2016

BIOMEDICAL WORD SENSE DISAMBIGUATION WITH NEURAL WORD AND CONCEPT EMBEDDINGS

AKM Sabbir
University of Kentucky, sabbirsourov@gmail.com
Digital Object Identifier: https://doi.org/10.13023/ETD.2016.490

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

AKM Sabbir, Student

Dr. Ramakanth Kavuluru, Major Professor

Dr. Miroslaw Truszczynski, Director of Graduate Studies

Biomedical Word Sense Disambiguation with
Neural Word and Concept Embeddings

THESIS

A thesis submitted in partial
fulfillment of the requirements for
the degree of Master of Science in
the College of Engineering at the
University of Kentucky

By
AKM Sabbir
Lexington, Kentucky

Director: Dr. Ramakanth Kavuluru, Professor of Internal Medicine and Computer Science
Lexington, Kentucky 2016

ABSTRACT OF THESIS

Biomedical Word Sense Disambiguation with
Neural Word and Concept Embeddings

Addressing ambiguity issues is an important step in natural language processing (NLP) pipelines designed for information extraction and knowledge discovery. This problem is also common in biomedicine where NLP applications have become indispensable to exploit latent information from biomedical literature and clinical narratives from electronic medical records. In this thesis, we propose an ensemble model that employs recent advances in neural word embeddings along with knowledge based approaches to build a biomedical word sense disambiguation (WSD) system. Specifically, our system identifies the correct sense from a given set of candidates for each ambiguous word when presented in its context (surrounding words). We use the MSH WSD dataset, a well known public dataset consisting of 203 ambiguous terms each with nearly 200 different instances and an average of two candidate senses represented by concepts in the unified medical language system (UMLS). We employ a popular biomedical concept, Our linear time (in terms of number of senses and context length) unsupervised and knowledge based approach improves over the state-of-the-art methods by over 3% in accuracy. A more expensive approach based on the k-nearest neighbor framework improves over prior best results by 5% in accuracy. Our results demonstrate that recent advances in neural dense word vector representations offer excellent potential for solving biomedical WSD.

KEYWORDS: word sense disambiguation, neural word embeddings, knowledge based systems, UMLS, MetaMap

Author's signature: AKM Sabbir

Date: December 11, 2016

Biomedical Word Sense Disambiguation with
Neural Word and Concept Embeddings

By
AKM Sabbir

Director of Thesis: Ramakanth Kavuluru

Director of Graduate Studies: Miroslaw Truszczynski

Date: December 11, 2016

To my friends and family, thank you!

ACKNOWLEDGMENTS

I would like to thank the following people for helping me complete the project presented in this thesis:

- First I would like to thank my advisor Dr. Ramakanth Kavuluru, for introducing me to the biomedical word sense disambiguation (WSD) problem and helping me all the way to end – giving me methodological feedback and making appropriate course corrections along the way and helping me with technical writing aspects relevant to the thesis manuscript.
- I would also like to thank student members in my advisor's group (the UKNLP lab) for engaging me in thoughtful technical discussions about machine learning and information extraction methods.
- Thanks to my parents for continually inspiring me.
- Thanks to Dr. Antonio Jimeno Yepes for helping me understand his WSD model, which really helped, in a complementary fashion, the approach I present in this thesis.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Introduction

Word Sense Disambiguation (WSD) [25] is the task of detecting the correct sense of a word based on the context among multiple senses it might assume. Thus sense represents one of the many meanings of an ambiguous word. For instance, the word 'cold' can mean different concepts based on the context. In "the air in the center of the vortex of a cyclone is generally very cold", cold refers to temperature. Let us consider the example, "I could not come to office last week because I had a cold". Here cold means fever or respiratory infection. Two completely different senses of the same word occur because the contexts are different. WSD is an important problem as it applies to a number of natural language processing (NLP) tasks such as text to speech conversion [1], machine translation [2,3], summary generation, information extraction, information retrieval, concept mapping and it also has impact on accuracy of biomedical applications such as biomedical coding and indexing [18, 34], relation extraction [5, 24], and knowledge discovery [8, 20]. One of the initial steps of any of the previously mentioned NLP tasks is to retrieve the correct sense of any ambiguous word present in input texts. Disambiguation is not as intuitive as in human communication because any ambiguity that arose while having human conversations is typically resolved as the conversation continues due to years of experience in the complexities of dealing with linguistic phenomena. However, with the development of modern technology, we now store information in computer systems so that we can retrieve the same information and use it later quickly. In order to retrieve the information correctly the computer system must be able to resolve any ambiguity that arises without any external help from the document creator. In order to assist computers to deal with this problem, NLP researchers have been working on WSD. This thesis focuses on developing new methods for biomedical WSD using a well known larger public dataset.

The methods we present in this thesis ensemble predictive information from two different approaches. These two approaches complement each other to form our final model. One of them is the Markov chain based conditional probability distribution approach that measures how much a document is associated with a given biomedical sense using Bayes rule. The second uses neural word and concept embeddings to measure the angular similarity between context and concept vectors; it also measures the magnitude of projection of document vectors along the concept vectors. We demonstrate that combinations of these three measurements lead to improvements over prior state-of-the-art in biomedical WSD on a particular public dataset called the MSH WSD dataset [16]. Details of our methods are in the final chapter. The fundamental contribution of this thesis can be outlined as follows.

- We develop weakly supervised methods that apply recent developments in deep neural networks for NLP tasks for biomedical WSD. Specifically, we learn word and concept (sense) vectors based on a large corpus of biomedical abstracts using a well known neural word embedding framework called word2vec [27]. We then

compare using different similarity metrics the concepts associated with each ambiguous term to the word context vector of testing instances and choose the best concept that maximizes this similarity. We improve our the best results generated through unsupervised approaches in published literature on a public dataset by over 3% accuracy using these linear time approaches.

- We propose a more expensive k-nearest neighbors approach that uses our linear time WSD methods to create a distantly supervised training dataset. By considering judgements of the top k neighbors that are most similar to a new test context, we predict its correct sense. Although expensive, this provides an absolute 2% improvement in accuracy over linear time methods.

The rest of the thesis is organized as follows. Chapter 2 discusses background information for WSD including different prior approaches to tackle it specifically machine learning and information extraction methods. Chapter 3 presents our core methods and results.

Chapter 2 Background and Prior Work

For this thesis, WSD specifically deals with identifying the correct sense of a term, among a set of given candidate senses for that term, when it is presented in a brief narrative along with its context (surrounding text). We start this chapter with a specific biomedical example. Consider the ambiguous word 'discharge'. It has two unique senses in biomedicine – (S1). The first is the administrative process of releasing a patient from a healthcare facility following an in-patient stay for some treatment or procedure. (S2). The second sense pertains to bodily secretions of certain fluids from an orifice or wound. In our task the ambiguous word discharge is specified along with the sense set {S1, S2} and an example context – "Low risk patients identified using CADILLAC risk score with STEMI treated successfully with primary PCI have a low adverse event rate on the third day or later of hospitalization suggesting that an earlier **discharge** is safe in properly selected patients." Our goal is to identify the correct sense S1 for this specific occurrence of 'discharge'.

For a thorough survey of approaches to WSD, please see the survey by Navigli [30], which suggests mainly three categories – supervised, knowledge-based, and unsupervised approaches. Supervised approaches for WSD [40,47] use a labeled dataset along with interesting lexical/syntactic features derived from the context around the term to build machine learned models that predict the correct sense in unseen test contexts. Knowledge based approaches [16, 26] do not use any corpus but solely rely on thesauri or sense inventories such as WordNet and the Unified Medical Language System (UMLS) that contain brief definitions of different senses and corresponding synonyms. Unsupervised approaches may employ topic modeling [21] based methods to disambiguate when the senses are known ahead of time. Some unsupervised approaches [42] are often referred to as performing word sense *discrimination* or *induction* as opposed to disambiguation because they employ clustering approaches where different clusters are expected to represent the different senses, which are not known a priori.

2.1 WSD in Biomedicine

In biomedicine, knowledge-based word sense disambiguation efforts mostly relied on the UMLS knowledge base [29], which contains over 3.4 million unique concepts expertly sourced from nearly 200 different terminologies in biomedicine and allied fields. The UMLS is maintained by the US National Library of Medicine (NLM) and is updated every year to reflect new concepts and other changes. For each concept in the UMLS, there is usually a brief definition and sometimes additional relations (both hierarchical and associative) connecting it with other concepts. Each concept has a unique ID called the concept unique identifier (CUI), an alphanumeric string that starts with a 'C'. For example, the sense S1 (administrative process) for 'discharge' discussed earlier is represented by CUI C0030685 and sense S2 (body substance) is represented by the CUI C0012621. S1 has a short definition "The

administrative process of discharging the patient, alive or dead, from hospitals or other health facilities". For S2 we notice the definition – "In medicine, a fluid that comes out of the body. Discharge can be normal or a sign of disease." In the MSH WSD dataset that we use in this thesis, the candidate senses for each ambitious word are represented in the form of these unique CUIs. The task is to identify the correct CUI given a particular context (few sentences) containing an ambiguous word. For the rest of the manuscript, we use the three terms *CUI*, *concept*, and *sense* synonymously as they refer to the same notion.

Schuemie et al. [38] present a nice survey of approaches and efforts in biomedical WSD until 2005 including the well known NLM WSD dataset [43], which has 50 ambiguous terms with 5000 test instances. Disambiguation efforts were also focused on a small set of 10–15 ambiguous abbreviations [32, 44] using combinations of supervised and unsupervised approaches. More recent approaches [23, 37] used supervised models including Naive Bayes, SVMs, logistic regressors, decision lists with a variety of features using both subsets of the NLM WSD dataset and other smaller datasets. McInnes and Pedersen [26] use the network structure of the UMLS (specifically the hyperemic trees) and concept definitions to devise concept relatedness measures which are in turn used for WSD for the MSH WSD dataset. Among all the datasets available, the MSH WSD that we use in our current effort is the largest publicly available dataset [16] for biomedical WSD.

In a recent approach [46], Jimeno-Yepes and Berlanga used a hybrid approach that combined a knowledge-based component that exploits the UMLS definitions and synonyms for different concepts with unlabeled biomedical narratives (from the biomedical abstract database Medline/PubMed) to derive word-concept probability estimates $P(w|c)$ for any word w and UMLS concept c. They exploited the Naive Bayes formulation and selected the correct sense as the CUI c that maximizes $P(T|c) = \prod_i P(w_i|c)$, where w_i is the i-th word in the test context T that contains the ambiguous word. With this approach they achieved an accuracy of 89.1% on the MSH WSD dataset [16]. This result corresponds to the best performance thus far on the MSH WSD dataset without using supervised models. In this thesis, we use recent approaches based on neural word embeddings to generate new state-of-the-art results on MSH WSD dataset achieved without supervised cross validation experiments on it. Our methods can be classified as weakly supervised given we employ a well known biomedical concept mapping tool MetaMap [4] to generate concept vectors and employ them in combination with our knowledge-based unsupervised methods [46].

2.2 Neural Embeddings for WSD

Neural word representations have been shown to capture both semantic and syntactic information and a few recent approaches learn word vectors [6, 12, 28] (as elements of $\mathbb{R}^d$, where d is the dimension) in an unsupervised fashion from textual corpora. These dense word vectors obviate the sparsity issues inherent to the so called *one-hot* representations of words that lead to very large dimensionality (typically the size of the vocabulary) resulting in further issues in similarity computations, a phenomenon often termed as the *curse of dimensionality* [7, Chapter 1.4]. Chen et al. [9] adapted the

neural word embedding approach to compute different sense embeddings (of the same word) and showed competitive performance on the SemEval 2007 WSD dataset [31]. Disambiguation is achieved by picking the sense that maximizes the cosine similarity of the corresponding sense vector with the context vector for an ambiguous word. Recently, Iacobacci et al. [15] evaluated and demonstrated the superiority of neural word embeddings as features in supervised WSD models on the same SemEval dataset.

In a very recent effort Pakhomov et al. [33] use word embeddings (without corpus enhanced concept embeddings) for the MSH WSD dataset but only report 77% accuracy. Their approach relies on vectors of words that co-occur with words in the definitions of different senses in the UMLS. In our current effort, we used a similar framework as Chen et al. [9] to directly learn biomedical sense vectors using a pure distributional semantics framework that doesn't rely on word vectors. Additionally, we employed complementary evidences beyond cosine similarity to achieve further improvements that rival performances typically reported using fully supervised approaches.

Chapter 3 NLP and Knowledge-Based Building Blocks

We alluded to most of the basic NLP components used in our methods in Chapter 2. In this chapter, we provide additional background about specific building blocks that are central to our main methods.

3.1 Unified Medical Language System

The UMLS is a large domain expert driven aggregation of nearly 200 biomedical terminologies and standards. It functions as a comprehensive knowledge base and facilitates interoperability between information systems that deal with biomedical terms. It has has three main components: Metathesaurus, Semantic Network, and SPECIALIST lexicon. The Metathesaurus has terms and codes, henceforth called *concepts*, from different terminologies. Biomedical terms from different vocabularies that are deemed synonymous by domain experts are mapped to the same Concept Unique Identifier (CUI) in the Metathesaurus. The semantic network acts as a typing system that is organized as a hierarchy with 133 *semantic types* such as *disease or syndrome, pharmacologic substance,* or *diagnostic procedure*. It also captures 54 important relationships (or relation types) between biomedical entities in the form of a relationship hierarchy with relationships such as *treats, causes,* and *indicates*.

The Metathesaurus currently has about 3.4 million concepts with more than 26 million relations connecting these concepts. Although relations in the Metathesaurus have relation types that are beyond the 54 available through the semantic network, the ones relevant to WSD are high level relation types such as *parent, child, rel_narrow,* and *rel_broad*. The high level relations can be represented as $C1 \rightarrow < rel - type > \rightarrow C2$ where $C1$ and $C2$ are concepts in the UMLS and $< rel - type > \in \{parent, child, rel\_narrow, rel\_broad\}$. The semantic interpretation of these relations (or triples) is that the $C1$ is related to $C2$ via the relation type $< rel - type >$. The *child* (resp. *parent*) relationship means that concept $C1$ has $C2$ as a child (resp. *parent*). The *rel_broad* (resp. *rel_narrow*) type means that $C1$ represents a broader (resp. narrower) concept than $C2$. For example, the concept *hypertensive disease* is a broader concept compared to *systolic hypertension*. These broad and narrow relationships are created by experts to capture those relationships that cannot be captured by the more rigid parent/child relationships in different source vocabularies. Knowledge based methods also exploit paths and their lengths in the UMLS relations graph to resolve word ambiguities. The third component, SPECIALIST lexicon, is useful for lexical processing and variant generation of different biomedical terms.

3.2 NLM's NER and Concept Mapping Program: MetaMap

Named entity recognition (NER) is a well known application of NLP techniques where different entities of interest such as people, locations, and institutions are automati-

cally recognized from mentions in free text. Named entity recognition in biomedical text is difficult because linguistic features that are normally useful (e.g., upper case first letter, prepositions before an entity) in identifying generic named entities are not useful when identifying biomedical named entities, several of which are not proper nouns. Hence, NER systems in biomedicine rely on expert curated lexicons and thesauri. In this work, we use MetaMap [4], a biomedical NER and concept mapping system developed by researchers at the NLM. MetaMap uses a dictionary based approach (using the UMLS concept names as the dictionary) in combination with shallow linguistic parsing (chunking) heuristics for partial match mapping (based on lexical information in the SPECIALIST lexicon) to extract UMLS concepts. MetaMap can process a textual document as a whole but can also generate UMLS concepts from individual noun phrases that are passed as input to it. The latter option is more helpful to identify more specific concepts from longer phrases. MetaMap also identifies negations of concepts and also has a WSD option which is based on concept profiles generated through words co-occurring with different concepts in biomedical literature [45]. We use MetaMap's WSD implementation in our approach to obtain concept vectors which are subsequently used to build superior WSD methods.

3.3 Neural Word Embeddings: Word2Vec

Representing words and documents as vectors has been a long standing approach in information retrieval and computational semantics [41]. Specifically, text corpora can be represented as the so called term-document matrices where each row represents a term (word) and each column represents a document. Each element in such a matrix typically contains the number of times the word corresponding to the row occurs in the document represented by the column. Additional weighting schemes such as the tf-idf heuristic are used instead of the raw frequencies to account for word frequency and informativeness. In this approach, the vector representation of a word is simply the row corresponding to that word. Similarity of two words can be computed by taking cosine similarity or some other metric that compares the corresponding vectors. This approach has two main issues: (1). The size of the vectors can be prohibitive given they are equal to the number of document, which could be very large. (2). The vectors are relatively sparse and do not accurately capture the lexical semantics as expected. To counter the second issue, latent semantic indexing [13] has been introduced and has been an excellent alternative. However, it typically involved expensive singular value decompositions due to which other approaches that also addressed the first issue were introduced. In particular, random indexing [10] has emerged as an alternative offering much less computational burden and is shown to be effective in cases when the corpora are large. Basically, random indexing projects sparse term vectors into dense vectors in a low dimensional space while also roughly preserving relative distances between the vectors. This has been used in many applications such as biomedical knowledge discovery [11] and multi-label classification [17, 19].

In 2013, researchers at Google introduced efficient approaches, release as a software program Word2Vec, that use neural networks to automatically learn low dimensional dense vector representations of words in an unsupervised manner from large

text corpora [27, 28]. In these methods, Mikolov et al. introduce the so called skip-gram model and its improvements to learn word vectors. This model is based on a neural network with three layers: the input layer, a low dimensional projection layer, and an output layer as shown in Figure 3.1. The central idea is to learn to predict neighbor words within a proximity context window of C words based on the current word being input to the neural network. As a sliding context window moves over a corpus of documents, at each position, the word at the center of the window becomes the target word and the words to the left and right of it become the context words to be predicted. The objective is to maximize the average log-likelihood of the corresponding context words given the input target word. Using back propagation, the gradients are used to modify the weight matrices of the neural network including the target word vector elements. Before training begins, the word vectors are typically initialized randomly with uniform selection from a small range $[-\frac{1}{2d}, \frac{1}{2d}]$ where d is the dimensionality of the word vectors [14]. As training proceeds, the word vectors are updated and the resultant vectors at the end are expected to have nice semantic properties. Mikolov et al. give several examples of such properties – they find the vector computed by vec("Madrid") - vec("Spain") + vec("France") is closest to vec("Paris") than any other word where proximity is measured using cosine distance. Although the idea of distributional semantics has been popular for quite some time, the advent of recent deep neural network based unsupervised pre-training as outlined here seems to have revitalized the field of dense vector representations. Full details of the derivations leading to parameter updates and other efficiency considerations are presented by Rong [36].

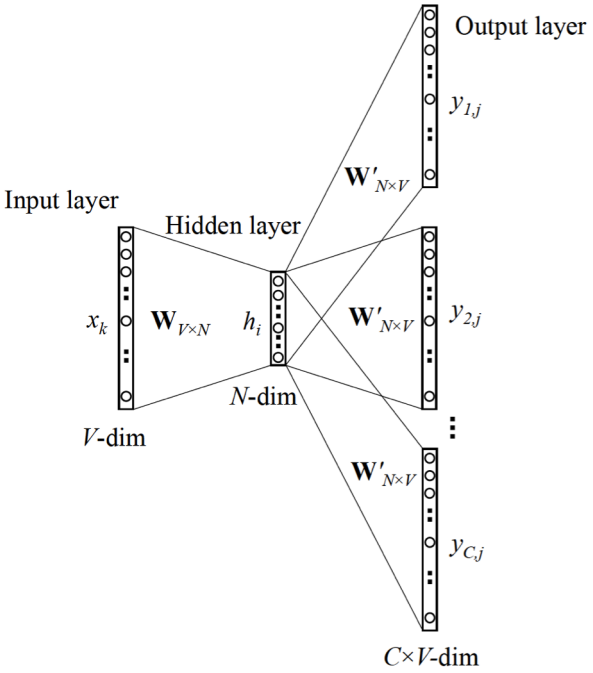

Figure 3.1: Skip-gram word embedding model architecture figure from Rong [36]

3.4 A Knowledge-Based Bayesian WSD Approach

As a component of our methods, we use the knowledge-based approach developed by Jimeno-Yepes and Berlanga [46] which we briefly discussed in Section 2.1. Here we give some additional details. The main idea is to model $P(c|T)$, probability of CUI c given a context T. If this is estimated accurately, our WSD solution is to simply pick the candidate sense c that maximizes $P(c|T)$. Using Bayes theorem, we have

$$P(c|T) = \frac{P(T|c)P(c)}{P(T)} \propto P(T|c) = \prod_{w_j \in T} P(w_j|c),$$

with the naive assumption of independence of tokens w_j that constitute the context T given the sense c. So our solution now depends on estimating the word-concept probabilities $P(w|c)$ for any word w and CUI c. The rest of this section outlines how Jimeno-Yepes and Berlanga accomplish that.

A straightforward first cut to obtain $P(w|c)$ is to simply model it as the MLE estimate

$$P(w|c) = \frac{count(w,c)}{\sum_{w' \in lex(c)} count(w',c)},$$

where $lex(c)$ is the synonymous name set of c in the UMLS. Instead of limiting the search of w to the lexical space of c, they propose to extend it to lexical spaces of concepts that are related to c based on the UMLS relations discussed in Section 3.1. That is, we now have $P_j(w|c_0)$, which denotes the probability of w being selected for the set of concepts $R_k(c_0)$ that are k hops away from the original concept $c_0 = c$. Specifically, they estimate

$$\begin{aligned}
P_k(w|c_0) &= \sum_{c_k \in R_k(c_0)} P_k(w, c_k, c_{k-1}, \ldots | c_0) \\
&= \frac{\sum_{c_k \in R_k(c_0)} P_k(w, c_k, c_{k-1}, \ldots, c_0)}{P(c_0)} \\
&= \frac{\sum_{c_k \in R_k(c_0)} P_0(w|c_k) \prod_{l=0,\ldots,k-1} P(c_{l+1}|c_l) P(c_0)}{P(c_0)} \\
&= \sum_{c_k \in R_k(c_0)} P_0(w|c_k) \prod_{l=0,\ldots,k-1} P(c_{l+1}|c_l),
\end{aligned}$$

where Markov assumption is used for estimating $P_k(w, c_k, \ldots, c_0)$ in terms of traversal probabilities, $P(c_{l+1}|c_l)$, of hopping from concept c_l to c_{l+1} in the UMLS relation graph. This is mathematically estimated as

$$P(c_{l+1}|c_l) = \frac{|r(c_{l+1}, c_l)|}{|r(*, c_l)|},$$

with $r(c_1, c_2)$ denoting the number of UMLS relations connecting c_1 and c_2 and the denominator indicating the number of relations where c_l participates.

The word concept probabilities $P_j(w|c_0)$ obtained at different values of $j = 0, \ldots, l$ are finally combined using coefficients are determined using

$$P(w|c_0) = \sum_{j=0,\ldots,l} \alpha_j P_j(w|c) \quad where \quad \alpha_0, \ldots, \alpha_l > 0 \ \ and \ \sum_{j=0,\ldots,l} \alpha_j = 1.$$

They start with each $\alpha_j = 1/l$ with l being the number of hops considered and update them using expectation-maximization, details of which are presented in their paper [46, Section 3.3].

Chapter 4 Dataset and Methods

There are 203 ambiguous terms in the MSH WSD dataset [16] with a total of 424 unique concept unique identifiers (CUIs) from the unified medical language system (UMLS [29]), each of which is a unique sense. Thus, on average, the dataset has 424/203 = 2.08 senses. There are a total of 38,495 test instances of contexts (a few sentences) each with one of the 203 ambiguous terms along with the correct sense (CUI). Besides being the largest biomedical WSD dataset, this also includes a richer set of ambiguities including 106 ambiguous abbreviations, 88 ambiguous noun phrases, and 9 that are combinations of both. Due to these features, the NLM encourages researchers to use this latest dataset over their older dataset (please see `https://wsd.nlm.nih.gov`). Our goal is to directly test on this dataset by employing distantly supervised or unsupervised approaches. To this end we learn vector representations of words and CUIs using well known approaches that apply deep neural networks to NLP tasks.

4.1 Neural Word and Concept Embeddings

We ran the Word2Vec [28] word embedding program (the skip-gram model) from Google on over 20 million biomedical citations (titles and abstracts) from PubMed to obtain word vector representations with a word window size of ten words and dimensionality $d = 300$ with all other parameters set to the default settings. To learn concept or CUI vectors of the same dimensionality, we curated a dataset of five million randomly chosen subset of nearly five million citations (published between 1998 and 2014). For this subset of PubMed, we ran MetaMap [4], a well known NER and concept mapping program from the NLM, with its WSD option turned on so we obtain unique CUIs for potential ambiguous terms. The text was passed through MetaMap two adjacent non-stop words at a time, to capture as many CUIs as possible. Next, we treated these sequences of CUIs in each citation thus obtained through NER as a semantic version of the free text corpus. We ran word2vec on this corpus of CUI texts, just like how we ran it on free text articles with the same parameters. As a result we obtained 300 dimensional word vectors for each CUI, including all 424 CUIs corresponding to the 203 ambiguous terms in our test dataset.

This component of our methodology to derive dense concept vectors involves distant supervision because although MetaMap with its WSD option is in and of itself not a powerful solution (see Chapter 5), it nevertheless was useful to learn concept vectors that in turn helped us achieve state-of-the-art results. This deep neural network based distributional semantics approach to learning CUI vectors aids in modeling complementary aspects of similarity given we use, as a component, the CUI definition based information via our earlier word-probability estimate based approach [46].

4.2 WSD with Concept Embeddings and Knowledge-Based Approaches

Our main idea is that besides comparing pairs of word vectors and concept vectors, we can also compare a word vector with a concept vector given at a high level there is a direct connection between words and concepts – words are often lexical manifestations of high level concepts. The fact that we simply replaced word sequences in free text with the corresponding concept sequences to generate CUI vectors of the same dimensionality as the word vectors also makes it feasible to compare word vectors and their compositions to concept vectors. As we show in Chapter 5, this intuition appears to work as well as other state-of-the-art approaches [46].

We establish some notation for the rest of the paper. In any WSD problem, a test instance corresponds to a three tuple $(T, w, C(w))$ where T is a context, typically a few sentences, that contains the ambiguous term w and $C(w)$ is the set of different senses that w can assume depending upon the context T[1]. Specifically, $C(w)$ in this thesis is the set of different CUIs that capture the different senses for w. Our WSD goal is to construct a function $f(T, w, C(w))$ that maps T to the CUI $c \in C(w)$ that corresponds to the correct sense in which w was used in T. We have four approaches that apply the embeddings from Section 4.1 to our test set. We specify them in terms of functions $f^{?}(T, w, C(w))$ where ? indicates symbols that identify the underlying method(s) used made clear as follows.

1. Our first approach is based on vector cosine similarity with

$$f^{c}(T, w, C(w)) = \arg\max_{c \in C(w)} \cos(\vec{T}_{avg}, \vec{c}),$$

   where $\vec{T}_{avg}$ is the simple average of non-stop words' vectors in the context T and $\vec{c}$ is the context vector for c.

2. Our second approach is based on vector projections with

$$f^{p}(T, w, C(w)) = \arg\max_{c \in C(w)} \left[\rho[\cos(\vec{T}_{avg}, \vec{c})] \cdot \frac{\|\mathcal{P}(\vec{T}_{avg}, \vec{c})\|}{\|\vec{c}\|} \right],$$

   where $\mathcal{P}(\vec{r}, \vec{s})$ refers to the projection of $\vec{r}$ on to $\vec{s}$, $\| \ \|$ is the Euclidean norm, and ρ is the sign function. Using straightforward manipulation based on vector projections in Euclidean spaces [22, Chapter 5], we have

$$\|\mathcal{P}(\vec{T}_{avg}, \vec{c})\| = \frac{|\vec{T}_{avg} \bullet \vec{c}|}{\|\vec{c}\|},$$

   which is used in our implementation (with $\bullet$ denoting vector dot product).

[1] In practice, there might be cases where the context in T is deemed insufficient even for human judges to pick the right sense. However, for this manuscript we assess our performance based on MSH WSD dataset where each instance is assigned a unique sense.

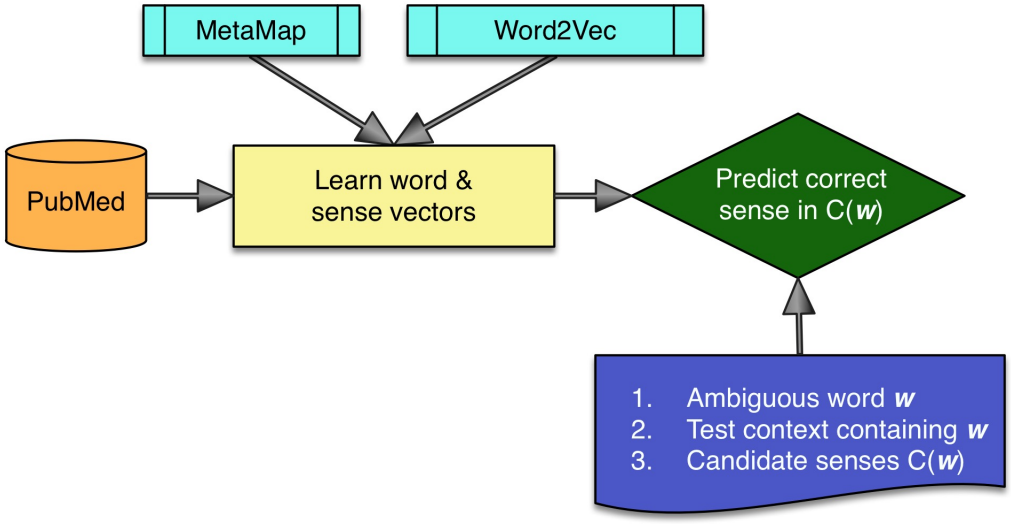

Figure 4.1: Architecture for WSD approaches from Sections 4.1 and 4.2

3. Our third approach is based on the first two approaches where we set

$$f^{c,p}(T, w, C(w)) = \arg\max_{c \in C(w)} \left[\cos(\vec{T}_{avg}, \vec{c}) \cdot \frac{\|\mathcal{P}(\vec{T}_{avg}, \vec{c})\|}{\|\vec{c}\|} \right].$$

We simply incorporate both evidences (magnitude and orientation of association) to compare different CUIs.

4. Our final approach uses a probabilistic model Jimeno-Yepes and Berlanga developed in an earlier effort [46], which as outlined in Section 2.1, selects the c that maximizes $P(T|c)$. We involve this knowledge based approach as a third scoring component and set

$$f^{c,p,k}(T, w, C(w)) = \arg\max_{c \in C(w)} \left[\cos(\vec{T}_{avg}, \vec{c}) \cdot \frac{\|\mathcal{P}(\vec{T}_{avg}, \vec{c})\|}{\|\vec{c}\|} + P(T|c) \right].$$

The approach in f^c is well known given cosine similarity is a popular approach to measure semantic similarity of entities (words, concepts, ...) represented by the corresponding vectors. Although f^c accounts for the overall directional similarity (thematic orientation) of the vectors, it does not account for the strength or magnitude of association, an aspect that seems ignored in others' efforts we reviewed for this paper. By considering the vector projection of the context vector onto the CUI vector $\vec{c}$, in f^p we also account for the magnitude of the context vector's projection in relation to that of the CUI vector. The sign function ρ is essentially to account for situations when $90 < \theta \leq 180$, the angle between $\vec{T}_{avg}$ and $\vec{c}$. The methods discussed thus far can be summarized using the schematic in Figure 4.1.

4.3 WSD with Distant Supervision

From methods in Section 4.2, we have multiple ways of disambiguating CUIs for any ambiguous term given a sample context. We exploit them to build a distantly supervised dataset for the 203 ambiguous terms in our test dataset. For each sentence in an independent corpus of biomedical citations that contains any ambiguous term from our dataset, we employ methods in Section 4.2 to assign the predicted correct CUI. Thus we can create a distantly supervised dataset for each ambiguous term with thousands of examples if we choose a large corpus. These examples can then be used to train traditional discriminative models or nearest neighbors models. We emphasize here that we are proposing to label arbitrary sentences (not our test sentences) in an external corpus based on our methods in Section 4.2. Hence we still have our full MSH WSD dataset to finally test the approach we propose here with other models in a fair way.

For the k nearest neighbor (k-NN) model, let $\mathcal{D}^w \subseteq \mathcal{D}$ be the set of instances for the ambiguous term w in the distantly supervised dataset $\mathcal{D}$. We rank instances $(D, w, c) \in \mathcal{D}^w$ for a given test instance T based on $\cos(\vec{T}_{avg}, \vec{D}_{avg})$, where c is the sense assigned to D from $C(w)$ based on methods in Section 4.2. Let $R_k(\mathcal{D}^w)$ be the set of top k instances in $\mathcal{D}^w$ when ranked in descending order based on $\cos(\vec{T}_{avg}, \vec{D}_{avg})$. Now the correct sense for T is assigned based on

$$f^{k-NN}(T, w, C(w)) = \arg\max_{c \in C(w)} \left[\sum_{(D,w,c) \in R_k(\mathcal{D}^w)} \cos(\vec{T}_{avg}, \vec{D}_{avg}) \right].$$

The expression in the arg max boils down to summing up the similarities of the test context with those contexts in the training dataset that have the same assigned CUI c. We subsequently pick the particular c that maximizes that summation. Intuitively, our approach aggregates evidence from training instances that are semantically most similar to our test instance. The choice of k also plays an important role in performance of k-NN approaches as we note in the next section.

Chapter 5 Results and Discussion

Our results are shown in Table 5.1 based on methods introduced in the previous section. MetaMap doesn't perform as well on this dataset (row 1) even with the WSD option achieving an accuracy of 81.77%. However, it may not be fair to compare MetaMap with our methods given it doesn't try to particularly disambiguate our specific 203 terms, for each of which we are already given candidate concepts that contain the correct sense. In row 2 of the table, we show the performance achieved by our prior work using word-concept probability estimates $P(w|c)$ derived from synonymous names of concepts in the UMLS Metathesaurus.

Table 5.1: Performance on MSH WSD Dataset

Method	Accuracy
MetaMap with WSD	81.77%
Jimeno-Yepes and Berlanga [46]	89.10%
Cosine similarity (f^c)	85.54%
Projection length proportion (f^p)	88.68%
Combining f^c and f^p ($f^{c,p}$)	89.26%
Combining f^c, f^p, and [46] ($f^{c,p,k}$)	92.24%
Convolutional neural networks	87.78%
k-NN with $k = 3500$ (f^{k-NN})	94.34%

Rows 3–6 show performances of methods we introduced based on neural word and concept representations in Section 4.2. The cosine similarity and projection approaches both score above 85% but when used together, they achieve an accuracy of 89.26% which is slightly better than the current best result [46] achieved through unsupervised approaches. Row 6 shows an accuracy of 92.24% achieved by our ensemble method that combines our word/concept vector approach with the knowledge based Bayesian approach [46]. The test time complexity of these methods is linear in terms of the number of words in the test context T and the number of senses $|C(w)|$ for the ambiguous term w considering the computation of $\vec{T}_{avg}$ and evaluation of the $\arg\max$ expressions for each $c \in C(w)$.

We created a distantly supervised dataset as outlined in Section 4.3 with the same corpus of five million biomedical citations used for training word and concept vectors (Section 4.1). From this corpus, we considered the so called utterances that represent clauses (from the input text) that MetaMap outputs as distinct fragments with the corresponding CUIs. For each utterance that contains an ambiguous term in our test

set, we apply our best linear method $f^{c,p,k}$ (corresponding to row 6 of Table 5.1) to assign one specific CUI from all possible candidates. There were seven million such utterances, with an average length of 18 words, that contained an ambiguous term out of a total of 78 million utterances from the corpus. Given our prior experiences in convolutional neural networks (CNNs) in biomedical text classification [35] that proved superior over traditional linear classifiers such as support vector machines and logistic regression models, we built 203 multi class CNN models, one for each ambiguous term based on this distantly supervised dataset. The configuration of the CNN and its various hyper parameters were determined as per our prior effort [35, Sections 3.2 and 4.2]. This setup however resulted in accuracy of 87.78% which doesn't match the performance of simpler approaches (rows 4–6 of Table 5.1).

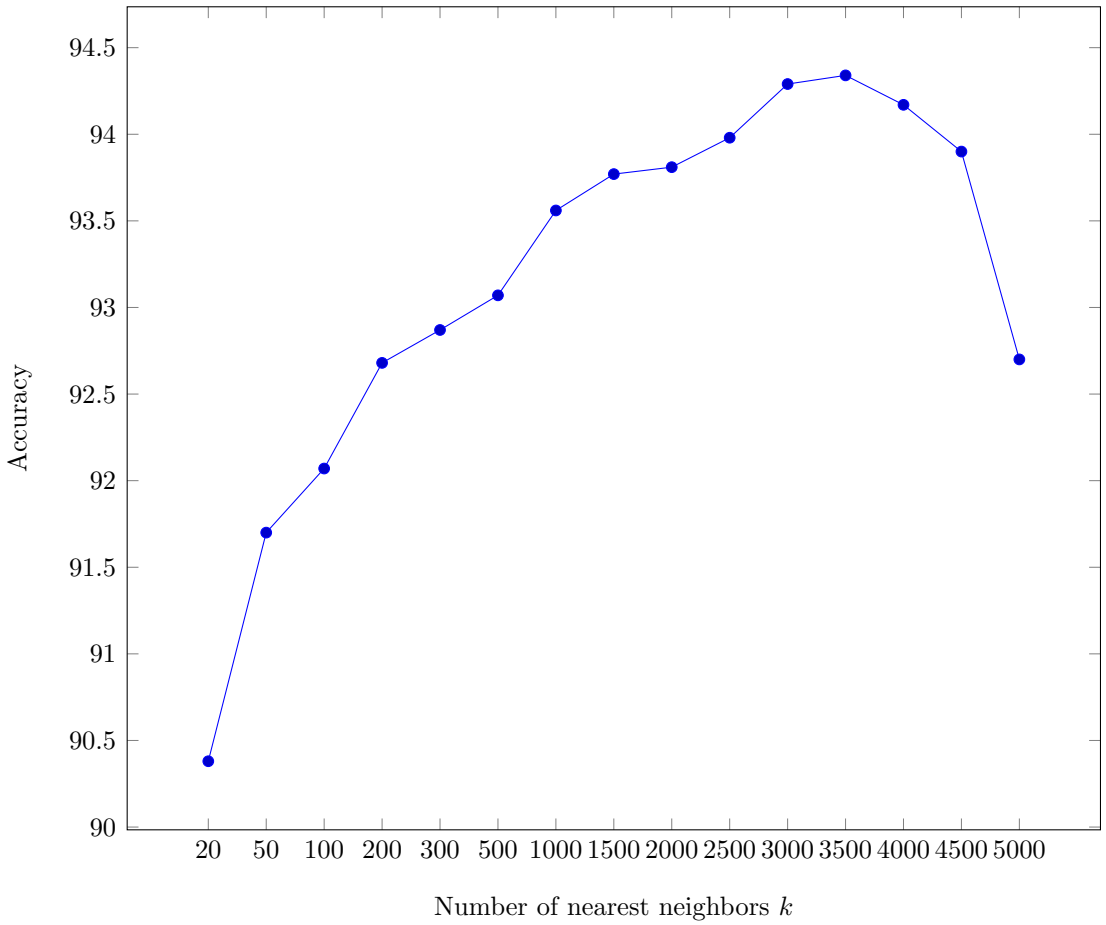

Figure 5.1: Accuracy of the k-NN approach with varying k

We finally applied the k-NN approach outlined in Section 4.3 with the distantly supervised dataset with the number of nearest neighbors k $\in \{20, 50, 100, 200, 300, 500, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}$. The corresponding accuracies are plotted as shown in Figure 5.1. We obtained the best accuracy of 94.34% when $k = 3500$ as shown in the last row of Table 5.1. Overall, the accuracy rapidly in-

creases as the numbers of neighbors used increase. The gains become smaller as more neighbors are added, reaching the top score at $k = 3500$ after which the accuracy descends abruptly. At $k = 5000$, the accuracy is same as that achieved with $k = 300$. This phenomenon is not surprising – at first more neighbors contribute to additional evidence, consistency, and robustness against noise in comparing the candidate concepts. However, considering an increasing number of neighbors at some point also leads to the semantic drift of their content from that of the test context. So neighbors ranked further down the list negatively affect the prediction given they are not as related to the test context, thus lowering overall accuracy. We realize that the value of $k = 3500$ is specific to this biomedical dataset and that there could be a value $3500 < k < 4000$ that achieves a slightly higher accuracy. Our analysis is essentially a proof of concept for the high level monotonous nature of performance of k-NN based approaches. Given that there are over 38,000 test instances in our dataset, we believe $k \approx 3500$ is appropriate in domains with similar characteristics (e.g., average number of senses per word, distributions of senses, and average length of test contexts). However, researchers may be able to derive more appropriate k values for their domains if they have access to relevant datasets.

Finally, it is well known that k-NN approaches are infamous for high test time complexity because of the nearest neighbor search in high dimensional space. Our implementation involves cosine similarity computation with all training instances for the corresponding ambiguous term. In this effort, on average there are nearly 40,000 training instances created through distant supervision per ambiguous term. So given a new test instance $(T, w, C(w))$, cosine similarity (of 300 dimensional vectors) needs to be computed for the test instance T with about 40,000 contexts to impose the ranking on these potential neighbors. The threshold of a chosen k (say, 3500) can only be applied after this ranking is created. However, this similarity computation can be parallelized in a straightforward manner by distributing the similarity computations across multiple processors and pooling the results to incrementally build the ranked neighbor list. Although real time disambiguation may not be feasible, having the k-NN models run overnight every day to address disambiguation in new articles may be practical. Alternative approaches such as locality sensitive hashing [39] that address the dimensionality problems without having to compute cosine similarities may be helpful to alleviate the situation. Overall, however, it is clear that k-NN based approaches with distantly supervised datasets offer an interesting alternative to purely supervised approaches in biomedical WSD.

Chapter 6 Conclusion

Biomedical WSD is an important initial task with implications for downstream components in NLP applications. In this effort, we applied recent approaches in neural word embeddings to construct concept embeddings. Our linear time method uses these embeddings to combine cosine similarity, projection magnitude proportion, and a prior knowledge based approach to produce an accuracy of 92.24%. This is an absolute 3% improvement over just using the knowledge based approach, which is the previous best result obtained without supervised learning. Based on predictions from our best linear method, we created a new distantly supervised dataset and built a k-NN model that achieves an accuracy of 94.34%. Our results rival performances achieved by supervised approaches – the best published supervised result achieves 93% macro accuracy over ten fold cross validation experiments on the MSH WSD dataset with the Naive Bayes model [16]. Overall, our results in this paper contribute new evidence that dense neural embeddings function as powerful representations of textual data for biomedical NLP applications. Furthermore, they also showcase the potential of knowledge-based approaches in learning better neural dense vector representations and their complementary contributions to WSD tasks.

Acknowledgment

This work is supported by the NIH National Library of Medicine through grant 1R21LM012274, and the Kentucky Lung Cancer Research Program through grant PO2 41514000040001. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Bibliography

[1] Eneko Agirre and Philip Edmonds. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media, 2007.

[2] Marianna Apidianaki. Data-driven semantic analysis for multilingual wsd and lexical selection in translation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 77–85. Association for Computational Linguistics, 2009.

[3] Marianna Apidianaki and Yifan He. An algorithm for cross-lingual sense-clustering tested in a mt evaluation setting. In *International Workshop on Spoken Language Translation (IWSLT-2010)*, pages 219–226, 2010.

[4] Alan R. Aronson and François-Michel Lang. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[5] Gokhan Bakal and Ramakanth Kavuluru. Predicting treatment relations with semantic patterns over biomedical knowledge graphs. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 586–596. Springer, 2015.

[6] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

[7] Christopher M Bishop. *Pattern recognition and machine learning*. Springer, 2006.

[8] Delroy Cameron, Ramakanth Kavuluru, Thomas C Rindflesch, Amit P Sheth, Krishnaprasad Thirunarayan, and Olivier Bodenreider. Context-driven automatic subgraph creation for literature-based discovery. *Journal of biomedical informatics*, 54:141–157, 2015.

[9] Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conf. on Empirical Methods in Natural Language Processing*, pages 1025–1035. ACL, 2014.

[10] Trevor Cohen, Roger Schvaneveldt, and Dominic Widdows. Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, 43(2):240–256, 2010.

[11] Trevor Cohen, Dominic Widdows, Roger W Schvaneveldt, Peter Davies, and Thomas C Rindflesch. Discovering discovery patterns with predication-based semantic indexing. *Journal of biomedical informatics*, 45(6):1049–1065, 2012.

[12] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[13] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391, 1990.

[14] Yoav Goldberg. A primer on neural network models for natural language processing. *J. Artif. Intell. Res. (JAIR)*, 57:345–420, 2016.

[15] Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Asso. for Computational Linguistics*, pages 897–907, 2016.

[16] Antonio J Jimeno-Yepes, Bridget T McInnes, and Alan R Aronson. Exploiting mesh indexing in medline to generate a data set for word sense disambiguation. *BMC bioinformatics*, 12(223), 2011.

[17] Ramakanth Kavuluru and Yuan Lu. Leveraging output term co-occurrence frequencies and latent associations in predicting medical subject headings. *Data & Knowledge Engineering*, 94(Part B):189–201, 2014.

[18] Ramakanth Kavuluru and Anthony Rios. Automatic assignment of non-leaf mesh terms to biomedical articles. In *AMIA Annual Symposium Proceedings*, volume 2015, page 697. American Medical Informatics Association, 2015.

[19] Ramakanth Kavuluru, Anthony Rios, and Yuan Lu. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial intelligence in medicine*, 65(2):155–166, 2015.

[20] Ramakanth Kavuluru, Christopher Thomas, Amit P Sheth, Victor Chan, Wenbo Wang, Alan Smith, Armando Soto, and Amy Walters. An up-to-date knowledge-based literature search and exploration framework for focused bioscience domains. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pages 275–284. ACM, 2012.

[21] Seonho Kim and Juntae Yoon. Link-topic model for biomedical abbreviation disambiguation. *Journal of biomedical informatics*, 53:367–380, 2015.

[22] Ron Larson and David C Falvo. *Elementary Linear Algebra*. Houghton Mifflin Harcourt Publishing Company, 2008.

[23] Hongfang Liu, Virginia Teller, and Carol Friedman. A multi-aspect comparison study of supervised word sense disambiguation. *Journal of the American Medical Informatics Association*, 11(4):320–331, 2004.

[24] Yuan Luo, Özlem Uzuner, and Peter Szolovits. Bridging semantics and syntax with graph algorithm – state-of-the-art of extracting biomedical relations. *Briefings in bioinformatics*, page bbw001, 2016.

[25] David Martinez, Oier Lopez De Lacalle, and Eneko Agirre. On the use of automatically acquired examples for all-nouns word sense disambiguation. *J. Artif. Intell. Res.(JAIR)*, 33:79–107, 2008.

[26] Bridget T McInnes and Ted Pedersen. Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of biomedical informatics*, 46(6):1116–1124, 2013.

[27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[28] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[29] National Library of Medicine. Unified Medical Language System Reference Manual. `http://www.ncbi.nlm.nih.gov/books/NBK9676/`.

[30] Roberto Navigli. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10, 2009.

[31] Roberto Navigli, Kenneth C Litkowski, and Orin Hargraves. Semeval-2007 task 07: Coarse-grained english all-words task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 30–35. Association for Computational Linguistics, 2007.

[32] Serguei Pakhomov, Ted Pedersen, and Christopher G Chute. Abbreviation and acronym disambiguation in clinical discourse. In *AMIA Annual Symposium Proceedings*, volume 2005, page 589, 2005.

[33] Serguei VS Pakhomov, Greg Finley, Reed McEwan, Yan Wang, and Genevieve B Melton. Corpus domain effects on distributional semantic modeling of medical terms. *Bioinformatics*, page In Press, 2016.

[34] Anthony Rios and Ramakanth Kavuluru. Analyzing the moving parts of a large-scale multi-label text classification pipeline: Experiences in indexing biomedical articles. In *Healthcare Informatics (ICHI), 2015 International Conference on*, pages 1–7. IEEE, 2015.

[35] Anthony Rios and Ramakanth Kavuluru. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 258–267. ACM, 2015.

[36] Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[37] Guergana K Savova, Anni R Coden, Igor L Sominsky, Rie Johnson, Philip V Ogren, Piet C De Groen, and Christopher G Chute. Word sense disambiguation across two domains: biomedical literature and clinical notes. *Journal of biomedical informatics*, 41(6):1088–1100, 2008.

[38] Martijn J Schuemie, Jan A Kors, and Barend Mons. Word sense disambiguation in the biomedical domain: an overview. *Journal of Computational Biology*, 12(5):554–565, 2005.

[39] Malcolm Slaney and Michael Casey. Locality-sensitive hashing for finding nearest neighbors [lecture notes]. *IEEE Signal Processing Magazine*, 25(2):128–131, 2008.

[40] Mark Stevenson, Yikun Guo, Robert Gaizauskas, and David Martinez. Disambiguation of biomedical text using diverse sources of information. *BMC bioinformatics*, 9(11), 2008.

[41] Peter D Turney, Patrick Pantel, et al. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188, 2010.

[42] Jing Wang, Mohit Bansal, Kevin Gimpel, Brian D Ziebart, and T Yu Clement. A sense-topic model for word sense induction with unsupervised data enrichment. *Transactions of the Association for Computational Linguistics*, 3:59–71, 2015.

[43] Marc Weeber, James G Mork, and Alan R Aronson. Developing a test collection for biomedical word sense disambiguation. In *Proceedings of the AMIA Symposium*, page 746. American Medical Informatics Association, 2001.

[44] Hua Xu, Peter D Stetson, and Carol Friedman. Combining corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1004. American Medical Informatics Association, 2012.

[45] Antonio Jimeno Yepes and Alan R Aronson. Integration of UMLS and Medline in unsupervised word sense disambiguation. In *2012 AAAI fall symposium series*, pages 26–31, 2012.

[46] Antonio Jimeno Yepes and Rafael Berlanga. Knowledge based word-concept model estimation and refinement for biomedical text mining. *Journal of biomedical informatics*, 53:300–307, 2015.

[47] Zhi Zhong and Hwee Tou Ng. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83. Association for Computational Linguistics, 2010.

Vita

AKM Sabbir received his bachelor in computer science and engineering from Khulna University of Engineering and Technology(Khulna, Bangladesh) in 2008. Right after graduation, he joined GrameenPhone Ltd the largest cellphone network provider in Bangladesh as a system engineering in Intelligent Network department. Subsequently, he worked in Accenture, Samsung as a software engineer before he started his graduate education in Computer Science department of University of Kentucky in 2013. He worked as Research Assistant in Bio-statistics department and Merkey Cancer Center under college of medicine.

Publication

Dr. Ramakanth Kavuluru, AKM Sabbir. Toward Automated E-cigarette Surveillance: Spotting E-cigarette Proponents on Twitter. In *Journal of Biomedical Informatics.*

AKM Sabbir, Dr. Antonio Jimino Yepes, Dr. Ramakanth Kavuluru. Knowledge-Based Biomedical Word Sense Disambiguation with Neural Concept Embeddings and Distant Supervision. In proceedings of *Journal of the American Medical Informatics Association.*

AKM Sabbir, Dr. Sally Ellingson. Side EffectTerm Matching for Computational Adverse Drug Reaction Predictions.In *BMC Bioinformatics 2016.*