXiv:1610.08557*, 2016.

[170] I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. Multiword expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, volume 2276 of *Lecture Notes in Computer Science (A. Gelbukh, editor)*, pages 1–15, 2002.

[171] S. K. Sahu and A. Anand. Recurrent neural network models for disease name recognition using domain invariant features. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, volume 1: Long Papers, pages 2216–2225, 2016.

[172] G. K. Savova, A. R. Coden, I. L. Sominsky, R. Johnson, P. V. Ogren, P. C. de Groen, and C. G. Chute. Word sense disambiguation across two domains: Biomedical literature and clinical notes. *Journal of Biomedical Informatics*, 41(6):1088–1100, 2008.

[173] G. K. Savova, J. J. Masanz, P. V. Ogren, J. Zheng, S. Sohn, K. C. Kipper-Schuler, and C. G. Chute. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association (JAMIA)*, 17(5):507–513, 2010.

[174] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe. Disease Ontology: A backbone for disease semantic integration. *Nucleic Acids Research*, 40(D1):D940–D946, 2012.

[175] I. Segura-Bedmar, V. Suárez-Paniagua, and P. Martínez. Exploring word embedding for drug name recognition. In *Proceedings of the International Workshop on Health Text Mining and Information Analysis*, pages 64–72, 2015.

[176] B. Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA)*, pages 104–107, 2004.

[177] B. Settles. ABNER: An open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192, 2005.

[178] N. H. Shah, N. Bhatia, C. Jonquet, D. Rubin, A. P. Chiang, and M. A. Musen. Comparison of concept recognizers for building the open biomedical annotator. *BMC Bioinformatics*, 10(9):S14, 2009.

[179] G. Sheikhshab, E. Starks, A. Karsan, A. Sarkar, and I. Birol. Graph-based semi-supervised gene mention tagging. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 27–35, 2016.

[180] D. Shen, J. Zhang, G. Zhou, J. Su, and C.-L. Tan. Effective adaptation of a hidden Markov model-based named entity recognizer for biomedical domain. In *Proceedings of the Association for Computational Linguistics (ACL) Workshop on Natural Language Processing in Biomedicine (BioMed)*, volume 13, pages 49–56, 2003.

[181] W. Shen, J. Wang, and J. Han. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460, 2015.

[182] C. Shivade, M.-C. de Marneffe, E. Fosler-Lussier, and A. M. Lai. Identification, characterization, and grounding of gradable terms in clinical text. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 17–26, 2016.

[183] A. Sil and A. Yates. Re-ranking for joint named-entity recognition and linking. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM)*, pages 2369–2374, 2013.

[184] M. S. Simpson and D. Demner-Fushman. Biomedical text mining: A survey of recent progress. In *Mining Text Data (C. C. Aggarwal, C. Zhai, editors)*, pages 465–517, 2002.

[185] N. Sioutos, S. d. Coronado, M. W. Haber, F. W. Hartel, W.-L. Shaiu, and L. W. Wright. NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *Journal of Biomedical Informatics*, 40(1):30–43, 2007.

[186] A. Siu, P. Ernst, and G. Weikum. Disambiguation of entities in MEDLINE abstracts by combining MeSH terms with knowledge. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 72–76, 2016.

[187] A. Siu, D. B. Nguyen, and G. Weikum. Fast entity recognition in biomedical text. In *Proceedings of the Workshop on Data Mining for Healthcare (DMH) at the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2013.

[188] A. Siu and G. Weikum. Semantic type classification of common words in biomedical noun phrases. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 98–103, 2015.

[189] L. Smith, L. K. Tanabe, R. Johnson née Ando, C.-J. Kuo, I.-F. Chung, C.-N. Hsu, Y.-S. Lin, R. Klinger, C. M. Friedrich, K. Ganchev, M. Torii, H. Liu, B. Haddow, C. A. Struble, R. J. Povinelli, A. Vlachos, W. A. Baumgartner, L. Hunter, B. Carpenter, R. T.-H. Tsai, H.-J. Dai, F. Liu, Y. Chen, C. Sun, S. Katrenko, P. Adriaans, C. Blaschke, R. Torres, M. Neves, P. Nakov, A. Divoli, M. Maña-López, J. Mata, and W. J. Wilbur. Overview of BioCreative II gene mention recognition. *Genome Biology*, 9(2):S2, 2008.

[190] P. Sondhi, M. Gupta, C. X. Zhai and J. Hockenmaier. Shallow information extraction from medical forum data. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1158–1166, 2010.

[191] Y. Song, E. Kim, G. G. Lee, and B.-K. Yi. POSBIOTM–NER: A trainable biomedical named-entity recognition system. *Bioinformatics*, 21(11):2794–2796, 2005.

[192] P. Stenetorp, S. Pyysalo, S. Ananiadou, and J. Tsujii. Generalising semantic category disambiguation with large lexical resources for fun and profit. *Journal of Biomedical Semantics*, 5:26, 2014.

[193] M. Stevenson, E. Agirre, and A. Soroa. Exploiting domain information for word sense disambiguation of medical documents. *Journal of the American Medical Informatics Association (JAMIA)*, 19(2):235–240, 2012.

[194] M. Stevenson and Y. Guo. Disambiguation in the biomedical domain: The role of ambiguity type. *Journal of Biomedical Informatics*, 43(6):972–981, 2010.

[195] M. Stevenson and Y. Guo. Disambiguation of ambiguous biomedical terms using examples generated from the UMLS Metathesaurus. *Journal of Biomedical Informatics*, 43(5):762–773, 2010.

[196] M. Stevenson, Y. Guo, and R. Gaizauskas. Acquiring sense tagged examples using relevance feedback. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 809–816, 2008.

[197] M. Stevenson, Y. Guo, R. Gaizauskas, and D. Martinez. Disambiguation of biomedical text using diverse sources of information. *BMC Bioinformatics*, 9(Suppl 11):S7, 2008.

[198] B. Strauss, B. E. Toma, A. Ritter, M.-C. de Marneffe, and W. Xu. Results of the WNUT16 named entity recognition shared task. In *Proceedings of the Workshop on Noisy User-generated Text (WNUT)*, pages 138–144, 2016.

[199] F. Suchanek, G. Kasneci, and G. Weikum. YAGO – A large ontology from Wikipedia and WordNet. *Elsevier Journal of Web Semantics*, 6(3):203–217, 2008.

[200] P. P. Talukdar and K. Crammer. New regularized algorithms for transductive learning. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (EMCL PKDD)*, volume: Part II, pages 442–457, 2009.

[201] M. A. Tanenblatt, A. Coden, and I. L. Sominsky. The ConceptMapper approach to named entity recognition. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 546–551, 2010.

[202] M. Theobald, J. Siddharth, and A. Paepcke. SpotSigs: Robust and efficient near duplicate detection in large web collections. In *Proceedings of the International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pages 563–570, 2008.

[203] P. Thomas, T. Rocktäschel, J. Hakenberg, Y. Lichtblau, and U. Leser. SETH detects and normalizes genetic variants in text. *Bioinformatics*, 32(18):2883, 2016.

[204] P. Thompson, J. McNaught, S. Montemagni, N. Calzolari, R. del Gratta, V. Lee, S. Marchi, M. Monachini, P. Pezik, V. Quochi, C. Rupp, Y. Sasaki, G. Venturi, D. Rebholz-Schuhmann, and S. Ananiadou. The BioLexicon: A large-scale terminological resource for biomedical text mining. *BMC Bioinformatics*, 12(1):397, 2011.

[205] K. S. E. F. Tjong. Introduction to the CoNLL-2002 Shared Task: Language-independent named entity recognition. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, volume 20, pages 1–4, 2002.

[206] K. S. E. F. Tjong and F. De Meulder. Introduction to the CoNLL-2003 Shared Task: Language-independent named entity recognition. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, volume 4, pages 142–147, 2003.

[207] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT)*, volume 1, pages 173–180, 2003.

[208] R. T.-H. Tsai, S.-H. Wu, W.-C. Chou, Y.-C. Lin, D. He, J. Hsiang, T.-Y. Sung, and W.-L. Hsu. Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, 7(1):92, 2006.

[209] E. Tseytlin, K. Mitchell, E. Legowski, J. Corrigan, G. Chavan, and R. S. Jacobson. NOBLE – flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics*, 17(1):1–15, 2016.

[210] Y. Tsuruoka, J. McNaught, J. Tsujii, and S. Ananiadou. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics*, 23(20):2768–2774, 2007.

[211] S. Tulkens, S. Šuster, and W. Daelemans. Using distributed representations to disambiguate biomedical and clinical concepts. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 77-82, 2016.

[212] A. Usié, R. Alves, F. Solsona, M. Vázquez, and A. Valencia. CheNER: chemical named entity recognizer. *Bioinformatics*, 30(7):1039–1040, 2014.

[213] Ö. Uzuner, I. Solti, and E. Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association (JAMIA)*, 17(5):514, 2010.

[214] Ö. Uzuner, B. R. South, S. Shen, and S. L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association (JAMIA)*, 18(5):552–556, 2011.

[215] K. Verspoor, K. B. Cohen, A. Lanfranchi, C. Warner, H. L. Johnson, C. Roeder, J. D. Choi, C. Funk, Y. Malenkiy, M. Eckert, N. Xue, W. A. Baumgartner, M. Bada, M. Palmer, and L. E. Hunter. 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*, 13(1):1–26, 2012.

[216] D. Vrandečić and M. Krötzsch. Wikidata: A free collaborative knowledgebase. *Communications of ACM*, 57(10):78–85, 2014.

[217] J. Wang, M. Bansal, K. Gimpel, B. Ziebart, and C. Yu. A sense-topic model for word sense induction with unsupervised data enrichment. *Transactions of the Association for Computational Linguistics (TACL)*, 3:59–71, 2015.

[218] X. Wang, J. Tsujii, and S. Ananiadou. Disambiguating the species of biomedical named entities using natural language parsers. *Bioinformatics*, 26(5):661–667, 2010.

[219] M. Weeber, J. G. Mork, and A. R. Aronson. Developing a test collection for biomedical word sense disambiguation. In *American Medical Informatics Association Annual Symposium Proceedings (AMIA)*, pages 746–750, 2001.

[220] C.-H. Wei, B. R. Harris, H.-Y. Kao, and Z. Lu. tmVar: A text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, 29(11):1433–1439, 2013.

[221] C.-H. Wei, H.-Y. Kao, and Z. Lu. SR4GN: A species recognition software tool for gene normalization. *PLOS ONE*, 7(6):1–5, 2012.

[222] C.-H. Wei, H.-Y. Kao, and Z. Lu. GNormPlus: An integrative approach for tagging genes, gene families, and protein domains. *BioMed Research International*, 2015:918710, 2015.

[223] C.-H. Wei, R. Leaman, and Z. Lu. SimConcept: A hybrid approach for simplifying composite named entities in biomedical text. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1385–1391, 2015.

[224] C.-H. Wei, R. Leaman, and Z. Lu. Beyond accuracy: Creating interoperable and scalable text-mining Web services. *Bioinformatics*, 2016.

[225] B. Wellner, J. Castaño, and J. Pustejovsky. Adaptive string similarity metrics for biomedical reference resolution. In *Proceedings of the Association for Computational Linguistics and the International Conference on Intelligent Systems for Molecular Biology (ACL-ISMB) Workshop on Linking Biological Literature, Ontologies and Databases: Mining Biological Semantics*, pages 9–16, 2005.

[226] R. W. White, R. Harpaz, N. H. Shah, W. DuMouchel, and E. Horvitz. Toward enhanced pharmacovigilance using patient-generated data on the Internet. *Clinical Pharmacology & Therapeutics*, 96(2):239–246, 2014.

[227] R. Winnenburg, A. Sorbello, A. Ripple, R. Harpaz, J. Tonning, A. Szarfman, H. Francis, and O. Bodenreider. Leveraging MEDLINE indexing for pharmacovigilance – inherent limitations and mitigation strategies. *Journal of Biomedical Informatics*, 57:425–435, 2015.

[228] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Research*, 36(Suppl 1):D901–D906, 2008.

[229] Y. Wu, J. Xu, Y. Zhang, and H. Xu. Clinical abbreviation disambiguation using neural word embeddings. In *Proceedings of the Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 171–176, 2015.

[230] I. Yamada, H. Shindo, H. Takeda, and Y. Takefuji. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 250–259, 2016.

[231] A. Yamaguchi, Y. Yamamoto, J.-D. Kim, T. Takagi, and A. Yonezawa. Discriminative application of string similarity methods to chemical and non-chemical names for biomedical abbreviation clustering. *BMC Genomics*, 13(Suppl 3):S8, 2012.

[232] D. Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, pages 189–196, 1995.

[233] A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman. BioCreAtIvE task 1A: Gene mention finding evaluation. *BMC Bioinformatics*, 6(1):S2, 2005.

[234] M. A. Yosef, S. Bauer, J. Hoffart, M. Spaniol, and G. Weikum. HYENA: Hierarchical type classification for entity names. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 1361–1370, 2012.

[235] S. Zhang and N. Elhadad. Unsupervised biomedical named entity recognition: Experiments with clinical and biological texts. *Journal of Biomedical Informatics*, 46(6):1088 – 1098, 2013.

[236] J. G. Zheng, D. Howsmon, B. Zhang, J. Hahn, D. McGuinness, J. Hendler, and H. Ji. Entity linking for biomedical literature. *BMC Medical Informatics and Decision Making*, 15(1):1–9, 2015.

[237] Z. Zhong and H. T. Ng. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, pages 78–83, 2010.

[238] X. Zhou, X. Zhang, and X. Hu. MaxMatcher: Biological concept extraction using approximate dictionary lookup. In *Proceedings of the Pacific Rim International Conference on Artificial Intelligence (PRICAI)*, pages 1145–1149, 2006.

[239] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMUCALD-02-107, School of Computer Science, Carnegie Mellon University, 2002.

[240] G. Zuccon, A. Holloway, B. Koopman, and A. Nguyen. Identify disorders in health records using conditional random fields and MetaMap: AEHRC at ShARe/CLEF 2013 eHealth evaluation lab task 1. In *Proceedings of Conference*

and Labs of the Evaluation Forum (CLEF) Workshop on Cross-Language Evaluation of Methods, Applications, and Resources for eHealth Document Analysis, 2013.

[241] P. Zweigenbaum, D. Demner-Fushman, H. Yu, and K. Cohen. Frontiers of biomedical text mining: Current progress. *Briefings in bioinformatics*, 8(5):358–375, 2007.

[242] S. Zwicklbauer, C. Seifert, and M. Granitzer. Do we need entity-centric knowledge bases for entity disambiguation? In *Proceedings of the International Conference on Knowledge Management and Knowledge Technologies (i-KNOW)*, pages 4:1–4:8, 2013.

[243] S. Zwicklbauer, C. Seifert, and M. Granitzer. Search-based entity disambiguation with document-centric knowledge bases. In *Proceedings of the International Conference on Knowledge Technologies and Data-driven Business (i-KNOW)*, pages 6:1–6:8, 2015.

List of Figures

List of Tables

Appendix A

List of suffixes removed from entity names in UMLS

(+)
(-)
(+-)
(+/-)
(#___)
[#/volume]
(___cm)
(___ cm)
(___ degree.)
(___ degrees)
(___ min)
(___ mm)
(& level)
(1)
(1:1)
(1-4)
(2)
(3)
(4)
(5)
(6)
(7)
(8)
(9)
(10)
(14)
(2010)
(2011)
(2012)
(a)
(activity)
(allelic variant)
[Ambiguous]
(Approved Lists 1980)
(assessment scale)
(attribute)
(b)
(Base Equivalent)
[BAU]
(biological function)
(body structure)
(Bulk)
(cell)
(cell structure)
[Chemical/Ingredient]
(combined site)
(context-dependent category)
(count/vol)
(CT Scan)
(D)
[D]
(diagnosis)
(diagnostic)
[Disease/Finding]
(discontinued)
(disorder)
(documented clinically or microbiologically)
(Drosophila)
[dup]
(E)
(EA)
(ENTERIC COATED)
(environment)
[EPC]
(etiology)
(event)
[FACIT]
(FC)
(FINAL DOSE FORM)
(finding)
(function)
(geographic location)
(GM)
(GRAM)
(GRAMS)
(H1N1)
(H3N2)
(HARD, SOFT, ETC.)
(Hensel, 1867)
(HDR)
(history)
[hp_C]
[hp_X]
[Identifier]

```
(I)
(II)
(III)
(INHALATION)
(IR)
[iU]
(L)
(L.)
(lab test)
(LDR)
(Linnaeus, 1758)
(M)
[M]
[Mass]
[Mass ratio]
[Mass/mass]
[Mass/time]
[Mass/volume]
(mammal)
(manifestation)
(mechanical)
(medication)
(MeSH Category)
(MILLILITERS)
(ML)
[MoA]
[Molar ratio]
[Moles/volume]
(morphologic abnormality)
(MRI)
(N)
(nail)
(navigational concept)
(non-specific)
nos
NOS
[nos]
[NOS]
(nos)
(NOS)
not otherwise specified
(observable entity)
(Obsolete)
(obsolete)
(occupation)
of unspecified site
(or disorder)
(organism)
[PE]
(person)
(pdr for recon)
[PhenX]
(physical finding)
(physical force)
(physical object)
[PNU]
[Presence]
(procedure)
(product)
(qualifier value)
(R)
[Ratio]
(regime/therapy)
-RETIRED-
(S)
(s)
(S/I)
(S. cerevisiae)
(SDV,MDV OR ADDITIVE)
(situation)
(specimen)
(structure)
(substance)
[Susceptibility]
(symptom)
(systemic)
(T)
(Titer)
[Titer]
[TREATMENT]
(treatment)
(tumor staging)
(unintentional)
[Units/volume]
unspecified
unspecified site
(USP)
[USP'U]
[V]
[VA Drug Interaction]
[VA Product]
[X]
(Z)
```

Appendix B

List of target nouns, custom semantic types, and seed phrases

Target noun	Custom semantic types	Seed phrases
activity	body and protein process	catalytic activity disease activity inflammatory activity kinase activity
	non-informative movement in general	of activity in of activity of
	physical activity	fetal activity physical activity
administration	applying medicine	intravenous administration medication administration topical administration
	bureaucracy	hospital administration safety administration
area	division of abstract entity	area of study priority area problem area research area
	division within a location	perioperative area work area
	geographical location	endemic area geographical area rural area urban area
	geometric surface	area under the curve
	location in body part	frontal area skin area tumor area tumour area
	physical surface	surface area total area

Target noun	**Custom semantic types**	**Seed phrases**
body	anatomical part	mandibular body vitreous body
	part of a cell	apoptotic body ketone body lewy body pineal body polar body
	rhetorical collection	body of research regulatory body
	whole physical being	body odor body odour body weight foreign body human body upper body
case	body part	brain case rib case
	english letters	lower case upper case

	instance of disease in patient	% of cases with
		[[number]] cases of
		case control
		case history
		case management
		case report
		case study
		confirmed case of

	legal matters	case law court case
	non-informative incidence in general	in case of in most cases in the case of
concentration	concentration camp	concentration camp
	density of matter	haemoglobin concentration high concentration of plasma concentration
	mental function	concentration loss mental concentration
	physical contraction	uterine concentration

Target noun	**Custom semantic types**	**Seed phrases**
condition	configuration or setting	chosen condition environmental condition experimental condition living condition under normal condition
	symptom or finding	chronic condition comorbid condition medical condition
control	about experiments	control group
	birth control	birth control
	disease control	disease control
	restriction in general	in the control of
culture	medical sample	blood culture cell culture culture medium tissue culture
	way of life	corporate culture hispanic culture language and culture
degree	academic degree	bachelor degree masters degree
	biomedical stage in progression	first degree relatives second degree burn third degree perineal tear
	degree of freedom	degree of freedom degrees of freedom
	edges out of a node in graph	degree distribution node degree
	metric for bending	[[number]] degree bevel [[number]] degree extension [[number]] degree flexion
	metric for temperature	[[number]] degree celcius [[number]] degree fahrenheit [[number]] degree fever
	non-informative rhetorical description of severity	to the degree of

Target noun	Custom semantic types	Seed phrases
development	creation of a drug	drug development
	disease progression	cancer development development of diabetes disease development
	growth of physical body	cell development delayed development embryonic development muscle development
	non-biomedical progress	community development economic development leadership development professional development software development urban development
	non-informative progression in general	advances in the development of continued development of future development
distribution	spread in general	in the distribution of size distribution tissue distribution
	statistical distribution	as a distribution of binomial distribution chi-square distribution gaussian distribution statistical distribution
effect	biomedical named entity	Bohr effect Haldane effect adverse effect cohort effect doppler effect generation effect late effect placebo effect side effect toxic effect
	non-informative impact in general	possible effect of the the effect of the
expression	communication	expression of feelings facial expression verbal expression
	english idiom	blank expression
	gene expression	gene expression
	manifestation of disease	disease expression

Target noun	**Custom semantic types**	**Seed phrases**
factor	a multiple of some quantity	by a factor of [[number]]
	impact factor	impact factor
	non-informative influence on consequence	factors impacting factors influencing
	related to protein	growth factor transcription factor
	risk factor	risk factor
	statistics	covariate factor multivariate factor univariate factor
flow	english idiom	flow of events
	flow cytometry analysis	flow cytometry analysis
	passing of fluid	blood flow flow rate flow volume fluid flow period flow
	passing of information	data flow information flow
	passing of other matter	flow cell nerve impulse flow
form	biomedical named entity	recessive form wave form
	dosage form	dosage form dose form oral form topical form
	non-informative type in general	a form of different forms of in the form of modified form of
	paper form	fill in the form
function	body function	metabolic function motor function pulmonary function renal function
	mathematical function	as a function of correlation function logistic function mathematical function
	non-informative utility in general	the function of the whose function is to

Target noun	Custom semantic types	Seed phrases
group	biomedical named entity	age group blood group control group ethnic group placebo group
	collection of subjects or cases in experiment	conservative group group [[number]] subjects group a group of [[number]] subjects treated group
	non-informative collection of items	group of drugs group of experts
information	biomedical information	drug information patient information personal information
	biomedical named entity	silent information regulators
	general use	additional information basic information sufficient information
	information technology	information retrievel information system
line	biomedical stage of events	first line treatment second line therapy
	body part boundary	jaw line joint line lip line
	cell line	cell line germ line
	geometric boundary	line in figure line of sight line tracing
	non-informative line in general	along those lines in line with
measure	actions in a strategy	control measure measures against preventive measure safety measures supportive measures
	quantitative measurement	interpretable measure performance measure similarity measure

Target noun	Custom semantic types	Seed phrases
mechanism	non-informative procedure in general	as a mechanism of probable mechanism
	specific biomedical mechanism	biochemical mechanism defense mechanism drug mechanism immune mechanism
model	manufactured objects	device model equipment model model number
	model organism	animal model cell line model mouse model
	simulated structure in experiment	model of chronic model of disease prediction model regression model statistical model
pattern	about biopsy	gleason pattern
	about ecg	ecg pattern
	mental activity	pattern recognition
	non-informative template in general	various patterns of whose pattern is
	physical activity	feeding pattern gait pattern sleep pattern
	specific characterization of biomedical entity	banding pattern binding pattern spatial pattern wavy pattern
period	biomedical duration	gestation period incubation period
	menstruation	menstrual period
	time span in general	[[number]] year period for a period of

Target noun	**Custom semantic types**	**Seed phrases**
point	about data and graphs	data point point in the graph
	abstract biomedical location	isoelectric point melting point
	metaphorical location	entry point starting point
	metric for scores	scored [[number]] points
	moment in time	time point
	opinion	key point point of view
	physical biomedical location	acupuncture point exit point needle point point mutation point of application
	specific mental state	at one point in at some point to the point of to the point where
pressure	non-physical pressure	selection pressure social pressure
	pressure from air	air pressure airway pressure
	pressure from blood	arterial pressure blood pressure
	pressure from body fliud	intracranial pressure pulmonary capillary wedge pressure
	pressure from sound	sound pressure
	pressure from weight	pressure sore
problem	english idiom	no problem teething problem
	non-informative issue in general	big problem same problem
	specific issue or difficulty	economic problem management problem mathematical problem problem resolution
	symptom	health problem medical problem problem behavior problem behaviour problem breathing

Target noun	**Custom semantic types**	**Seed phrases**
process	biomedical process	ageing process aging process binding process healing process thought process
	body part	articular process condyloid process
	information process	data process information process markov process review process
	non-informative procedure in general	during the process of
product	blood product	blood product
	cell function	degradation product end product gene product metabolic product
	food product	food products meat products milk products
	health care product	health care product medicinal product
	mathematics	inner product outer product
	other output	product ion
profile	about proteins and genes	expression profile lipid profile
	biomedical characterization	activity profile adverse effect profile clinical profile pharmacokinetic profile risk profile safety profile
	physical shape	facial profile shape profile

Target noun	Custom semantic types	Seed phrases
program	computer program	computer program computer programme program computing programme computing
	genetic program	gene expression program gene expression programme genetic program genetic programme
	medical treatment	dialysis program dialysis programme exercise program exercise programme intervention program intervention programme relaxation program relaxation programme
	social assistance scheme	government program government programme medicaid program medicaid programme
	training regimen	athletic program athletic programme degree program degree programme education program education programme training program training programme
rate	biomedical named entity	death rate flow rate heart rate
	proportion	mortality rate rate of cancer
	speed of activity	rate of increase
reaction	body function	adverse reaction allergic reaction drug reaction immune reaction
	chemical reaction	chemical reaction
	lab technique	polymerase chain reaction
	mental reaction	emotional reaction reaction to threat
	reaction time	reaction time

Target noun	**Custom semantic types**	**Seed phrases**
reduction	medical procedure	closed reduction open reduction
	opposite of oxidation	oxidation and reduction
	reducing in quantity	quantity reduction size reduction volume reduction
region	about proteins	c-terminal region n-terminal region promoter region
	body part	cubital region frontal region midline region sacral region
	geographical area	appalachian region polluted region of tropical region
	other physical area in general	region in the graph
report	biomedical named entity	assessment report case report laboratory report
	non-informative communication in general	detailed report first report on preliminary report published reports
resistance	chemical or protein repulsion	drug resistance immune resistance insulin resistance
	physical repulsion	airway resistance vascular resistance
	psychological repulsion	resistance to the idea
response	body function	behavioral response behavioural response immune response inflammatory response inhibitory response response to glucose
	emergency response	emergency response
	other reaction in general	response to criticism response to financial
	reaction to treatment	dose response treatment response
	response time	response time

Target noun	**Custom semantic types**	**Seed phrases**
result	english idiom	as a result
	findings from experiment or research	functional result laboratory result negative result test result
	non-informative outcome	promising results previous results unpublished results
	treatment outcome	surgical result treatment result
role	biomedical function	catalytic role functional role gender role regulatory role
	non-informative rhetorical use	critical role role of the
	role model	role model
sequence	about proteins	chromosome sequence protein sequence sequence alignment
	series in general	sequence of events
set	division of data	data set test set training set validation set
	non-informative collection in general	[[number]] set of is a set of large set of
site	anatomical location	body site distal site injection site site of injury tumor site tumour site
	general physical location	construction site remote site
	location in protein	binding site phosphorylation site splice site target site transcription start site
	website	internet site web site

Target noun	**Custom semantic types**	**Seed phrases**
solution	chemical in liquid form	aqueous solution dialysis solution nacl solution ophthalmic solution oral rehydration solution powder for solution saline solution
	solution to problem	best solution possible solution
state	geographical entity	county and state state or local united states
	non-informative status in general	future state of in a state of
	specific biomedical status	disease state functional state mental state steady state
	state of the art	state of the art
strain	body part strain	eye strain shoulder strain
	mental strain	mental strain psychological strain
	organism strain	virus strain
system	about information technology	automated system billing system
	collection of body parts	cardiac system immune system
	organized collection in general	a system of scoring system system performance
	systems biology	systems biology
technique	non-informative know-how in general	[[number]] different techniques a novel technique for other techniques the technique of
	specific technique	diffusion technique imaging technique

Appendix C

Distribution of entity types in different corpora

UMLS semantic group: Activities & Behaviors

	Scientific literature		**Encyclopedic Web portals**						**Online discussion forums**		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs .com**	**Mayo Clinic**	**Medline-Plus**	**RxList**	**UpTo-Date**	**Wikipedia Health**	**eHealth-Forum**	**Health-Boards**	**patient .co.uk**
Activity	2.53%	2.64%	1.91%	2.19%	1.19%	1.98%	1.80%	1.80%	1.87%	2.00%	1.96%
Behavior	0.04%	0.03%	0.03%	0.10%	0.16%	0.02%	0.03%	0.05%	0.13%	0.08%	0.08%
Daily or Recreational Activity	0.22%	0.20%	0.19%	0.43%	0.23%	0.13%	0.22%	0.37%	0.77%	0.98%	1.07%
Event	0.09%	0.12%	0.05%	0.02%	0.01%	0.14%	0.08%	0.11%	0.06%	0.06%	0.05%
Governmental or Regulatory Activity	0.17%	0.15%	0.14%	0.02%	0.35%	0.27%	0.08%	0.12%	0.05%	0.06%	0.04%
Individual Behavior	0.35%	0.24%	0.15%	0.26%	0.11%	0.19%	0.16%	0.25%	0.48%	0.39%	0.37%
Machine Activity	0.03%	0.03%	0.00%	0.00%	0.00%	0.01%	0.01%	0.02%	0.03%	0.03%	0.02%
Occupational Activity	0.46%	0.39%	0.53%	0.30%	0.22%	0.57%	0.45%	0.52%	0.58%	0.83%	1.11%
Social Behavior	0.72%	0.65%	0.19%	0.35%	0.18%	0.14%	0.36%	0.64%	0.81%	0.78%	0.67%
Total	4.61%	4.45%	3.19%	3.67%	2.45%	3.45%	3.19%	3.88%	4.78%	5.21%	5.37%

UMLS semantic group: Anatomy

	Scientific literature		**Encyclopedic Web portals**						**Online discussion forums**		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs .com**	**Mayo Clinic**	**MedlinePlus**	**RxList**	**UpToDate**	**Wikipedia Health**	**eHealthForum**	**HealthBoards**	**patient .co.uk**
Anatomical Structure	0.01%	0.01%	0.01%	0.01%	0.00%	0.01%	0.01%	0.01%	0.01%	0.01%	0.02%
Body Location or Region	0.25%	0.21%	0.41%	0.79%	0.39%	0.22%	0.37%	0.47%	1.23%	1.24%	1.49%
Body Part, Organ, or Organ Component	1.37%	1.20%	1.32%	2.91%	1.96%	0.72%	1.74%	2.64%	2.68%	2.37%	2.19%
Body Space or Junction	0.10%	0.27%	0.24%	0.32%	0.30%	0.17%	0.24%	0.50%	0.66%	0.45%	0.56%
Body Substance	0.27%	0.22%	0.33%	0.32%	0.18%	0.37%	0.37%	0.27%	0.35%	0.21%	0.27%
Body System	0.09%	0.06%	0.27%	0.41%	0.30%	0.24%	0.16%	0.20%	0.11%	0.12%	0.09%
Cell	0.97%	1.37%	0.08%	0.17%	0.13%	0.13%	0.39%	0.52%	0.10%	0.06%	0.02%
Cell Component	0.31%	0.53%	0.04%	0.02%	0.01%	0.06%	0.11%	0.29%	0.03%	0.03%	0.02%
Embryonic Structure	0.06%	0.08%	0.04%	0.05%	0.03%	0.07%	0.08%	0.07%	0.03%	0.01%	0.00%
Fully Formed Anatomical Structure	0.00%	0.00%	0.01%	0.00%	0.00%	0.02%	0.00%	0.00%	0.00%	0.00%	0.00%
Tissue	0.40%	0.38%	0.20%	0.47%	0.26%	0.16%	0.36%	0.53%	0.30%	0.25%	0.17%
Total	3.83%	4.33%	2.95%	5.47%	3.56%	2.17%	3.83%	5.50%	5.50%	4.75%	4.83%

UMLS semantic group: Chemicals & Drugs

	Scientific literature		Encyclopedic Web portals						Online discussion forums		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs .com**	**Mayo Clinic**	**Medline-Plus**	**RxList**	**UpTo-Date**	**Wikipedia Health**	**eHealth-Forum**	**Health-Boards**	**patient .co.uk**
Amino Acid, Peptide, or Protein	2.40%	2.59%	1.02%	0.45%	0.90%	1.63%	1.37%	1.75%	0.54%	0.66%	0.56%
Antibiotic	0.12%	0.10%	0.67%	0.30%	0.71%	0.58%	0.29%	0.23%	0.21%	0.25%	0.32%
Biologically Active Substance	1.76%	1.95%	0.58%	0.24%	0.64%	0.78%	0.77%	1.39%	0.16%	0.26%	0.28%
Biomedical or Dental Material	0.29%	0.31%	1.75%	0.63%	0.84%	1.76%	0.39%	0.32%	0.52%	0.38%	0.66%
Carbohydrate	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Chemical	0.08%	0.07%	0.04%	0.05%	0.05%	0.04%	0.03%	0.09%	0.03%	0.03%	0.03%
Chemical Viewed Functionally	0.08%	0.07%	0.22%	0.09%	0.11%	0.14%	0.11%	0.10%	0.02%	0.03%	0.02%
Chemical Viewed Structurally	0.17%	0.14%	0.06%	0.04%	0.05%	0.08%	0.08%	0.13%	0.03%	0.03%	0.02%
Clinical Drug	0.01%	0.01%	0.18%	0.03%	0.18%	0.19%	0.04%	0.01%	0.02%	0.02%	0.02%
Eicosanoid	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Element, Ion, or Isotope	0.33%	0.30%	0.22%	0.19%	0.35%	0.20%	0.26%	0.38%	0.09%	0.12%	0.10%
Enzyme	0.56%	0.55%	0.17%	0.04%	0.09%	0.30%	0.27%	0.41%	0.35%	0.35%	0.36%
Hazardous or Poisonous Substance	0.30%	0.24%	0.28%	0.35%	0.28%	0.25%	0.41%	0.43%	0.14%	0.15%	0.12%
Hormone	0.19%	0.14%	0.59%	0.32%	0.47%	0.85%	0.45%	0.35%	0.21%	0.24%	0.20%
Immunologic Factor	0.59%	0.73%	0.43%	0.29%	0.32%	0.71%	0.63%	0.40%	0.09%	0.11%	0.06%
Indicator, Reagent, or Diagnostic Aid	0.27%	0.43%	0.18%	0.16%	0.04%	0.18%	0.12%	0.18%	0.05%	0.05%	0.05%
Inorganic Chemical	0.29%	0.25%	0.61%	0.34%	0.58%	0.49%	0.23%	0.29%	0.20%	0.20%	0.21%
Lipid	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Neuroreactive Substance or Biogenic Amine	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

(continued on next page)

	Scientific literature		Encyclopedic Web portals						Online discussion forums		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs .com**	**Mayo Clinic**	**Medline-Plus**	**RxList**	**UpTo-Date**	**Wikipedia Health**	**eHealth-Forum**	**Health-Boards**	**patient .co.uk**
Nucleic Acid, Nucleoside, or Nucleotide	0.50%	0.89%	0.17%	0.07%	0.17%	0.36%	0.17%	0.39%	0.02%	0.02%	0.15%
Organic Chemical	2.14%	1.76%	8.09%	4.55%	8.97%	7.44%	3.18%	3.37%	1.62%	2.23%	2.52%
Organo-phosphorus Compound	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Pharmacologic Substance	2.54%	2.37%	12.46%	9.36%	13.43%	10.73%	5.13%	4.58%	1.80%	2.55%	2.87%
Receptor	0.24%	0.22%	0.03%	0.01%	0.01%	0.04%	0.09%	0.20%	0.01%	0.01%	0.01%
Steroid	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Vitamin	0.09%	0.07%	0.35%	0.14%	0.63%	0.13%	0.16%	0.16%	0.08%	0.11%	0.08%
Total	12.95%	13.19%	28.10%	17.65%	28.82%	26.88%	14.18%	15.16%	6.19%	7.80%	8.64%

UMLS semantic group: Concepts & Ideas

	Scientific literature		Encyclopedic Web portals						Online discussion forums		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs .com**	**Mayo Clinic**	**Medline-Plus**	**RxList**	**UpTo-Date**	**Wikipedia Health**	**eHealth-Forum**	**Health-Boards**	**patient .co.uk**
Classification	0.23%	0.22%	0.03%	0.02%	0.01%	0.06%	0.19%	0.19%	0.02%	0.02%	0.01%
Conceptual Entity	1.01%	1.20%	1.47%	2.56%	3.35%	0.98%	0.84%	0.99%	1.05%	1.08%	1.18%
Functional Concept	10.56%	10.63%	10.18%	11.21%	10.42%	9.26%	11.25%	10.44%	5.95%	6.38%	5.60%
Group Attribute	0.03%	0.02%	0.00%	0.01%	0.01%	0.00%	0.01%	0.02%	0.00%	0.00%	0.00%
Idea or Concept	4.06%	3.93%	2.25%	2.66%	2.24%	2.52%	3.68%	3.01%	2.98%	3.02%	2.94%
Intellectual Product	3.46%	3.85%	2.78%	2.23%	2.88%	2.19%	2.28%	2.58%	4.05%	4.38%	4.04%
Language	0.02%	0.02%	0.00%	0.01%	0.13%	0.00%	0.01%	0.03%	0.01%	0.01%	0.01%
Qualitative Concept	12.39%	11.75%	8.37%	10.17%	6.59%	9.01%	10.73%	10.37%	10.53%	10.68%	10.32%
Quantitative Concept	7.13%	8.26%	5.71%	4.60%	2.80%	7.79%	8.22%	5.97%	5.70%	6.24%	5.64%
Regulation or Law	0.05%	0.06%	0.03%	0.00%	0.01%	0.03%	0.05%	0.07%	0.02%	0.02%	0.01%
Spatial Concept	3.46%	4.11%	1.47%	1.69%	1.55%	1.75%	2.77%	3.21%	3.62%	3.59%	3.21%
Temporal Concept	3.30%	3.28%	3.72%	3.55%	2.71%	4.64%	4.52%	3.45%	10.15%	9.35%	10.60%
Total	45.70%	47.33%	36.01%	38.71%	32.70%	38.23%	44.55%	40.33%	44.08%	44.77%	43.56%

UMLS semantic group: Devices

	Scientific literature		**Encyclopedic Web portals**						**Online discussion forums**		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs .com**	**Mayo Clinic**	**Medline-Plus**	**RxList**	**UpTo-Date**	**Wikipedia Health**	**eHealth-Forum**	**Health-Boards**	**patient .co.uk**
Drug Delivery Device	0.00%	0.00%	0.01%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%
Medical Device	0.59%	0.73%	0.56%	0.77%	0.64%	0.58%	0.64%	0.68%	0.82%	0.95%	0.70%
Research Device	0.05%	0.05%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%
Total	0.64%	0.78%	0.57%	0.77%	0.64%	0.59%	0.64%	0.69%	0.82%	0.95%	0.70%

UMLS semantic group: Disorders

	Scientific literature		**Encyclopedic Web portals**						**Online discussion forums**		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs .com**	**Mayo Clinic**	**Medline-Plus**	**RxList**	**UpTo-Date**	**Wikipedia Health**	**eHealth-Forum**	**Health-Boards**	**patient .co.uk**
Acquired Abnormality	0.04%	0.03%	0.02%	0.06%	0.03%	0.02%	0.07%	0.09%	0.05%	0.06%	0.06%
Anatomical Abnormality	0.05%	0.03%	0.01%	0.06%	0.01%	0.02%	0.09%	0.10%	0.06%	0.05%	0.06%
Cell or Molecular Dysfunction	0.15%	0.19%	0.01%	0.00%	0.01%	0.02%	0.04%	0.07%	0.00%	0.00%	0.00%
Congenital Abnormality	0.13%	0.13%	0.07%	0.23%	0.10%	0.08%	0.18%	0.37%	0.30%	0.33%	0.51%
Disease or Syndrome	2.06%	1.35%	2.86%	3.25%	2.80%	3.14%	4.74%	4.12%	2.09%	2.30%	2.56%
Experimental Model of Disease	0.03%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Finding	2.49%	2.40%	4.07%	4.98%	4.17%	3.44%	4.10%	3.69%	5.71%	4.66%	4.57%
Injury or Poisoning	0.37%	0.22%	0.31%	0.39%	0.39%	0.31%	0.50%	0.53%	0.34%	0.28%	0.25%
Mental or Behavioral Dysfunction	0.32%	0.18%	0.40%	0.54%	0.55%	0.45%	0.34%	0.91%	0.70%	0.69%	0.88%
Neoplastic Process	0.87%	0.46%	0.27%	0.51%	0.56%	0.38%	0.82%	0.80%	0.32%	0.32%	0.25%
Pathologic Function	0.59%	0.36%	0.94%	0.85%	0.68%	1.33%	1.43%	0.86%	0.50%	0.32%	0.37%
Sign or Symptom	0.35%	0.26%	3.11%	4.61%	3.41%	1.80%	1.59%	1.40%	2.51%	2.14%	2.92%
Total	7.45%	5.62%	12.07%	15.48%	12.71%	10.99%	13.90%	12.94%	12.58%	11.15%	12.43%

UMLS semantic group: Genes & Molecular Sequences

	Scientific literature		**Encyclopedic Web portals**						**Online discussion forums**		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs .com**	**Mayo Clinic**	**Medline-Plus**	**RxList**	**UpTo-Date**	**Wikipedia Health**	**eHealth-Forum**	**Health-Boards**	**patient .co.uk**
Amino Acid Sequence	0.04%	0.17%	0.00%	0.00%	0.00%	0.00%	0.02%	0.06%	0.00%	0.00%	0.00%
Carbohydrate Sequence	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Gene or Genome	2.07%	3.15%	0.52%	0.40%	0.65%	0.39%	0.59%	1.17%	1.84%	1.89%	1.93%
Molecular Sequence	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Nucleotide Sequence	0.09%	0.20%	0.00%	0.00%	0.00%	0.00%	0.01%	0.04%	0.00%	0.00%	0.00%
Total	2.20%	3.52%	0.52%	0.40%	0.65%	0.39%	0.62%	1.27%	1.84%	1.89%	1.93%

UMLS semantic group: Geographic Areas

	Scientific literature		**Encyclopedic Web portals**						**Online discussion forums**		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs .com**	**Mayo Clinic**	**Medline-Plus**	**RxList**	**UpTo-Date**	**Wikipedia Health**	**eHealth-Forum**	**Health-Boards**	**patient .co.uk**
Geographic Area	0.58%	0.59%	0.17%	0.10%	0.39%	0.14%	0.34%	1.07%	0.56%	0.50%	0.44%
Total	0.58%	0.59%	0.17%	0.10%	0.39%	0.14%	0.34%	1.07%	0.56%	0.50%	0.44%

UMLS semantic group: Living Beings

	Scientific literature		Encyclopedic Web portals						Online discussion forums		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs .com**	**Mayo Clinic**	**Medline-Plus**	**RxList**	**UpTo-Date**	**Wikipedia Health**	**eHealth-Forum**	**Health-Boards**	**patient .co.uk**
Age Group	0.33%	0.20%	0.73%	0.89%	0.44%	0.51%	0.77%	0.44%	0.81%	0.40%	0.24%
Amphibian	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%
Animal	0.21%	0.69%	0.20%	0.28%	0.07%	0.26%	0.18%	0.37%	0.16%	0.16%	0.11%
Archaeon	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Bacterium	0.25%	0.18%	0.14%	0.07%	0.11%	0.19%	0.16%	0.20%	0.04%	0.03%	0.03%
Bird	0.05%	0.04%	0.01%	0.01%	0.01%	0.00%	0.01%	0.05%	0.02%	0.03%	0.03%
Eukaryote	0.30%	0.30%	0.04%	0.09%	0.05%	0.02%	0.07%	0.29%	0.36%	0.32%	0.39%
Family Group	0.18%	0.19%	0.06%	0.17%	0.08%	0.12%	0.15%	0.27%	1.02%	0.95%	0.74%
Fish	0.08%	0.05%	0.02%	0.01%	0.02%	0.02%	0.01%	0.03%	0.03%	0.03%	0.05%
Fungus	0.07%	0.06%	0.02%	0.01%	0.02%	0.02%	0.03%	0.05%	0.02%	0.02%	0.01%
Group	0.07%	0.06%	0.01%	0.01%	0.00%	0.01%	0.02%	0.06%	0.01%	0.01%	0.02%
Human	0.32%	0.32%	0.16%	0.02%	0.06%	0.31%	0.18%	0.38%	0.05%	0.03%	0.03%
Mammal	0.48%	0.51%	0.14%	0.02%	0.04%	0.36%	0.05%	0.28%	0.10%	0.08%	0.09%
Organism	0.16%	0.14%	0.04%	0.02%	0.01%	0.06%	0.07%	0.10%	0.01%	0.00%	0.00%
Patient or Disabled Group	1.11%	0.56%	1.09%	0.34%	0.15%	2.18%	2.09%	0.66%	0.12%	0.14%	0.39%
Plant	0.30%	0.28%	0.05%	0.06%	0.35%	0.03%	0.08%	0.29%	0.15%	0.15%	0.20%
Population Group	0.92%	0.71%	0.66%	0.95%	1.16%	0.76%	0.73%	1.25%	1.58%	1.21%	0.93%
Professional or Occupational Group	0.43%	0.43%	1.40%	1.35%	1.89%	0.44%	0.53%	0.67%	0.82%	0.94%	1.06%
Reptile	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.01%	0.00%	0.00%	0.00%
Vertebrate	0.01%	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Virus	0.21%	0.21%	0.07%	0.08%	0.10%	0.10%	0.18%	0.22%	0.08%	0.06%	0.02%
Total	5.51%	4.97%	4.84%	4.38%	4.56%	5.39%	5.31%	5.63%	5.38%	4.56%	4.34%

UMLS semantic group: Objects

	Scientific literature		**Encyclopedic Web portals**						**Online discussion forums**		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs.com**	**Mayo Clinic**	**MedlinePlus**	**RxList**	**UpToDate**	**Wikipedia Health**	**eHealthForum**	**HealthBoards**	**patient.co.uk**
Entity	0.08%	0.16%	0.29%	0.19%	0.32%	0.12%	0.07%	0.12%	0.65%	0.76%	0.56%
Food	0.45%	0.43%	0.46%	0.81%	0.94%	0.19%	0.35%	0.42%	0.73%	0.82%	0.86%
Manufactured Object	1.24%	1.63%	0.99%	1.14%	1.32%	0.93%	0.91%	1.66%	1.95%	2.08%	1.94%
Physical Object	0.06%	0.04%	0.05%	0.06%	0.07%	0.04%	0.09%	0.07%	0.02%	0.02%	0.02%
Substance	0.47%	0.47%	0.26%	0.23%	0.28%	0.18%	0.18%	0.28%	0.11%	0.11%	0.08%
Total	2.30%	2.73%	2.05%	2.43%	2.93%	1.46%	1.60%	2.55%	3.46%	3.79%	3.46%

UMLS semantic group: Occupations

	Scientific literature		**Encyclopedic Web portals**						**Online discussion forums**		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs.com**	**Mayo Clinic**	**MedlinePlus**	**RxList**	**UpToDate**	**Wikipedia Health**	**eHealthForum**	**HealthBoards**	**patient.co.uk**
Biomedical Occupation or Discipline	0.30%	0.19%	0.24%	0.05%	0.26%	0.27%	0.19%	0.32%	0.07%	0.07%	0.06%
Occupation or Discipline	0.25%	0.23%	0.09%	0.11%	0.07%	0.07%	0.12%	0.23%	0.13%	0.13%	0.11%
Total	0.55%	0.42%	0.33%	0.16%	0.33%	0.34%	0.31%	0.55%	0.20%	0.20%	0.17%

UMLS semantic group: Organizations

	Scientific literature		**Encyclopedic Web portals**						**Online discussion forums**		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs.com**	**Mayo Clinic**	**MedlinePlus**	**RxList**	**UpToDate**	**Wikipedia Health**	**eHealthForum**	**HealthBoards**	**patient.co.uk**
Health Care Related Organization	0.20%	0.17%	0.50%	0.22%	0.49%	0.27%	0.23%	0.31%	0.17%	0.20%	0.33%
Organization	0.19%	0.24%	0.13%	0.04%	0.37%	0.08%	0.12%	0.40%	0.19%	0.17%	0.15%
Professional Society	0.01%	0.00%	0.00%	0.00%	0.00%	0.00%	0.02%	0.01%	0.00%	0.00%	0.00%
Self-help or Relief Organization	0.01%	0.00%	0.00%	0.02%	0.00%	0.00%	0.00%	0.01%	0.01%	0.01%	0.02%
Total	0.41%	0.41%	0.63%	0.28%	0.86%	0.35%	0.37%	0.73%	0.37%	0.38%	0.50%

UMLS semantic group: Phenomena

	Scientific literature		Encyclopedic Web portals						Online discussion forums		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs .com**	**Mayo Clinic**	**Medline-Plus**	**RxList**	**UpTo-Date**	**Wikipedia Health**	**eHealth-Forum**	**Health-Boards**	**patient .co.uk**
Biologic Function	0.09%	0.06%	0.03%	0.00%	0.01%	0.06%	0.05%	0.05%	0.01%	0.00%	0.00%
Environmental Effect of Humans	0.03%	0.01%	0.01%	0.02%	0.02%	0.01%	0.02%	0.02%	0.05%	0.04%	0.04%
Human-caused Phenomenon or Process	0.09%	0.07%	0.12%	0.05%	0.05%	0.13%	0.05%	0.08%	0.03%	0.03%	0.02%
Laboratory or Test Result	0.07%	0.07%	0.05%	0.04%	0.03%	0.07%	0.10%	0.06%	0.03%	0.03%	0.02%
Natural Phenomenon or Process	0.68%	0.59%	0.43%	0.49%	0.38%	0.33%	0.38%	0.55%	0.50%	0.45%	0.42%
Phenomenon or Process	0.46%	0.47%	0.22%	0.27%	0.15%	0.15%	0.32%	0.48%	0.42%	0.39%	0.30%
Total	1.42%	1.27%	0.86%	0.87%	0.64%	0.75%	0.92%	1.24%	1.04%	0.94%	0.80%

UMLS semantic group: Physiology

	Scientific literature		Encyclopedic Web portals						Online discussion forums		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs .com**	**Mayo Clinic**	**Medline-Plus**	**RxList**	**UpTo-Date**	**Wikipedia Health**	**eHealth-Forum**	**Health-Boards**	**patient .co.uk**
Cell Function	0.48%	0.45%	0.05%	0.01%	0.02%	0.09%	0.12%	0.20%	0.01%	0.01%	0.00%
Clinical Attribute	0.47%	0.34%	0.41%	0.55%	0.29%	0.56%	0.58%	0.28%	0.11%	0.10%	0.11%
Genetic Function	0.47%	0.60%	0.02%	0.02%	0.01%	0.04%	0.14%	0.27%	0.01%	0.01%	0.01%
Mental Process	1.65%	1.48%	1.10%	1.94%	1.55%	0.87%	1.05%	1.54%	5.89%	6.42%	6.13%
Molecular Function	0.48%	0.47%	0.27%	0.33%	0.09%	0.27%	0.16%	0.24%	0.03%	0.06%	0.05%
Organ or Tissue Function	0.25%	0.16%	0.23%	0.21%	0.12%	0.26%	0.31%	0.29%	0.29%	0.20%	0.19%
Organism Attribute	0.71%	0.61%	0.21%	0.23%	0.19%	0.37%	0.37%	0.39%	0.61%	0.58%	0.50%
Organism Function	0.67%	0.54%	0.87%	1.35%	0.86%	0.92%	0.74%	0.84%	2.30%	1.38%	1.18%
Physiologic Function	0.15%	0.13%	0.19%	0.22%	0.22%	0.23%	0.19%	0.21%	0.32%	0.41%	0.35%
Total	5.33%	4.78%	3.35%	4.86%	3.35%	3.61%	3.66%	4.26%	9.57%	9.17%	8.52%

UMLS semantic group: Procedures

	Scientific literature		Encyclopedic Web portals						Online discussion forums		
UMLS semantic type	**PubMed MEDLINE**	**PubMed Central**	**Drugs .com**	**Mayo Clinic**	**Medline-Plus**	**RxList**	**UpTo-Date**	**Wikipedia Health**	**eHealth-Forum**	**Health-Boards**	**patient .co.uk**
Diagnostic Procedure	0.61%	0.35%	0.14%	0.47%	0.04%	0.17%	0.75%	0.39%	0.36%	0.36%	0.36%
Educational Activity	0.09%	0.06%	0.01%	0.04%	0.01%	0.01%	0.04%	0.08%	0.05%	0.06%	0.04%
Health Care Activity	1.27%	0.90%	1.89%	1.92%	3.57%	1.21%	1.37%	0.83%	1.76%	1.90%	2.25%
Laboratory Procedure	1.36%	1.41%	0.55%	0.48%	0.26%	1.03%	0.96%	0.62%	0.37%	0.34%	0.35%
Molecular Biology Research Technique	0.15%	0.22%	0.01%	0.00%	0.00%	0.02%	0.03%	0.03%	0.00%	0.00%	0.00%
Research Activity	1.14%	1.10%	0.41%	0.09%	0.30%	0.94%	0.61%	0.40%	0.11%	0.12%	0.10%
Therapeutic or Preventive Procedure	1.91%	1.54%	1.33%	1.73%	1.23%	1.87%	2.84%	1.88%	1.02%	1.14%	1.15%
Total	6.53%	5.58%	4.34%	4.73%	5.41%	5.25%	6.60%	4.23%	3.67%	3.92%	4.25%

Appendix D

List of ambiguous entity names and candidates for the political domain

Candidate names are spelled exactly as their corresponding Wikipedia page titles.

List of abbreviations:

Ambiguous entity name	Candidates
ACP	African, Caribbean and Pacific Group of States ACP Magazines Alliance for Climate Protection American Communist Party A Connecticut Party Alliance for Climate Protection American Communist Party Armenian Communist Party Panama Canal Authority
ADL	Anti-Defamation League Armenian Democratic Liberal Party
ADM	Admiral Archer Daniels Midland Assyrian Democratic Movement
AFP	Armed Forces of the Philippines Argentine Federal Police Australian Federal Police Austrian Federal Police Alliance of the Forces of Progress (Benin) Alliance of the Forces of Progress (Senegal) America First Party (disambiguation) Americans for Prosperity Australia First Party
ANC	African National Congress Armenian National Congress Advisory Neighborhood Commission Assemblea Nacional Catalana ABS-CBN News Channel
ANZ	Antarctica New Zealand Australia and New Zealand Banking Group ANZ Bank New Zealand ANZ (Fiji)
AUS	Army of the United States Australia Aus, Namibia
ACT	Australian Capital Territory Allied Command Transformation ACT Alberta ACT Alliance

Ambiguous entity name	**Candidates**
	ACT New Zealand America Coming Together Alliance for Change and Transparency
BBC	Banahaw Broadcasting Corporation British Broadcasting Company Bangkok Bank of Commerce Biplobi Bangla Congress
CAC	Corporate Affairs Commission, Nigeria Canadian Aviation Corps Campaign Against Censorship Central Advisory Commission Coalition against Communalism
CDU	Cameroon Democratic Union Christian Democratic Union (Germany) Christian Democratic Union (Ukraine) Croatian Democratic Union Democratic Unitarian Coalition United Christian Democrats United Democratic Centre (El Salvador)
CHP	Christian Heritage Party of Canada Christian Heritage Party of British Columbia Christian Heritage Party of New Zealand Republican People's Party (Turkey)
COE	Center of excellence NATO Centres of Excellence Afghanistan-Pakistan Center of Excellence Church of England Council of Europe United States Army Corps of Engineers
CPL	Communist Party of Latvia Communist Party of Lithuania Corporal
CPP	Cambodian People's Party Communist Party of Pakistan Communist Party of the Philippines Convention People's Party Patriotic Pan-African Convergence Centre for Public Policy
DNC	Delaware North Companies Democratic National Committee Democratic National Convention
DPA	Deutsche Presse-Agentur Democratic Party of Albanians Democratic Progressive Alliance Doctor of Public Administration United Nations Department of Political Affairs

Ambiguous entity name	**Candidates**
DPP	Danish People's Party Democratic Party of the Philippines Democratic Progressive Party Democratic Progressive Party (Malawi) Democratic Progressive Party (Singapore)
ECB	European Central Bank European Chemicals Bureau Equatorial Commercial Bank
GCC	Gulf Cooperation Council Garde côtière canadienne Glasgow City Council Global Climate Coalition
ICA	ICA AB Ica, Peru Immigration and Checkpoints Authority International Court of Arbitration Islamic Consultative Assembly
IND	India Indianapolis
ITC	ITC Entertainment ITC Limited International Trade Centre Intertropical Convergence Zone International Teledemocracy Centre Independent Television Commission Information and communications technology
MAS	Monetary Authority of Singapore Malaysia Mouvement pour une Alternative Socialiste Movement for Socialism (Argentina) Movement toward Socialism (Bolivia) Broad Social Movement Movement toward Socialism (Venezuela)
NBC	NBC Norwegian Broadcasting Corporation Namibian Broadcasting Corporation Nation Broadcasting Corporation National Bank of Canada National Business Center Naval Base Coronado
NPD	National Democratic Party of Germany New Democratic Party of Canada
NPR	NPR New Port Richey, Florida
NYC	New York City

Ambiguous entity name	**Candidates**
	Youth Development Administration North York Centre
ODA	Civic Democratic Alliance Organization for Democratic Action
PFP	Party of Freedom and Progress People First Party (South Korea) People First Party (Republic of China) Peace and Freedom Party Progressive Federal Party Popular Front Party Federal Preventive Police Partnership for Peace
PGA	Parliamentarians for Global Action Peoples' Global Action
PLA	Palestinian Liberation Army Party of Labour of Albania People's Liberation Army People's Liberation Army of Manipur ProLife Alliance
PPI	Italian People's Party (1994) Peace Party of India Pirate Party (Iceland) Pirate Parties International Professionals Party of India
PTE	Workers' Party of Ecuador Party of Labour of Spain Private (rank)
ROK	Republic of Korea Republic of Kosovo Rok River
SAIC	Leidos South African Indian Congress
SBS	SBS Broadcasting Group Special Broadcasting Service Seoul Broadcasting System Spanish Broadcasting System

List of organizations:

Ambiguous entity name	**Candidates**
Blackberry	BlackBerry (company) BlackBerry Blackberry Township, Itasca County, Minnesota Blackberry Township, Kane County, Illinois

Ambiguous entity name	**Candidates**
	Blackberry
Bosch	Bosch, Netherlands Bosch en Duin Den Bosch Villa Bosch Robert Bosch GmbH
Broad Front	Broad Front (Argentina) Broad Front (Paraguay) Broad Front (Uruguay) Broad Front (Costa Rica) Broad Left Front (Peru) Broad Front for Democracy Socialist Party – Broad Front of Ecuador
Bundesrat	Federal Council (Austria) Bundesrat of Germany Federal Council (Switzerland)
Democratic Progressive Party	Democratic Progressive Party (Argentina) Democratic Progressive Party (Austria) Democratic Progressive Party (Malawi) Democratic Progressive Party (Singapore) Democratic Progressive Party Progressive Democratic Party (Tunisia) Sammarinese Democratic Progressive Party
Democratic Rally	Democratic Rally Democratic Rally (France) Democratic Rally (Senegal) Central African Democratic Rally Martinican Democratic Rally Oceanian Democratic Rally
Dow	Dow Jones Industrial Average Dow, California Dow, Kentucky
Front line	Front Line Defenders Frontline States Front Line (political party)
Gallup	Gallup (company) Gallup International Association Gallup, Kentucky Gallup, New Mexico Gallup, South Dakota Alec Gallup David Gallup George Gallup Gallup Glacier
House	Lower house House of Commons

Ambiguous entity name	Candidates
	House of Representatives United States House of Representatives Upper house House of Lords House, New Mexico House, North Carolina Douglas House (Arkansas politician)
J. P. Morgan	J. P. Morgan J. P. Morgan, Jr. JPMorgan Chase
Junta	Military junta Junta (Habsburg)
Millennium	Millennium Development Goals Millennium Kids White House Millennium Council Millennium Stadium Millennium Summit
NATO headquarters	NATO Supreme Headquarters Allied Powers Europe Allied Command Transformation
New Era	New Era, Indiana New Era, Michigan New Era, Oregon New Era Park, Sacramento, California A New Era New Era Party
Panther	Panther, Daviess County, Kentucky Black Panther Party Gray Panthers White Panther Party Black Panthers (Israel) Polynesian Panthers
Rouge	Khmer Rouge Rõuge Rouge, Toronto The Rouge Baton Rouge, Louisiana
Rover	Rover (marque) Rover Company Rover Group MG Rover Group Land Rover Group Freight Rover Land Rover
Schindler	Schindler Group Emilie Schindler

Ambiguous entity name	Candidates
	Oskar Schindler
Shalom	Silvan Shalom Brit Tzedek v'Shalom Brit Shalom (political organization) Gush Shalom Hevel Shalom Neve Shalom
Shell	Royal Dutch Shell Shell Oil Company Shell Canada Shell Nigeria Shell corporation Shell, Ecuador Shell, California Shell, Wyoming
Thomas Cook	Thomas Cook Group Thomas Cook AG Thomas Cook & Son Thomas Cook Airlines Belgium Thomas Cook Airlines Thomas Cook Airlines Scandinavia T. Cooke & Sons
Yoruba	Yoruba people Yoruba language Yoruba culture Yoruba religion
Young America	Young America, Indiana Young America, Wisconsin Norwood Young America, Minnesota Young America Township, Carver County, Minnesota Young America Township, Edgar County, Illinois Young America Lake Young America movement Young America's Foundation Young Americans for Freedom

List of persons:

Ambiguous entity name	Candidates
Bains	Bains, Haute-Loire Bains (Mirpur) Bains, Louisiana Hardial Bains Harry Bains Navdeep Bains

Ambiguous entity name	Candidates
Bush	George H. W. Bush George W. Bush Jeb Bush Bush family Bush, Cornwall Bush, Saskatchewan Bush Island (Nunavut) Bush, Illinois Bush, Louisiana
Calhoun	Bootsie Calhoun Charles Calhoun, Jr. Calhoun, Georgia Calhoun, Illinois Calhoun, Kentucky Calhoun, Missouri Calhoun, South Carolina Calhoun, Tennessee Calhoun, West Virginia
Constance	Konstanz Lake Constance Constance Bay, Ottawa Constance, Minnesota Constance (Portugal) Mount Constance Andrew Constance Angela Constance Ansley Constance
Faris	Faris Glubb Faris Odeh Al-Faris Faris Faris Island
Gaillard	Gaillard Château-Gaillard, Ain Château Gaillard Gaillard Island La Gaillarde Brive-la-Gaillarde Claude Gaillard Geneviève Gaillard Micha Gaillard
Gloria	Gloria, Lafayette Parish, Louisiana Gloria, Oriental Mindoro Gloria Cultural Arena Gloria Macapagal-Arroyo
Herzog	Chaim Herzog Aura Herzog

Ambiguous entity name	Candidates
	Gustav Herzog Isaac Herzog Maurice Herzog Roman Herzog Herzog Mountains
Jeb	Jeb Bardon Jeb Bradley Jeb Bush Jeb Hensarling Jeb Stuart Magruder Jeb Spaulding
Jamieson	Cathy Jamieson Don Jamieson (politician) Margaret Jamieson Norma Jamieson Stuart Jamieson
Karamanlis	Caramania Konstantinos Karamanlis Kostas Karamanlis Karamanlı, Burdur Qaramanlı
Lance	Bert Lance Leonard Lance Łańce
MacLeod	Fort Macleod, Alberta McLeod (Edmonton) McLeod County, Minnesota McLeod, North Dakota Macleod, Victoria McLeod Ganj McLeod, Texas Macleod (electoral district) Macleod (provincial electoral district) Lake Macleod
Michael Collins	Michael Collins (Irish leader) Michael Collins (Limerick politician)
Olli	Egil Olli Olli Rehn
Prentice	Prentice, Wisconsin Bridget Prentice Christopher Prentice Jim Prentice
Prince of Wales	Prince of Wales Prince of Wales, New Brunswick Prince of Wales Strait Cape Prince of Wales

Ambiguous entity name	**Candidates**
	Prince of Wales Mountains Prince of Wales Range
Ricardo	Ricardo Hausmann Ricardo Lagos John Lewis Ricardo
Sana	Sana’a Sana, Haute-Garonne Sana, Bhutan Sana, Chalkidiki Sana (river) Saña, Peru
Skelton	Skelton, Cumbria Skelton, East Riding of Yorkshire Skelton, North Yorkshire Skelton-on-Ure Skelton, York Skelton-in-Cleveland North Skelton Skelton, Indiana Ike Skelton
Soledad	Soledad Alvear Soledad, California Soledad, Atlántico Soledad Atzompa Soledad de Doblado La Soledad, Tamaulipas
Stockwell	Stockwell Stockwell, Indiana Stockwell, South Australia Chris Stockwell Stockwell Day
Summers	Summers, Arkansas Summers, California Summers, West Virginia Summers County, West Virginia
Theodore	Theodore, Alabama Theodore, Australian Capital Territory Theodore, Queensland Theodore, Saskatchewan Theodore Roosevelt
Yolanda	Yolanda King Typhoon Haiyan

List of geographical places:

Ambiguous entity name	**Candidates**
Anaconda	Anaconda, Missouri Anaconda, Montana Anaconda, British Columbia Anaconda Range
Ankara	Ankara Ankara Province Ankara University Ankara Central railway station Ankara Castle Ankara River Greater Ankara
Bombay	Bombay Bombay State Isle of Bombay New Bombay Bombay, New York Bombay Beach, California
Cartagena	Cartagena, Chile Cartagena, Colombia Cartagena Province Cartagena del Chairá Cartagena, Spain Campo de Cartagena Carthagena, Ohio Carlos Mauricio Funes Cartagena Nicolás Nogueras Cartagena
Delphi	Delphi Delphi, Indiana Delphi, County Mayo
Estonia	Estonia Estonia, Abkhazia Estonia, Altai Krai Estonia Mine Estonia (peak)
Fargo	Fargo, North Dakota Fargo, Arkansas Fargo, California Fargo, Georgia Fargo, Oklahoma Fargo, Wisconsin Wells Fargo
Georgia	Georgia (country) Georgia (U.S. state) Georgia, Indiana Georgia, New Jersey Georgia, Vermont New Georgia

Ambiguous entity name	**Candidates**
	South Georgia and the South Sandwich Islands Strait of Georgia
Great Lakes	Great Lakes African Great Lakes Great Lake (Britain) Great Lake, Tasmania Great Lakes region
Hampden	Hampden, New Zealand Hampden, Newfoundland and Labrador Hampden, Quebec Hampden Park Hampden Park, Eastbourne Hampden, Maine Hampden, Baltimore Hampden, Massachusetts Hampden, North Dakota
Jaya	Petaling Jaya Putrajaya Seberang Jaya Puncak Jaya
Kremlin	Kremlin The Kremlin Government of the Soviet Union Government of Russia Kremlin, Montana Kremlin, Oklahoma Kremlin, Virginia Kremlin, Wisconsin
Missouri	Missouri Missouri River Missouri City, Texas Missouri Rhineland Missouri Territory Missouri bellwether
Niger	Niger Niger River Niger State
Paisley	Paisley Paisley, Florida Paisley, Oregon Paisley Caves Paisley, Ontario Paisley, Edmonton Paisley Islet Paisley (UK Parliament constituency) Ian Paisley Ian Paisley, Jr.

Ambiguous entity name	**Candidates**
Palo Alto	Palo Alto, California East Palo Alto, California Palo Alto County, Iowa Palo Alto, Mississippi Palo Alto, Pennsylvania Palo Alto, Texas Palo Alto, Aguascalientes
Potsdam	Potsdam Potsdam-Mittelmark Potsdam (Papua New Guinea) Potsdam, Eastern Cape Potsdam (town), New York Potsdam (village), New York Potsdam, Ohio
RO	Ro, Greece Ro, Emilia–Romagna Rø Romania Rondônia Reserve Officers' Training Corps
Saxony	Saxony Saxony (wine region) Sachsen bei Ansbach
South of the Border	England and Wales Mexico United States
Southland	Dunbar–Southlands Southland, New Zealand Southland District Southland Plains Chicago Southland Greater Los Angeles Area Los Angeles metropolitan area Southland, Texas
Stade	Stade Stade (district) Stade (region) Stade de France Stade Louis II
Thames	River Thames Thames Estuary Thames River (Ontario) Thames River (Connecticut) Waihou/Thames River Thames, New Zealand
Zaire	Zaire

Ambiguous entity name	Candidates
	Zaire Province Congo River
Zanzibar	Zanzibar Unguja Zanzibar City Zanzibar Archipelago Zanzibar Protectorate Zanzibar Sultanate