BIOMEDICAL SEMANTIC EMBEDDINGS: USING HYBRID SENTENCES TO CONSTRUCT BIOMEDICAL WORD EMBEDDINGS AND ITS APPLICATIONS

Arshad Shaik

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

December 2019

APPROVED:

Wei Jin, Major Professor
Xuan Guo, Committee Member
Bill Buckles, Committee Member
Barrett Bryant, Chair of the Department of Computer Science and Engineering
Hanchen Huang, Dean of the College of Engineering
Victor Prybutok, Dean of the Toulouse Graduate School

Shaik, Arshad. *Biomedical Semantic Embeddings: Using Hybrid Sentences to Construct Biomedical Word Embeddings and Its Applications*. Master of Science (Computer Science), December 2019, 50 pp., 8 tables, 4 figures, 62 numbered references.

Word embeddings is a useful method that has shown enormous success in various NLP tasks, not only in open domain but also in biomedical domain. The biomedical domain provides various domain specific resources and tools that can be exploited to improve performance of these word embeddings. However, most of the research related to word embeddings in biomedical domain focuses on analysis of model architecture, hyper-parameters and input text. In this paper, we use SemMedDB to design new sentences called `Semantic Sentences'. Then we use these sentences in addition to biomedical text as inputs to the word embedding model. This approach aims at introducing biomedical semantic types defined by UMLS, into the vector space of word embeddings. The semantically rich word embeddings presented here rivals state of the art biomedical word embedding in both semantic similarity and relatedness metrics up to 11%. We also demonstrate how these semantic types in word embeddings can be utilized.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

This chapter is dedicated to familiarizing the research related to word embeddings in open domain, how it has been incorporated in the biomedical domain, and how different biomedical tools have been utilized in previous research work. This section will provide information on the following topics below:

- Word Embeddings(Open Domain)
- Biomedical Word Embeddings
- Biomedical Resources
- Unifying biomedical tools and word embeddings

1.1. Word Embeddings(Open Domain)

Word embeddings is an alternative name for feature learning techniques in which words and phrases are mapped as multidimensional one-hot encoded vectors. Word embeddings are a sub-category of distributional semantics in linguistics. In large sets of linguistic language data, word embeddings quantify and classify semantics similarities between linguistic items based on their distributions. This enables the model to capture the context, relation, semantic and syntactic similarities with other units in the dataset for effective predictions. Lavelli et al. [34] studied two different forms of word embedding, one in which words are represented as vectors of co-occurring words and second in which words are represented as vectors of linguistic contexts in which words occur in the dataset. Neural Networks are the most popular and efficient approach to map words to vectors but dimensionality reduction, probabilistic models, explainable knowledge base methods can be employed to obtain similar results.

Word embeddings have been the focus of research since the early 1990s. However, more noticeable results were achieved after 2010 partly due to the advancement in computational speed and neural networks and partly because of the advances made on the training speed of the model and quality of vectors. Bengio et al. [8] provided a neural probabilistic

language model which was able to learn a distributed representation for words tackling the curse of dimensionality. Generalization is achieved easily for continuous variables but for discrete spaces, the structure is difficult to obtain as a change in a variable has a drastic impact on the estimated value. N-gram models that obtained generalization by concatenating short overlapping sequences in the training set were fairly successful before the boom in Neural networks. Collobert et al. [14] trained a deep neural network with semi-supervised and multitask learning for various NLP tasks such as prediction of part-of-speech tags, named entity tags, chunks, semantic roles, etc. Researchers have opted for divide and conquer approach and divided the implementation process into several subtasks improving the overall efficiency one step at a time. Joseph et al. [58] took previously trained word embeddings and provided a comparison of these embeddings when used as input features for NLP tasks such as named entity recognition and chunking. The result was that learning word features using unsupervised learning and integrating them into an existing supervised NLP model produces less accuracy compared to a semi-supervised model that learn supervised and unsupervised tasks simultaneously. Socher et al. [54] introduced a recursive neural network (RNN) architecture based on context-sensitive recursive neural networks for parsing natural language and learning vector space representations for variable sized inputs. The networks induce distributed feature representations for unseen phrases and provide syntactic as well as semantic information. Mikolov et al. [39] proposed continuous bag-of-words (CBOW) and skip-gram architectures for computing the word vectors from very large data sets. These model architectures provided a significant improvement in accuracy when tested for syntactic and semantic word similarities with much lower computational cost, both these models still stand as state of the art for training word embeddings. Word2Vec, the product of these models remains the most widely used algorithm to produce word embeddings. Recently, a surge of algorithms has been proposed in the literature which utilized co-occurrence information and matrix factorization for learning distributed representation of words [36], [47], [45], [49]. Pennington et al. [47] introduced GloVe a global log-bilinear regression model for unsupervised learning of word representations. They argued that the two main approaches for learning distributional

word representations, the count-based method, and the prediction based method, both produce similar results as they probe the underlying co-occurrence statistics of the corpus but the count-based method provides better efficiency for global statistics used in GloVe. Although algorithms have different underlying architectures, the resulting embeddings usually give similar performance when leveraged for various NLP tasks.

The introduction of CBOW and skip-gram architectures led to an exponential growth of the applications of word embeddings. Milkov et al. [40] researched several implementations to improve the training speed of skip-gram architecture and obtained speedup ranging from 100 percent to 900 percent on several datasets by sub-sampling similar linguistic terms and replacing the hierarchical softmax by Negative sampling. Negative sampling algorithm learns accurate representations for frequent words while subsampling resulted in better representation of uncommon words. Milkov et al. [41] researched vector space word representations and found that these are capable of capturing syntactic and semantic similarities in language. The study also demonstrated the ability of an RNN to encode similarities between pair of words using vector arithmetic, which were termed as linguistic similarities. Levy et al. [35] improved on this model and showed that analogous to the neural embedding space, the explicit vector space encodes a vast amount of relational similarity that can be recovered similarly. This implies that the novel approach in neural word embeddings is not to discover patterns but to preserve the patterns in the word-context co-occurrence matrix. A key insight to understanding the approach is divide and conquer, by decomposing vector arithmetic into a linear combination of three pairwise similarities. Using a modified Optimization Objective for the pairwise similarities resulted in significant improvement in performance. The study revealed that finding analogies through vector arithmetic under certain conditions performs equally as neural word embeddings. Word embedding requires large training sets to predict accurately however embeddings for domain-based datasets can be obtained using Domain Adaptation. The drawback of open domain pre-trained embedding is that the corpora used for training have a significant influence on resulting word representations, Hence it cannot be used for domain-specific tasks.

1.2. Biomedical Word Embeddings

The rapid increase in scientific biomedical literature makes it difficult, even for domain experts, to keep the current knowledge updated. Web engines and information retrieval (IR) models have made significant advances but fall behind in making accurate predictions resulting in many misinterpretations by the general users and these are major causes of conditions such as cyberchondria. The need to make biomedical word embeddings a key research topic have increased sharply over the years.

The publicly available biomedical literature contains billions of words in abstracts and texts waiting to be utilized in statistical models, which are inputs for text classification. Similarly, there has been some research for the creation of word embeddings in the biomedical domain. Stenetorp et al. [55] were the first to present an analysis of various word representations in biomedical domain NLP, demonstrating substantial benefit from word representations trained on in-domain texts compared to out-of-domain texts for entity recognition and classification tasks. The manually annotated corpora are dwarfed by unannotated citations present in large scale databases such as PubMed literature database. Supervised learning algorithms are successful, however, they fail to incorporate the large raw data available and thus researchers opt for semi-supervised or support supervised approaches. Stenetorp et al. [55] analyzed support supervised methods extrinsically, by studying the capacity of in-duced representations to support machine learning-based natural language processing tasks, specifically named entity recognition on three different corpora and semantic category disambiguation on a large automatically acquired corpus.

Pyysalo et al. [43] provided distributional semantic resources for biomedical text processing and produced the word representations induced from the entire biomedical literature. They utilized these resources to preprocess openly available biomedical literature, i.e. PubMed and PMC OA, and finally apply word2vec to train word embeddings from these preprocessed texts. These word embeddings are openly available and have been used as features for various BioNLP studies. This produced a spark in models of meaning with unannotated text. Chiu et al. [12] prepared a study on how to train good word embeddings

for biomedical NLP, and their research provided a comparison on how the quality of word embeddings differed based on pre-processing of text, model selection(skip-grams vs CBOW), hyper-parameters(sampling, min-count, learning rate, vector dimension, and context window size) and corpus selection(PMC, PubMed). The research also concluded that the skip-gram model produces better results than the Global Vector model (GloVe) on word similarity task and the performance of skip-gram is improved by the usage of higher dimensional vectors. It is observed that larger corpus does not necessarily lead to better performance in biomedical NLP and that optimization of hyperparameters boosts the performance of vectors. In conclusion, most of the word embeddings related research in the biomedical domain is restricted to the utilization of openly available biomedical text and analysis studies. Muneeb et al. [56] made a comparative study of the latest state of the art architectures, GloVe and Word2Vec, on a task to output semantic similarity and relatedness between biomedical texts utilizing a corpus of size greater than one million clinical research articles. The performance of the models varied with different hyperparameters but the study concluded with Word2Vec model performing better and capturing more lexico-semantic similarities than others. Jiang et al [27] proposed a domain-specific biomedical word embedding model that chunk and entity information present in the corpus. The results showed that domain-specific word embeddings outperformed other open domain word embedding models. De Vine et al. [16] compared Skip-Gram to other benchmark approaches and the results displayed that Skip-Gram performs better in capturing the semantic similarities of the biomedical texts in the corpus used and co-relates closely with expert human assessors.

1.3. Biomedical Resources

The popularly used text data in training word embeddings provides considerable accuracy on NLP tasks but lacks the utilization of domain-specific resources to increase the accuracy of the existing models. Furthermore, the biomedical domain is rich in semantic resources that have been used to improve various tasks such as text classification, named entity recognition, information retrieval, Question Answering systems. These biomedical tools were developed with the advancement in NLP in the biomedical field and have been

used directly or indirectly for various NLP tasks. Sarker et al. [50] used MetaMap [5] to extract semantic types and concepts from the text for Adverse Drug Reaction (ADR) detection. The feature-rich classification approach used resulted in greater F-scores on all the test datasets and combining multiple similar corpora resulted in a significant increase in the ADR F-scores. The multi corpora approach is useful in imbalanced datasets and reduces time and cost to annotate the data. Cao, et al. [10] extracted similar features for complex question classification based on topics and created AskHERMES, a clinical question answering system to perform robust semantic analysis and output answers in the form of extractive summaries.

Weiming et al. [59], Cao, et al. [11] and Hritovski et al. [25] make use of SemRep [48], Metamap and UMLS [26] for biomedical question answering system. The United Medi-cal Language System (UMLS) Metathesaurus, the largest biomedical thesaurus, provides a representation of biomedical concepts semantically classified and both hierarchical and non-hierarchical relationships among them. MetaMap is an application developed by researchers at the National Library of Medicine (NLM) that maps biomedical text to Metathesaurus or presents data related to the metathesaurus in the text. Initially developed to improve the retrieval of bibliographic materials such as MedLine citations, it is now used extensively for data mining and information retrieval and is a key aspect in NLMs Medical Text Indexer.

Hristovski et al. [25] created a web-based application, SemBT, that provides answers instantaneously. The application utilizes the semantic relations extracted with SemRep from the entire MedLine citations up to 2012 and the instances were organized in a relational database from which answers are drawn. Rindflesch et al. [48] described a method for interpreting linguistic structures that encode hypernymic propositions. The method combines underspecified syntactic analysis and structured domain knowledge of the ULMS. To ensure the compatibility of the two processes, semantic groups from the Semantic Network are used. The semantic resources have been a good addition to improve on various NLP tasks, however, it remains an underexplored domain in the field of biomedical word embeddings.

1.4. Unifying Biomedical Tools and Word Embeddings

With the steady rise in the utilization of Electronic Health Records (EHRs) by the medical community, there is a possibility of better health care through the data obtained from the EHRs because these records provide important information related to patients which can be used across health care applications. This has been a driving factor for research in the biomedical Text Classification and Question Answering systems. With the increasing population, the workload of clinicians has increased significantly. Online biomedical corpora have answers to clinic questions when utilized properly. Jonnalagadda et al. [28] studied the feasibility of an automatic summary generating model for topics, composed of sentences extracted from the MedLine database. Although individual research has been conducted in the area of word embeddings to solve BioNLP problems and utilization of domain knowledge to improve various BioNLP tasks, there remains a large gap of research in the area of word embeddings utilizing semantic resources in the biomedical domain. Some of the research involving both is done by Yu, et al. [62], Cohen, et al. [13] and Abdeddaim, et al. [2] who have made use of UMLS/Mesh lexicon to improve word embeddings. Cohen et al. [13] proposes the Embedding of Semantic Predications (ESP), a probabilistic approach for encoding predications, a variant to their Predication-based Semantic Indexing (PSI). ESP has better performance for lower dimensionalities and exhibits comparable correlation with human judgment across the dimensionalities tested. PSI provides better performance when retrieving explicit relationships in larger dimensionality datasets like the SemMedDB. Yu et al. [62] introduced a semantic similarity measure utilizing both the biomedical taxonomy and vector space word representations to determine the degree of semantic and syntactic similarity between words. Their results displayed that the 'retrofitting model' proposed by Faruqui et al. [21] that incorporated information from semantic lexicons into word representations such that similar words have similar word representations produces higher correlation with the judgment of doctors compared to other existing techniques. Abdeddaim, et al. [2] proposed a MeSH gram neural model that is a variant of Skip-Gram model and utilizes the Medical Subject Headings (MeSH) descriptors instead of words. Implementation of the

model resulted in the performance equal to state-of-the-art models and techniques. Our paper is unique in utilizing SemMedDB information and constructing new biomedical word embeddings which are semantically rich and more useful compared to previous biomedical word embeddings.

The remainder of this paper's chapters discuss the preparatory research, background, method, and results of this thesis. Chapter 2 is dedicated to the preliminary research that was conducted and how through those findings I arrived to my final thesis research. Chapter 3 provides background information, tools, the implementation of those tools and resources that were used in this approach. Chapter 4 delves into word embedding design and efficiency. Chapter 5 discusses the dataset used for experiments and the results of the evaluation. Chapter 6 concludes the research and gives way to future work.

CHAPTER 2

PRELIMINARY RESEARCH

This chapter provides information for the initial research that was carried out and how it helped in arriving at my final thesis approach. The initial research focus was a Biomedical Question Answering System. However, through the progression of this research, we encountered many failures and learnings that directed this research towards biomedical semantic embeddings, which resulted into the final thesis. This chapter provides literature on biomedical question answering and question classification, which is of interest to many researchers in the NLP and biomedical domain. This chapter details the preliminary research and how I arrived at final thesis, it is divided into the following sections:

- Biomedical Question Answering
- Question/Text Classification
- Transition to biomedical semantic embeddings

2.1. Biomedical Question Answering

A Question Answering(QA) system takes the search engine to the next level, by providing exact answers as output for a given natural language question. Such a system is built by making efficient use of Natural Language Processing, Information Retrieval, and Machine Learning techniques. The answers are generated by querying through a knowledge base like Wikipedia or specific web pages on the World Wide Web.

QA systems are broadly classified as open or closed domain based on the knowledge base. Closed domain systems answer questions that are specific to particular topics like baseball, medicine and sometimes have constraints on the type of question that can be given as input. These exploit domain-specific knowledge formalized in ontologies and academic text. Open-domain systems do not have any constraints on the type of question that can be asked and as a result, have a very broad knowledge base covering a huge chunk horizontal domain.

In QA systems, it is convenient for a user to input a question in Natural Language but this increases the complexity of the model as it has to identify the correct question among the several types and extract the answer accordingly. There are several methods to accomplish this task. Extracting information from the question is the first step in any QA systems, which involve tasks like question classification(question type identification) or Keyword extraction. After keyword extraction and question type identification, an information retrieval system is used to retrieve documents based on the information extracted in the first step. Once the relevant documents are retrieved, top relevant paragraphs are selected from these documents and finally, candidate answers are selected from these paragraphs. Finally, several techniques are used to validate the candidate answers and a score is assigned to each one based on relevance to the question's context. The answer with the highest score is chosen, converted into a presentable form using parsing and is produced as output by the system. Galitsky et al. [23] worked on QA system design and proposed a multi-agent system in which each domain is represented by an agent which tries to answer questions of that domain. A meta-agent controls the co-operation between the agents and chooses the most optimal answers. By this process, the system provides the preciseness of a vertical domain model and the breadth of a horizontal domain model. A common issue that affects the accuracy of QA systems are questions which combine two or more domains. The presence of meta-agent and multiple agents ensures that questions combining several domains are answered adequately.

Biomedical QA is a domain-specific system where the model tries to answer questions specific to the biomedical domain. Need for such a system arises from the availability of an enormous amount of biomedical literature in several databases like MeSH, Metathesaurus, PMC open access subset, MEDLINE and its use by various users to answer questions related to biomedical research. Majority of the current Biomedical QA systems are used for clinical question answering.There has been an increase in research in building various biomedical QA systems [7], [1], [11], [59], [25], [24], [3]. Some of the major ones are discussed below:

- Cao et al. [11] created AskHERMES, a clinical question answering system, tested by physicians on ease of usage and accuracy of answering questions. It was trained using

MEDLINE, eMedicine, PubMed and Wikipedia documents. The model was built on semantic analysis of words and focused on retrieval of semantic content. The main process of answering questions in AskHERMES is Question Analysis, Document Retrieval, Passage Retrieval, and Summarization, in that exclusive order. Testing over several parameters resulted in AskHERMES performing on equal footing as state of the art models like Google and UptoDate.

- Abacha et al. [1] proposed MEANS, a semantic QA system for the medical domain. The system integrated NLP techniques and semantic web technologies for deep analysis of questions and documents in the MEDLINE knowledge base. The overall performance of the system was improved using query-relaxation. The process of answer generation follows NL question processing followed by answer search and ranking using query relaxation and semantic search. The paper successfully tackled automatic question-answering in Biomedical NLP.
- Hristovski et al. [24] launched SemBT, which uses MEDLINE citations extracted using SemRep as its knowledge base. The semantic instances extracted using SemRep collectively formed a database and the system answered questions by searching through this database. The system was released as a web-based application that provided an accuracy of 68%.
- Weiming et al. proposed a QA system based on ULMS relations. The system generated phrase-level answers after searching through the knowledge base. SemRep was used to identify relations in the database but failed to completely extract all the relationships. comparison with other model shows that this system produced high precision and recall.

The first step in building a QA system is question processing to extract information that will be used by other components of the system. Lexical Answer Type(LAT) is one of the most significant information that should be identified from a question, which can significantly reduce the number of documents to be retrieved for a given question and quickly help find sentences that contain answers to the question. Although, there has been an increase in

research in building various biomedical QA systems, very less work has been done that focuses on the LAT Prediction component. Therefore, my initial research focus was to improve the question classification which would help in improving the overall performance of a Question Answering system.

2.2. Question/Text Classification

Answer type prediction is the first step in a Question Answering(QA) System design, aiming to predict answer types with a goal of reducing the number of documents searched to find an answer for a natural language question. Lexical Answer Type(LAT) information can also be used to rank the list of possible answers generated for a question. Answer Type prediction is a text classification problem that provides useful information, which is used by QA systems.

Although, there has been an increase in research in building various biomedical QA systems, very less work has been done that focuses on the LAT Prediction component. Related work on this task includes Yang, Zi et al. [60], who provided a series of features such as tokens, semantic types, head tokens, etc. for question classification that were inspired from Li and Roth [37]. Also, they used UMLS semantic types to label questions. Cao, Yong-gang, et al. [10] and Kobayashi and Shyu [33] have also used UMLS semantic types to improve question classification, but they used it in a different context [18], with the aim of classifying questions into Biomedical topics and Taxonomies [20] rather than LATs. Most open domain QA systems make use of resources such as WordNet [42], to improve QA systems. But for biomedical question answering, mastering domain knowledge in QA becomes a necessity and challenge due to its richer domain-dependent terminologies and definitions.

2.2.1. Method

My preliminary research involved utilizing UMLS resources to improve biomedical QA performance, in particular, utilizing Semantic Network and SemMedDB provided by UMLS to generate better features for LAT prediction. A generalization of semantic types was done into general labels and SemMedDB was used for feature engineering.

Semantic Types(label) Generalization: Most existing methods directly use semantic types as labels for prediction in classification model. However, due to complexity of potential answers and lack of training data in reality, this may not be a good choice. In my work, I choose to cluster relevant UMLS semantic types potentially associated with the answers and use a grouping of such clusters as labels for question classification.

Feature Engineering: The features previously used in question classification are bag of words (BOW), bigrams, Part-of-Speech (POS), semantic types, head words, etc. In this research new features were created from Semantic Network(SemNet) and SemMedDB(SemMed). These features were generated based on the idea that if a question contains a concept belonging to one particular semantic type, then it is more likely that the user is asking about the information carried by concepts belonging to its closely related semantic type(s).

2.2.2. Results

Drug Questions Data Set: I have used the dataset provided by BioASQ [57], containing questions, in English, along with gold standard answers constructed by a team of biomedical experts. BioASQ organizes challenges on biomedical semantic indexing and question answering (QA). A set of 170 drug-related questions, including 107 factoid questions and 63 list questions were gathered and used for classification and analysis. Factoid questions refer to those that have single concepts as answers and list questions are questions that have a list of concepts as answers. My initial research interest was to answer drug questions which would be beneficial and helpful for common public. With this focus unrelated questions was filtered out by setting up MetaMap to maintain question and answer pairs that contain concepts belonging to the semantic type [pharmacologic substance] in either question or answer sentence. This has resulted in 170 drug related questions.

BioMedLAT Questions Data Set: BioMedLAT corpus is provided by Neves, et al. [44]. This corpus consists of 643 list/factoid questions from BioASQ training data and the class labels of these data are semantic types from UMLS. We have taken those questions which have at least 10 instances for each class label. This leads to 515 questions with 15 class labels from the BioMedLAT corpus.

Experiments: To analyze the significance of proposed features, different models were created using different combinations of possible features and the results are presented in Table 2.1. Logistic Regression(LR), Support vector Machine(SVM) and Multi Layer Perceptron(MLP) were used to train models and the average accuracies were calculated using 10 fold cross validation. For Correlation analysis, I compared the performance of using single features, including the bag-of-words(BOW), semantic types(Sem), feature generated from Semantic Network(SemNet), feature derived from SemMedDB (SemMed), features derived by using both Semantic Network and SemMedDB (here SemNet+SemMed refers to the feature vector combining features from SemNet and SemMed). It can be seen that other than the content-based feature (which is also domain independent), i.e., the BOW feature, within the choices of domain specific features, a newly proposed features, SemMed, significantly performs better than using semantic types (Sem) feature. In particular, the SVM model trained on SemMed gave 80.00% accuracy which has a 9.42% increase from using Sem, and the MLP model trained using SemMed achieved an accuracy of 81.76%, an 8.24% increase from using Sem. This shows proposed new features are good additions to the list of domain specific features that can be used for LAT prediction in biomedical QA systems, compared with the work done by Cao, et al. and Kobayashi and Shyu where the Semantic Type feature is the only domain specific feature that has been designed.

To further demonstrate the power of these domain specific features, Support Vector Machine based model was trained on a set of 515 questions from the BioMedLAT corpus. It is evident in Table 2.2 that new proposed features performed consistently well with this larger dataset. When SemMed and SemNet features were added to commonly used features, such as bag-of-words(BOW) and semantic types(Sem), there have been a significant improvement of performance (14.48% and 8.3% increases compared with using BOW and BOW+POS+Sem, respectively) This is also consistent with the findings from Cao, et al. and Kobayashi and Shyu related to topic classification, where they claim that when combining text based features with additional domain specific features, the system performance can be improved. It is evident from these results that features created from Semantic Network and SemMedDB

	BOW	**Sem**	**SemNet**	**SemMed**	**SemNet +SemMed**	**BOW +POS +Sem**	**BOW+POS +Sem +SemNet**	**BOW+POS +Sem +SemMed**	**BOW+POS +Sem +SemNet +SemMed**
LR	84.11%	57.05%	60.00%	69.41%	68.23%	83.52%	82.35%	88.23%	**88.82%**
SVM	86.47%	70.58%	61.17%	80.00%	78.23%	90.00%	86.47%	**90.00%**	**90.00%**
MVP	85.88%	73.52%	70.00%	81.76%	80.58%	85.88%	85.29%	**88.82%**	87.05%

TABLE 2.1. Performance comparison of different feature sets using Logistic Regression(LR), Support Vector Machine(SVM) AND Multi Layer Perceptron(MLP) on drug questions where features bag Of Word(BOW), Semantic Types(Sem), Semantic Network(SemNet), SemMedDB(SemMed), and combination of these features(SemNet + SemMed) were used

BOW	**Sem**	**SemNet**	**SemMed**	**SemNet +SemMed**	**BOW +POS +Sem**	**BOW+POS +Sem +SemNet**	**BOW+POS +Sem +SemMed**	**BOW+POS +Sem +SemNet +SemMed**
61.16%	48.73%	37.47%	27.57%	44.97%	66.91%	69.55%	**70.54%**	**75.64%**

TABLE 2.2. Performance comparison of different feature sets using Support Vector Machine(SVM) on BioMedLat questions where features bag Of Word(BOW), Semantic Types(Sem), Semantic Network(SemNet), SemMedDB(SemMed), and combination of these features(SemNet + SemMed) were used

provided better results compared to Semantic features used by Cao, et al. and Kobayashi and Shyu and our features are good additions to list of features that can be used for LAT prediction in QA systems.

2.3. Transition to Biomedical Semantic Embeddings

Although, the results attained in this preliminary research outperformed previous work, it failed to address some issues. Firstly, the features created from semmeddb and semantic network used hot-one encoding and as a result, the number of features increased, thereby resulting in the problem of overfitting. Secondly, The test set used was very small and was not able to provide a definitive effectiveness of new features. Lastly, most recently, state of the art results in open domain text classification has been attained by making use of neural network model with word embeddings as features.

Fortunately, using domain specific semantic resources was the right step towards improving text classification in biomedical domain, even though the way it was utilized in this research was not satisfactory. This preliminary research helped in understanding biomedical resources such as semantic network and SmemMedDB and provided with a venue to explore these tools further.

Keeping in mind the state of the art text classification models i.e. neural networks using word embeddings as features, we explored how these word embeddings can be trained efficiently utilizing biomedical resources that were explored in preliminary research for better question classification. This results in our thesis on Bioemdical Word Embeddings.

CHAPTER 3

BACKGROUND

Considering the extensive research in model architecture to train biomedical word embeddings, our motivation of the study was not to tweak these already established state of the art architectures. Rather, the novelty of our research lies in the usage of domain-specific semantic resources to provide better inputs to these models. We were also influenced by the fact that there is minimal research in biomedical word embeddings exploiting these semantic resources, especially about using SemMedDB to improve the quality of biomedical word embeddings. Even though these biomedical semantic resources have been used to solve NLP problems that are tackled by word embeddings, these semantic resources have not been used in unison with word embeddings. We took this opportunity to reap the benefits of both, the biomedical domain knowledge and state of the art in NLP research word embeddings.

We use information in SemMedDB to engineer hybrid sentences which we call "semantic sentences" for the scope of this study. These semantic sentences in addition to biomedical text are provided as input to our word2vec(skip-gram) to train our semantic word embeddings. As a result, we create biomedical word embeddings whose vectors are not only words but also semantic types. Later in this paper, we explain the benefit of inducing these semantic types into vector space.

This section further provides an outline of the tools and resources utilized in research to achieve semantic word embeddings. The section further provides a brief description of how the tools are implemented in this novel approach. These domain-specific tools and resources are provided by the National Library of Medicine (NLM) and available for use.

3.1. Word2vec

Creating word embeddings or distribution representation of words is based on the idea that we can know a word by the company it keeps. The main intuition is that if two different words have very similar contexts (that is, what words are likely to appear around them), then those two words are similar in meaning. For example, we could expect that syn-

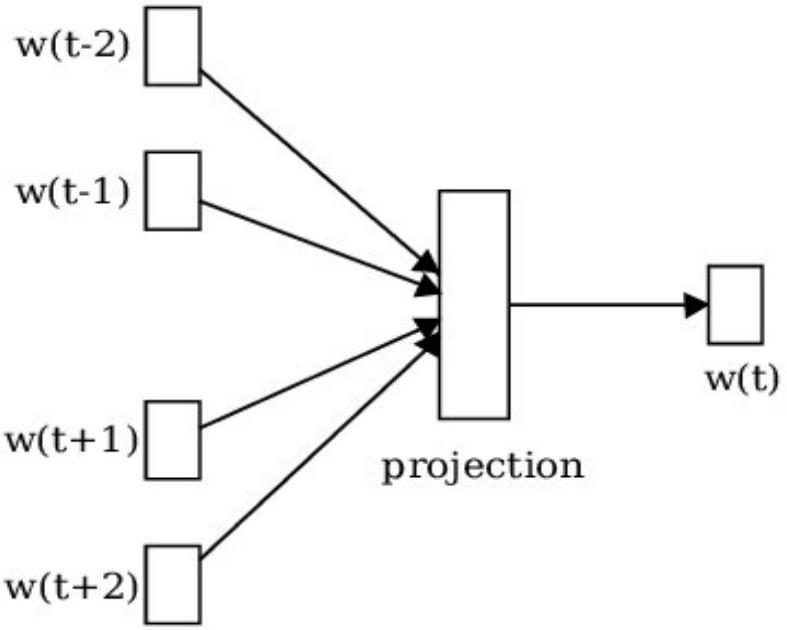

FIGURE 3.1. The CBOW architecture taken from Mikolov et al. [39]

onyms like intelligent and clever would have very similar contexts or words that are related, like engine and motor, would probably have similar contexts as well.

Mikolov et al. [39] proposed two architectures as part of their word2vec tool to learn distributed representation of words: continuous bag-of-words(CBOW) and skip-gram for computing the word vectors from very large data sets. These model architectures are two-layered shallow neural networks and yield a remarkable enhancement in accuracy when tested for syntactic and semantic word similarities with much inferior computational cost, and both these models still stand as state of the art for training word embeddings. Continuous Bag of Words architecture predicts the target word based on the source context words. It means that weights are obtained from the surrounding words and the probabilistic model then generates the output word. The output is not influenced by the order in which context words occur in the dataset. The Skip-gram architecture is a generalization of the n-gram and provides a way to overcome the issue of data sparsity in traditional n-gram analysis. In its primal form skip-gram is the exact opposite of the CBOW and uses the current word to predict sur-

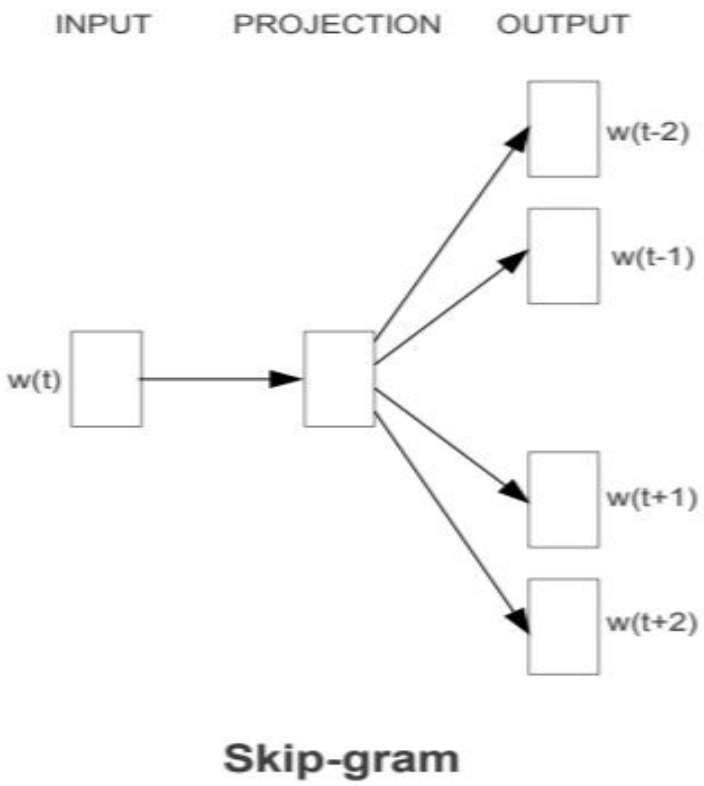

FIGURE 3.2. The Skip-gram architecture taken from Mikolov et al. [39], which predicts surrounding words given the current word.

rounding context words. The output of skip-gram is influenced by the positioning of similar context words in the corpus with closer words having more weights than the distant context words. Despite being slow, skip-gram performs better than CBOW for infrequent words.

skip-gram. Given a text corpus, skip-gram targets at deducing word representations that are good at estimating the context words given a target word in a sliding window of text. Specifically, skip-gram takes each word in the corpus (denoted as w_t) and its surrounding words within a window of defined size (denoted as C_t) as input. The model then feeds each pair (w_t, w_c), where $w_c \in C_t$ into a neural network that is trained to maximize the log probability of neighboring words in the corpus. More formally, given a training corpus

represented as a sequence of words $w_1, w_2, ..., w_T$, the objective of the skip-gram model is to maximize the following function:

$$O = \sum_{t=1}^{T} \sum_{w_c \in C_t} logP(w_c|w_t)$$

For both its performance and popularity, we make use of skip-gram shown in Figure 3.2 to train our word embeddings model.

3.2. UMLS

The aim of the National Library of Medicine's(NLM) Unified Medical Language System (UMLS) [26] [9] resource is to ease the development of conceptual relationships between users and machine-readable data. The UMLS model comprises of three centrally developed Knowledge Sources: a Metathesaurus, a Semantic Network, and SPECIALIST Lexicon and Lexical Tools. It integrates and delivers key terminology, coding and classification standards, and associated resources for the creation and advancement of biomedical information systems and services which include EHRs. Metathesaurus consists of biomedical terms and codes from many vocabularies, including CPT, ICD-10-CM, LOINC, MeSH, RxNorm, and SNOMED CT. This information is gathered by making use of Semantic Network and lexical tools to group synonymous terms, categorize biomedical concepts by semantic types, link health information, medical terms, drug names, and billing codes across different computer systems. Semantic Network consists of broad categories of semantic types and relationships between them. SPECIALIST Lexicon and Lexical Tools consist of Natural language processing tools such as SemRep, Metamap, etc. Details of each resource that is used in this research have been provided in the below sections.

3.3. Semantic Network

Semantic Network [38] [51] provides a categorization of all concepts represented in the UMLS Metathesaurus and a set of relationships that exist between these concepts. It is a knowledge base that contains semantic relations between Biomedical terms present in ULMS. Semantic Network contains information about the set of semantic types, or categories, which

may be assigned to a biomedical concept, and it defines the set of relationships that may hold between these semantic types. The Semantic Network contains 133 different semantic types like anatomical structure, biological functions, chemicals, events, conditions, etc., and 54 relationships that can exist between these concepts. The semantic types are represented as nodes in the Network, and the relationships between them are the links.

These semantic types are represented using unique identifiers comprising of four characters, ex: phsu for "Pharmacologic Substance". We make use of these unique identifiers in our paper and also provide its full form as necessary. Example of a relationship in a semantic network is "phsu—affects—dsyn" where "phsu"(Pharmacologic Substance) and "dsyn"(Disease or Syndrome) are semantic types and "affects" is the relationship between them. A prime example is the assignment of smoking to phsu and lung cancer to dsyn creating "smoking causes lung cancer" with causes being the relationship. These semantic types and relationships are defined as follows by the semantic network:

phsu: A substance used in the treatment or prevention of pathologic disorders. This includes substances that occur naturally in the body and are administered therapeutically.

dsyn: A condition which alters or interferes with a normal process, state, or activity of an organism. It is usually characterized by the abnormal functioning of one or more of the host's systems, parts, or organs. Included here is a complex of symptoms descriptive of a disorder.

affects: Produces a direct effect on. Implied here is the altering or influencing of an existing condition, state, situation, or entity. This includes has a role in, alters, influences, predisposes, catalyzes, stimulates, regulates, depresses, impedes, enhances, contributes to, leads to, and modifies.

below are some examples out of 6217 total semantic relations defined in the semantic network:

- Acquired Abnormality(acab)—affects—Human(humn)
- Disease or Syndrome(dsyn)—isa—Biologic Function(biof)
- Disease or Syndrome(dsyn)—occurs_in—Family Group(famg)
- Enzyme(enzy)—ingredient_of—Clinical Drug(clnd)

- Enzyme(enzy)—interacts_with—Receptor(rcpt)
- Food(food)—affects—Organism Function(orgf)
- Gene or Genome(gngm)—part_of—Human(humn)
- Hazardous or Poisonous Substance(hops)—complicates—Neoplastic Process(neop)
- Hormone(horm)—disrupts—Cell Function(celf)
- Neoplastic Process(neop)—result_of—Genetic Function(genf)
- Nucleic Acid, Nucleoside, or Nucleotide(nnon)—interacts_with—Enzyme(enzy)
- Organic Chemical(orch)—affects—Disease or Syndrome(dsyn)
- Organism(orgm)—interacts_with—Virus(virs)
- Pathologic Function(patf)—affects—Animal(anim)
- Pharmacologic Substance(phsu)—diagnoses—Disease or Syndrome(dsyn)

We use information in semantic network as standard guidance for working with semantic type vectors that are introduced as part of our research. This information will be used during our applications, similarity and relatedness related tasks which will come in later sections of this paper.

3.4. MetaMap

MetaMap [5] [4] [6] is a tool developed to map biomedical texts to biomedical concepts. These concepts are classified by semantic types[1] defined by the UMLS Metathesaurus, such as Body Part, Organ, or Organ Component (T023, A.1.2.3.1) and Anatomical Structure (T017, A1.2).

MetaMap uses a knowledge-intensive approach based on symbolic, natural language processing (NLP) and computational linguistic techniques to identify biomedical concepts in the text. For this reason, it has been used in many applications such as information retrieval, data mining and decision support systems. It breaks the text into phrases and then based on mapping strength, for each phrase, return a ranked list of mapping options. It allows the user to use domain-specific customized dictionaries.

[1] `https://metamap.nlm.nih.gov/Docs/SemanticTypes_2013AA.txt`

Given the biomedical text, MetaMap outputs the biomedical concepts in the text and provide details about these biomedical concepts, such as concept id, concept name, preferred name, semantic type, etc.

For Example: when we provide "Histochemical changes in psoriasis treated with triamcinolone" as input text to MetaMap, the following biomedical concepts are identified in the text along with its information:

Psoriasis [Disease or Syndrome]

changes [Quantitative Concept]

Treated with [Therapeutic or Preventive Procedure]

Triamcinolone [Organic Chemical, Pharmacologic Substance]

We only show the concept [semantic type] information above, since these are the two values that we make use of in our research, however, MetaMap provides many other details related to biomedical concepts identified. We make use of the concept and semantic types identified in the biomedical text to use them as input features for our text classification task.

3.5. SemRep

SemRep [48] is another useful tool provided by NLM that extracts semantic predications from sentences in biomedical text. It is a multilingual graph-based platform that integrates concepts and their semantic relations extracted from various databases such as Wikipedia, UMLS, OpenThesaurus, and WordNet. This makes SemRep a powerful tool for data integration tasks based on different background knowledge like finding semantic correspondences between schemas and ontologies or semantic enriching of fairly straightforward mappings. SemRep maps relations and concepts similar to the nodes approach and if a relationship does not exist, different techniques are used to derive it. Semantic Predication consists of a subject argument, an object argument, and the relation that exist between them. The subject and object argument of each predication are concepts from the UMLS Metathesaurus and their binding relationship is a relation from the UMLS Semantic Network. Table 3.1 shows an example of relations that are extracted by passing the biomedical text through SemRep.

Input Sentence	**Output Relations**
We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia.	• Hemofiltration-TREATS-Patients • Digoxin overdose-PROCESS_OF-Patients • Hyperkalemia-COMPLICATES-Digoxin overdose • Hemofiltration-TREATS-Digoxin overdose
Treatment of tumors of the eyeball with radium and radiotherapy.	• Neoplasm-PART_OF-Eye • Radiation therapy-TREATS-Neoplasm • Radium-TREATS-Neoplasm
Significance of Alimemazine in the treatment of delirium tremens	• Alimemazine-TREATS-delirium tremens
Surgical treatment of congenital cardiovascular anomalies accompanied by cyanosis.	• Cyanosis-COEXISTS_WITH-Cardiovascular Abnormalities

Hypophyseal tumors induced by estrogenic hormone	• Neoplasm-PART_OF-Pituitary Gland • Estrogenic-CAUSES-Neoplasm
Myoclonus and epilepsy appearing in various patients during the administration of chlorpromazine	• Epilepsy-PROCESS_OF-Patients • Myoclonus-PROCESS_OF-Patients

TABLE 3.1. Examples of semantic relations extracted from biomedical sentences

3.6. SemMedDB

The Semantic MEDLINE Database (SemMedDB) [31] [30] [22] is a database of semantic predications extracted by running SemRep [48] on PubMed. SemMedDB currently contains information about approximately 94.0 million predications from all of PubMed citations (about 27.9 million citations, as of December 31, 2017) and provides valuable information for extraction purposes. It is used as a knowledge resource to assist in hypothesis generation and literature-based discovery in biomedical text resources. The presence of semantic types in SemMedDB makes it an excellent input source to increase the performance of Machine learning models. This database consists of 8 tables out of which SENTENCE, PREDICATION, and PREDICATION_AUX tables are used for our research. More specifically we use row data present in SENTENCE, SUBJECT_SEMTYPE, OBJECT_SEMTYPE, SUJECT_TEXT and OBJECT_TEXT columns of these tables. Figure 3.3
shows the entity-relationship diagram of SemMedDB, and this figure provides details of all the tables and columns present in the SemMedDB along with the relationship between these tables.

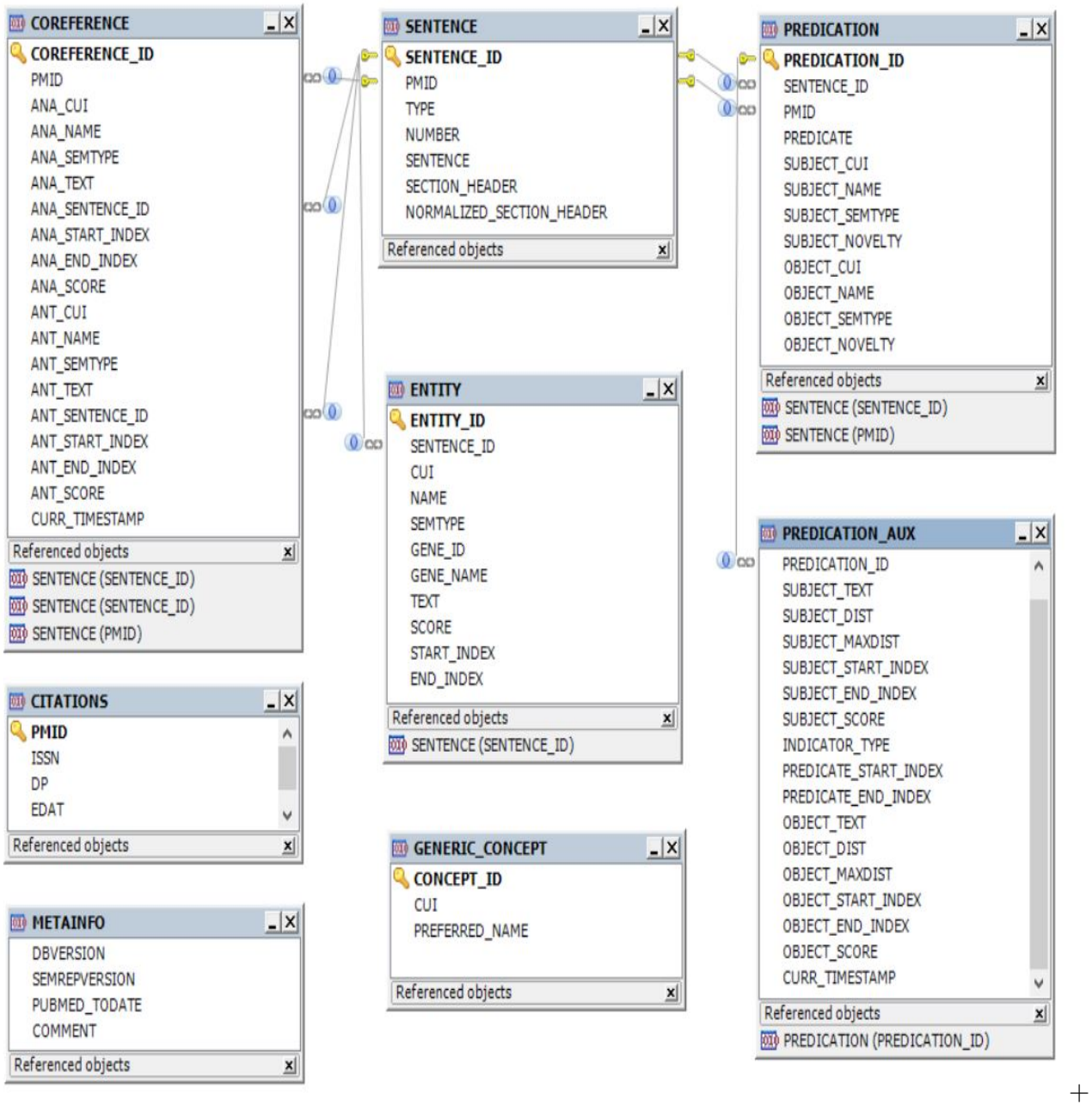

FIGURE 3.3. The Entity-Relationship diagram of SemMedDB

The most critical part of our research, i.e., "Semantic Sentences" are created by extracting information from these columns of SemMedDB.

CHAPTER 4

METHOD

4.1. Semantic Embeddings

Our research idea can be explained in two simple steps:

(1) Use SemMedDB to create hybrid sentences or "semantic sentences".

(2) Use these hybrid sentences as input to our word embedding training model.

Word embeddings are created by providing a list of sentences as input to the skip-gram model, and Skip-gram model aims to deduce word representations based on context words. The main goal of our research is to introduce semantic types into vector space of word embeddings. To do so it is essential that input to the skip-gram model has sentences or text that contain semantic types. With this aim of providing semantic types and its context information to the skip-gram model, we first create "semantic sentences" which have semantic types and their context information. By training skip-gram model on these semantic sentences, we get word embeddings with a distributed representation of words and semantic types.

SemMedDB provides biomedical domain-specific semantic data related to predicate sentences. Semantic sentences are constructed using SENTENCE, SUBJECT_SEMTYPE, OBJECT_SEMTYPE, SUJECT_TEXT, and OBJECT_TEXT columns in SemMedDB tables. This is done by replacing subject and object text in a sentence with their respective semantic types. Algorithm 1 illustrates how semantic sentences are created from sentences. Table 4.1 shows examples of semantic sentences created using semantic information from SemMedDB.

sentence	**subject**	**object**	**subject semantic type**	**object semantic type**	**semantic sentence**
Histochemical changes in **psoriasis** treated with **triamcinolone**.	triamcinolone	psoriasis	phsu	dsyn	Histochemical changes in **dsyn** treated with **phsu**.
Fulminating hepatic necrosis in a **patient** with **multiple myeloma** treated with urethan.	multiple myeloma	patient	neop	humn	Fulminating hepatic necrosis in a **humn** with **neop** treated with urethan.
Clinical importance of the Russian **spasmolytic** preparation Etaphen in visceral-reflex **stenocardia**.	spasmolytic	stenocardia	phsu	sosy	Clinical importance of the Russian **phsu** preparation Etaphen in visceral-reflex **sosy**.
Effect of cyclic progestin-estrogen therapy on sebum and **acne** in **women**.	acne	women	dsyn	humn	Effect of cyclic progestin-estrogen therapy on sebum and **dsyn** in **humn**.

Surgical treatment of **congenital cardiovascular anomalies** accompanied by **cyanosis**.	congenital cardiovascular anomalies	cyanosis	dsyn	sosy	Surgical treatment of **dsyn** accompanied by **sosy**.
Significance of **Alimemazine** in the treatment of **delirium tremens**.	Alimemazine	delirium tremens	phsu	dsyn	Significance of **phsu** in the treatment of **dsyn**.
Activation of **Hageman factor** by **L-homocystine**.	L-homocystine	Hageman factor	aapp	gngm	Activation of **gngm** by **aapp**.
Acute suppurative infections of the **salivary glands** in the newborn.	salivary glands	Acute suppurative infections	bpoc	dsyn	**dsyn** of the **bpoc** in the newborn.
Hypophyseal **tumors** induced by **estrogenic** hormone.	estrogenic	tumors	horm	neop	Hypophyseal **neop** induced by **horm** hormone.

FACTORS AFFECTING **MOTOR PERFORMANCE** IN FOUR-MONTH-OLD **INFANTS**.	MOTOR PERFORMANCE	INFANTS	orgf	humn	FACTORS AFFECTING **orgf** IN FOUR-MONTH-OLD **humn**.

TABLE 4.1. Examples of semantic sentences created from semantic information of sentences in SemMedDB, here "phsu"(Pharmacologic Substance), "dsyn"(Disease or Syndrome), "neop"(Neoplastic Process), "humn"(Human), "sosy"(Sign or Symptom), "aapp"(Amino Acid, Peptide, or Protein), "gngm"(Gene or Genome), "bpoc"(Body Part, Organ, or Organ Component), "horm"(Hormone) and "orgf"(Organism Function) are biomedical sematic types defined by UMLS

We use information in SemMedDB to construct 'semantic sentences'. Using these 'semantic sentences' as input to skip-gram, word embeddings are created, which consist of both biomedical terms and UMLS semantic types as explained by Arshad et al. [52]. Figure 4.1 shows how the word embeddings are created. Finally, the word embeddings created in this process not only have distributed representation of words but also semantic types, which are at our disposal for various tasks such as calculating relatedness between two biomedical terms and text classification utilizing the vectors of these semantic types as input features.

For analysis purpose we create two-word embeddings SMDB and SMDB+ using the skip-gram model from the gensim library, and we use the default parameters provided by gensim library such as window size of 5 and vector dimension of 200. The training sentences which we use for these embeddings are what differentiate them, detailed as below:

Algorithm 1 Semantic Sentence Creation

```
1: procedure createSemanticSentences
2:     n ← total no. of Sentences
3:     sentences ← list of sentences
4:     subjectSemType ← semantic type of subject
5:     objectSemType ← semantic type of object
6:     subjectText ← subject text
7:     objectText ← object text
8:     semanticSentences ← [ ]
9:     for i ← 1 to n do
10:        tempSentence ← sentences[i]
11:        [t] tempSentence ← replace(
             sentences[i], subjectText[i],
             subjectSemType[i])
12:        [t] tempSentence ← replace(
             tempSentence, objectText[i],
             objectSemType[i])
13:        semanticSentences.append(tempSentence)
14:    return semanticSentences
```

4.1.1. SMDB

This is the word embeddings trained using 94 million sentences in SemMedDB as input to the skip-gram model, and we use this as a baseline to compare with semantic embeddings created in this research. The vector space of this embedding has biomedical terms but not the semantic types defined in UMLS.

4.1.2. SMDB+

This is our special word embedding created using 94 million sentences in SemMedDB and respective 94 million semantic sentences which we created utilizing semantic types in-

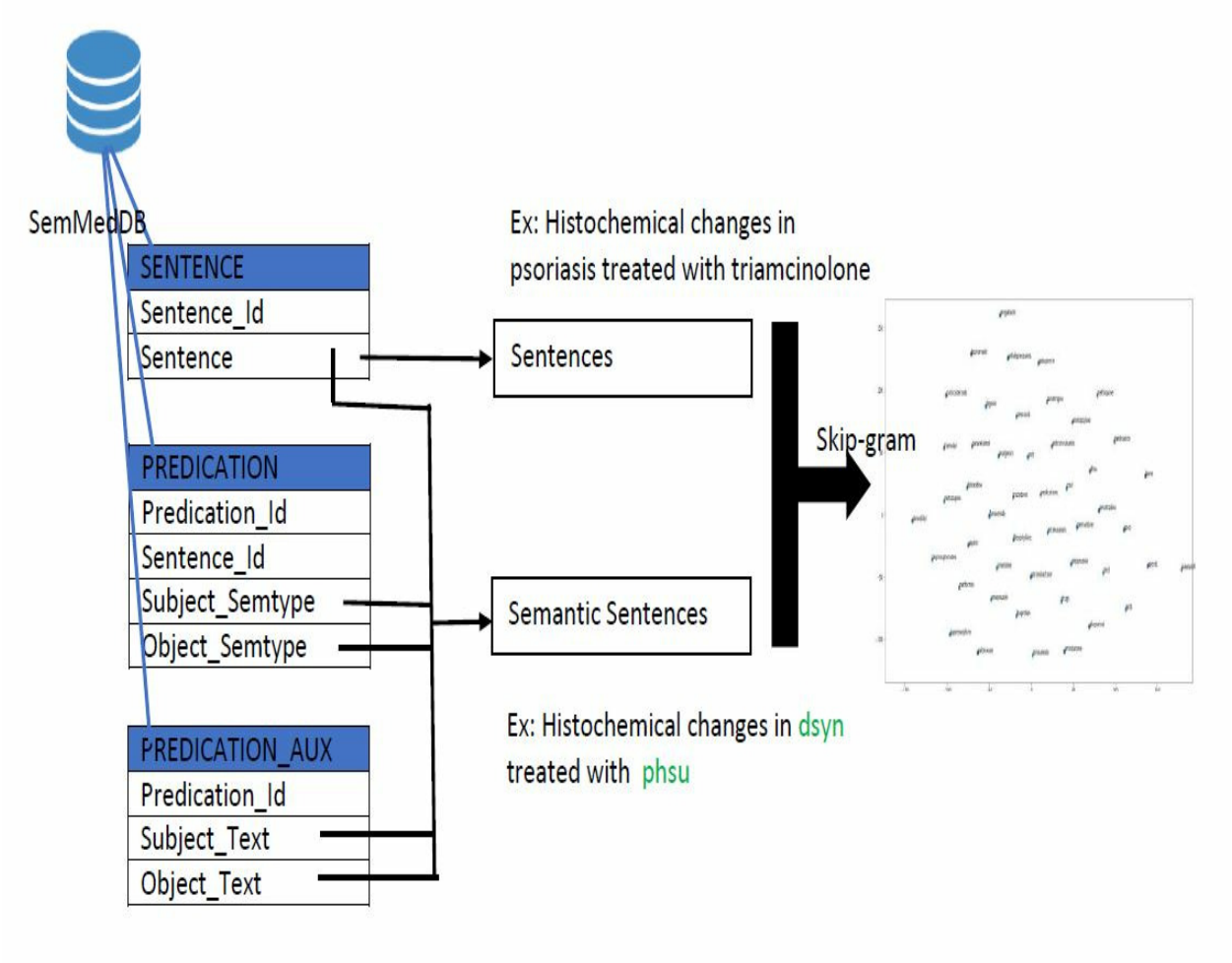

FIGURE 4.1. biomedical semantic embeddings are created from sentences and "semantic sentences", where "semantic sentences" are created by utilizing semantic information in SemMedDB taken from Arshad et al. [52]. Here "phsu"(Pharmacologic Substance) and "dsyn"(Disease or Syndrome) are biomedical semantic types defined by UMLS

formation. It consists of both biomedical terms and semantic types in vector space.

4.2. Applications

Word embeddings are a powerful tool because of the relationship it captures in different words. For example: on taking vectors of "queen", "woman" and "man" and computing vector("queen") - vector("woman") + vector("man"), the output generated is the vector of

"king". Similarly, the standard UMLS semantic types introduced in our semantic word embedding captures various semantic relations between different biomedical terms. To further demonstrate this idea we show below how we can compute a disease vector given the value of its related drug.

Drug-Disease pairs: using vectors of a drug and semantic types of drug and disease, a set of related diseases are generated. For example: calculating vector("bortezomib") - vector("orch") + vector("neop") gives a vector X and on computing the nearest word vector to X we get "myeloma", in terms of word embeddings. Bortezomib is a drug used to cure multiple myeloma, the semantic type of bortezomib in UMLS is defined as orch(organic chemical), and neop(neoplastic process) is the semantic type which is assigned to cancer diseases.

By using the knowledge of semantic types from the semantic network and utilizing the semantic types vectors from our word embeddings we were able to arrive at Myeloma cancer disease just by knowing the name of the drug Bortezomib. Similarly, there are 133 different semantic types in the semantic network and 54 relationships that exist between these semantic types, and using our semantic word embeddings these semantic types can be exploited to get pairs such as drug-disease, drug-target, etc.

4.2.1. Intrinsic Evaluation

Pakhomov, et al. [46] provides similarity and relatedness for clinical terms culminating in a set of biomedical pairs and scores associated with them based on the degree of similarity and relatedness. This dataset acts as a benchmark for the intrinsic evaluation of the embeddings presented here.

To compute similarity between vectors A and B, cosine similarity is used which is defined as follows:

$$\cos(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{AB}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} \mathbf{A}_i \mathbf{B}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{A}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{B}_i)^2}}$$

Similarity: From Pakhomov, et al. [46] observations and result data we can conclude that similarity between two biomedical terms exist when these two terms belong to the same

semantic types. For example, Medrol and prednisolone are the names of drugs which belong to the same drug class, treats similar diseases and both of them are assigned the same semantic type of phsu(Pharmacologic Substance) by UMLS. Similarity score in such pairs can easily be determined by finding cosine similarity between two biomedical terms. We compute the similarity measure using the cosine similarity as it has been used previously [43].

Relatedness: Relatedness between two biomedical terms is a difficult task to define computationally. From Pakhomov, et al. [46] observations and result data we can conclude that two semantically related terms can be of same or different semantic types. In the previous research [43] problem of relatedness is targeted by using the same cosine similarity score. Although cosine score is a good measure if the semantic types of two terms are the same, it is not a good measure to compute relatedness between two terms which are of different semantic types. For example, diabetes and insulin are semantically related but both of these biomedical terms are of different semantic types, where diabetes has the semantic type of disease or syndrome and insulin has the semantic type of pharmacologic substance. There exists a relatedness between these two terms because insulin is the drug which is used to treat diabetes.

Given the distribution of words in word embeddings, similar words are close to each other but words with different semantic types are far apart. Based on this phenomenon we define a better measurement of semantic relatedness here, and this measure of relatedness is only possible because of the special nature of our word embeddings which contains vectors of semantic types in addition to biomedical terms. This more accurate Relatedness score is determined by utilizing semantic types in the word embeddings to capture the relationship between two biomedical terms. For a given biomedical pair (term1,term2) it is possible to calculate relatedness using the formula:
relatedness(term1,term2)=cos(X,Y) where X="term1"-"sem1"+"sem2" and Y="term2", where "sem1" and "sem2" are semantic types of "term1" and "term2" respectively.
For example: relatedness("diabetes","insulin")=cos(X,Y) where X="diabetes"-"dsyn"+"aapp" and Y="insulin", here dsyn(disease or syndrome) is the semantic type of diabetes and

aapp(amino acid, peptide or protein) is semantic type of insulin. Relatedness score algorithm 2 is provided below for illustration:

Algorithm 2 Relatedness Calculation

procedure RELATEDNESS(TERM1,TERM2,SMDB+)
  $sem1 \leftarrow getSem(term1)$ ▷ use MetaMap to get semantic type
  $sem2 \leftarrow getSem(term2)$ ▷ use MetaMap to get semantic type
  $vectorTerm1 \leftarrow getVector(term1, SMDB+)$
  $vectorTerm2 \leftarrow getVector(term2, SMDB+)$
  $vectorSem1 \leftarrow getVector(sem1, SMDB+)$
  $vectorSem2 \leftarrow getVector(sem2, SMDB+)$
  $relatedVector \leftarrow vectorTerm1 - vectorSem1 + vectorSem2$
  $relatedness \leftarrow cos(relatedVector, vectorTerm2)$
  return *relatedness*

Since the distribution of words captures relations between different words, semantic types yield a better calculation of relatedness. Note in case there is no semantic type for a biomedical concept we can just use the word vector value ignoring the semantic type vector in the formula, which is same as the similarity measure. The results achieved in the relatedness score are comparably higher than those in other biomedical word embeddings.

4.2.2. Text Classification

Recently, CNNs and other neural networks have been widely used for various NLP tasks including sentence classification [61], [53], [29], [15]. Pre-trained word embeddings have shown effective results when used as input feature to train a sentence classification model [29], [32]. Applying Yoon Kim's [32] CNN model showcases the power of the seman-tic embeddings(SMDB+) compared to a normal word embedding(SMDB) which does not contain semantic types vector.

Normally, vectors of words present in a text are used as input features to a CNN text classifier. However, due to the nature of word embeddings developed here, the vector of

semantic types present in a text can be used as an additional feature to the CNN classifier. Semantic types in the text are determined by using MetaMap, and semantic type vectors are then appended to vectors of words in the text as input. Results of text classification are discussed in details in the following section.

CHAPTER 5

RESULTS

5.1. Data Sets

5.1.1. UMNSRS-Rel and UMNSRS-Sim

UMNSRS-Rel and UMNSRS-Sim are the datasets compiled by Pakhomov, et al. [46] which consist of semantically related and semantically similar biomedical terms. The degree of similarity and relatedness is captured by a score assigned by a group of eight medical residents. Based on these scores assigned we use top 12 semantically similar pairs from UMNSRS-Sim for similarity measure and top 12 related pairs from UMNSRS-Rel for relatedness measure.

5.1.2. Clinical Questions

A subset of clinical question dataset provided by the National Library of Medicine, which were accumulated by Ely et al. and D'Alessandro et al. [18], [19], [17]. This dataset consists of 4,654 clinical questions that arose during patients care and visit. This dataset contains information on topics assigned to each question, each question is assigned one or more topics to them from a set of 12 topics. This paper makes use of question set that belongs to the top five most recurring topics namely "Pharmacological", "Management", "Diagnosis", "Treatment & Prevention" and "Test". Distribution of these questions across topics is illustrated in Table 5.1.

5.2. Semantic Similarity

Similarities are computed between the two terms by finding cos(Term1, Term2) where Term1 and Term2 are word vectors from different word embeddings. Table 5.2 shows the results of similarity scores computed on top 12 similar terms from UMNSRS-Sim dataset using Pyysalo et al. [43], SMDB, and SMDB+ word embeddings. Word embeddings created from SemMedDB sentences(SMDB) and SemMedDB+semantic sentences(SMDB+) provided approximately 3.5% increase in average score demonstrating the effectiveness of this technique.

Question Topic	No. of Questions
Pharmacological	1594
Management	1403
Diagnosis	994
Treatment & Prevention	868
Test	746

TABLE 5.1. Question topics distribution

One interesting thing to note here is SMDB performed equally well, and this could be due to the quality of sentences in SemMedDB. Since SemMedDB has sentences which have a subject-object relationship in them, the similarity between words was captured better than Pyysalo et al. [43] embeddings which were trained on the entire PubMed.

5.3. Semantic Relatedness

The same cosine score for computing relatedness in Pyysalo et al. and SMDB was applied here. However, SMDB+ provides with semantic types vectors which can be exploited to better capture the relatedness between two terms. We calculate relatedness in SMDB+ word embeddings using the formula defined in the Method section. Table 5.3 shows relatedness score on top 12 biomedical pairs from UMNSRS-Rel which are different in semantic types but semantically related to each other. The 11% increase in the average score of SMDB+ from Pyssalo et al. showcases the power of introducing semantic types in the word embeddings and further bolsters the formula used to compute the relatedness score. Again SMDB performed better than Pyysalo et al. [43] due to a better quality of sentences. Cosine score is not a good measure of relatedness because words with similar semantic types are grouped but words with different semantic types are far apart in vector space. However, by using semantic types in SMDB+ vector space we were able to arrive at better relatedness measures.

Term 1	**Term 2**	**Pyysalo et al.**	**SMDB**	**SMDB+**
medrol	prednisolone	**0.60804**	0.53884	0.50556
lipitor	zocor	0.66682	**0.85673**	0.84450
thalassemia	hemoglobinopathy	**0.70729**	0.62486	0.61177
convulsion	epilepsy	**0.54465**	0.52118	0.51423
emaciation	cachexia	0.49607	**0.57806**	0.57375
dizziness	vertigo	0.72978	0.80141	**0.81400**
mycosis	histoplasmosis	0.55776	**0.61603**	0.59177
enalapril	lisinopril	0.94660	0.93651	**0.94842**
actonel	fosamax	0.67757	0.76767	**0.82370**
carboplatin	cisplatin	**0.87725**	0.85271	0.86052
xanax	ativan	0.72460	0.80707	**0.80830**
ethanol	alcohol	0.57237	0.61436	**0.62272**
	Average	0.67573	0.70962	**0.70994**

TABLE 5.2. comparison of cosine scores between Term 1 and Term 2 vectors from pyysalo et al., SMDB and SMDB+ word embeddings. Values in bold indicates the highest scores.

5.4. Text Classification

Using clinical questions dataset, text classification was performed to classify the questions into labelled topics. This questions dataset was prepared in accordance with Cao, et al. [10] for binary classification. The accuracies of SVM classifier trained on a set of domain-independent and domain-specific features as provided by Cao, et al. [10] were compared to the accuracies of CNN classifiers trained using vectors in SMDB and SMDB+ word embeddings as features. Cao et al. used a combination of BOW", "POS", "CSTY" and "BIGRAM" where "BOW" is a bag of words, "POS" is Part of Speech, and "CSTY" is Concept Semantic Type. Here SMDB classifier uses vectors of words in the text as input features, whereas SMDB+ classifier uses both vectors of the word and semantic type in the question text. To

Term 1	**Sem 1**	**Term 2**	**Sem 2**	**Pyysalo et al.**	**SMDB**	**SMDB+**
diabetes	dsyn	insulin	aapp	0.43701	0.34331	**0.43770**
meningitis	dsyn	headache	sosy	0.31292	0.32223	**0.51885**
nausea	sosy	zofran	orch	0.35318	**0.46599**	0.37232
hypothyroidism	dsyn	synthroid	aapp	0.27220	**0.32810**	0.26122
pain	sosy	morphine	orch	0.34102	0.44372	**0.54410**
diabetes	dsyn	polydipsia	sosy	0.27831	0.25547	**0.37107**
hyperemesis	sosy	zofran	orch	0.09222	0.19064	**0.20726**
diabetes	dsyn	polydipsia	sosy	0.30673	0.32209	**0.35027**
obesity	dsyn	snoring	sosy	0.37785	0.40361	**0.45213**
dyslipidemia	dsyn	lipitor	orch	0.22700	0.24698	**0.40581**
headache	sosy	tylenol	orch	**0.25602**	0.10493	0.25559
ataxia	sosy	ethanol	orch	0.03230	0.00496	0.**36580**
			Average	0.26851	0.28517	**0.37851**

TABLE 5.3. comparison of relatedness scores between Term 1 and Term 2 vectors from pyysalo et al., SMDB and SMDB+ word embeddings. Here Sem 1 and sem 2 are semantic types for term 1 and term 2 respectively. Values in bold indicates the highest scores. aapp(Amino Acid, Peptide, or Protein), dsyn(Disease or Syndrome), orch(Organic Chemical), sosy(Sign or Symptom) are semantic types defined in UMLS

get semantic types in question text we make use of Metamap. Table IV shows the accuracies comparison of SVM classifiers trained on different features to CNN classifiers trained using word embeddings as features. It is evident from the result that use of both words and semantic types(SMDB+) provided a 2.35% increase in classification accuracy compared to Cao, et al. [10], and a 1% increase in SMDB+ from SMDB classifier illustrates the effectiveness of using semantic types vectors as additional features to CNN model.

	BOW+ BIGRAM	**BOW+ BIGRAM +POS**	**BOW+ BIGRAM +CSTY**	**BOW+ BIGRAM +CSTY +POS**	**SMDB**	**SMDB+**
Pharmacological	86.99	87.22	87.81	87.97	90.35	**91.3**
Management	67.06	67.09	67.02	67.02	65.99	**68.51**
Diagnosis	76.82	76.61	76.97	77.13	77.19	**78.97**
Treatment & Prevention	71.09	71.20	71.26	71.03	74.59	**75.28**
Test	79.71	79.78	80.79	80.79	82.55	**83.98**
Average:	76.33	76.38	76.77	76.78	78.13	**79.60**

TABLE 5.4. Accuracy comparision of SVM classifier trained on a set of features i.e. "BOW+BIGRAM", "BOW+BIGRAM+POS", "BOW+BIGRAM+CSTY", "BOW+BIGRAM+CSTY+POS" as trained by Cao, et al. [10] to CNN classifiers trained using SMDB and SMDB+ word embeddings as features. Values in bold indicates the highest scores.

CHAPTER 6

CONCLUSION AND FUTURE WORK

This work takes a new and effective approach towards the training of biomedical word embeddings, utilizing semantic information in SemMedDB. The research does not aim to produce a new architecture for generating word embeddings but to check the effect of utilizing semantic inputs on the current state of the art models. In this research, a background of biomedical tools and resources useful for NLP tasks were provided. The research bridges the gap between the utilization of biomedical semantic resources and word embeddings. Using semantic information from SemMedDB semantic sentences were created, using these sentences, semantic types were introduced into the vector space of word embeddings. Skip-gram model was trained to build biomedical semantic embeddings from our new sentences. This research further illustrates the use of semantic type vectors in various tasks:

- Finding biomedical pairs such as drug-disease. Similarly, we could find relations between different biomedical pairs such as drug-target, symptom-disease, drug-side effects, etc.
- Calculating better relatedness scores using UMLS as standard for getting relations between biomedical terms.
- Improving text classification accuracy. Similarly, vectors of our semantic types could be used as additional input features wherever word vectors are used.

The introduction of semantic types in the vector spaces results in better similarity and relatedness between words compared to regular word embeddings that do not have semantic inputs. When an input is fed to the system the embedding utilizes the semantic type of the input to generate supportive and related text. The output of this research are two word embeddings SMDB(trained on SemMedDB sentences) and SMDB+(trained on SemMedDB+Semantic sentences), generated using the genism library. The SMDB is a regular word embedding whereas SMDB+ utilizes semantic types. In terms of semantic similarity, the performance of SMDB and SMDB+ was equivalent, but both of them performed

better than the model used by Pyysalo el al. [43]. This can be attributed to the fact that SemMedDB database contain only filtered sentences from PubMed which has subject-object pairs in them, while PubMed used by Pyysalo comprises of all the biomedical text. In terms of semantic relatedness, SMDB+ performed better compared to SMDB, reinforcing the idea that semantic sentences created as part of this research improves word embeddings. Accuracies of CNN classifiers trained using SMDB and SMDB+ were compared to the SVM classifiers used by Cao et al. The results displayed that SMDB+ has better accuracy, followed by SMDB and lastly SVM classifiers.

The research acts as a foundation and creates many opportunities for further improvement in Biomedical Natural Language Processing. The prospects of further study include:

- The addition of inputs from other clinical text resources (such as PubMed and Medline) to create more powerful word embeddings with a larger vocabulary and better performance accuracy.
- Analyse the effectiveness of other word embeddings model such as GloVe, CBOW, trained using our semantic sentences. Compare the performances of these models to provide the best possible solution to the problems in Biomedical NLP.
- Experimenting with different hyper-parameters such as vector dimension, window size to improve the accuracy of the embedding.

Two main approaches to prepare word embedding include using an embedding layer in the Neural Network and generating the embedding from scratch, which is the most popular choice utilized by major Deep Learning packages, and another approach is to use a pre-trained embedding and train the model with it as the base. The approaches when tested reveal that the training loss decays more rapidly in almost all the cases of the later on different training sets. Due to the lack of pre-trained word embeddings with semantic inputs, the word embedding was generated using the former method, however, is made publicly available for other researchers to utilize our semantic embeddings in their research.

REFERENCES

[1] Asma Ben Abacha and Pierre Zweigenbaum, *Means: A medical question-answering system combining nlp techniques and semantic web technologies*, Information processing & management 51 (2015), no. 5, 570–594.

[2] Saïd Abdeddaïm, Sylvestre Vimard, and Lina Fatima Soualmia, *The mesh-gram neural network model: Extending word embedding vectors with mesh concepts for umls semantic similarity and relatedness in the biomedical domain*, 2015.

[3] Enrique Alfonseca, Marco De Boni, José-Luis Jara-Valencia, and Suresh Manandhar, *A prototype question answering system using syntactic and semantic information for answer retrieval*, NIST SPECIAL PUBLICATION SP (2002), no. 250, 680–685.

[4] A R Aronson, *Metamap: Mapping text to the umls metathesaurus*, Bethesda, MD: NLM, NIH, DHHS (2006), 1–26.

[5] Alan R Aronson, *Effective mapping of biomedical text to the umls metathesaurus: the metamap program.*, Proceedings of the AMIA Symposium, American Medical Informatics Association, 2001, p. 17.

[6] Alan R Aronson and François-Michel Lang, *An overview of metamap: historical perspective and recent advances*, Journal of the American Medical Informatics Association 17 (2010), no. 3, 229–236.

[7] Sofia J Athenikos and Hyoil Han, *Biomedical question answering: A survey*, Computer methods and programs in biomedicine 99 (2010), no. 1, 1–24.

[8] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin, *A neural probabilistic language model*, Journal of machine learning research 3 (2003), no. Feb, 1137–1155.

[9] Olivier Bodenreider, *The unified medical language system (umls): integrating biomedical terminology*, Nucleic acids research 32 (2004), no. suppl_1, D267–D270.

[10] Yong-gang Cao, James J Cimino, John Ely, and Hong Yu, *Automatically extracting*

information needs from complex clinical questions, Journal of biomedical informatics 43 (2010), no. 6, 962–971.

[11] Yonggang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J Cimino, John Ely, and Hong Yu, *Askhermes: An online question answering system for complex clinical questions*, Journal of biomedical informatics 44 (2011), no. 2, 277–288.

[12] Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo, *How to train good word embeddings for biomedical nlp*, Proceedings of the 15th Workshop on Biomedical Natural Language Processing, 2016, pp. 166–174.

[13] Trevor Cohen and Dominic Widdows, *Embedding of semantic predications*, Journal of biomedical informatics 68 (2017), 150–166.

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

[15] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa, *Natural language processing (almost) from scratch*, Journal of Machine Learning Research 12 (2011), no. Aug, 2493–2537.

[16] Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza, *Medical semantic similarity with a neural language model*, Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (New York, NY, USA), CIKM '14, ACM, 2014, pp. 1819–1822.

[17] Donna M DAlessandro, Clarence D Kreiter, and Michael W Peterson, *An evaluation of information-seeking behaviors of general pediatricians*, Pediatrics 113 (2004), no. 1, 64–69.

[18] John W Ely, Jerome A Osheroff, Mark H Ebell, George R Bergus, Barcey T Levy, M Lee Chambliss, and Eric R Evans, *Analysis of questions asked by family doctors regarding patient care*, Bmj 319 (1999), no. 7206, 358–361.

[19] John W Ely, Jerome A Osheroff, Kristi J Ferguson, M Lee Chambliss, Daniel C Vinson,

and Joyce L Moore, *Lifelong self-directed learning using a computer database of clinical questions*, Journal of family practice 45 (1997), no. 5, 382–389.

[20] John W Ely, Jerome A Osheroff, Paul N Gorman, Mark H Ebell, M Lee Chambliss, Eric A Pifer, and P Zoe Stavri, *A taxonomy of generic clinical questions: classification study*, Bmj 321 (2000), no. 7258, 429–432.

[21] Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith, *Retrofitting word vectors to semantic lexicons*, Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Denver, Colorado), Association for Computational Linguistics, May–June 2015, pp. 1606–1615.

[22] Marcelo Fiszman, Thomas C Rindflesch, and Halil Kilicoglu, *Abstraction summarization for managing the biomedical research literature*, Proceedings of the HLT-NAACL workshop on computational lexical semantics, Association for Computational Linguistics, 2004, pp. 76–83.

[23] Boris Galitsky and Rajesh Pampapathi, *Can many agents answer questions better than one?*, First Monday 10 (2005), no. 1.

[24] José Luis Vicedo González and Antonio Ferrández Rodríguez, *A semantic approach to question answering systems.*, TREC, 2000.

[25] Dimitar Hristovski, Dejan Dinevski, Andrej Kastrin, and Thomas C Rindflesch, *Biomedical question answering using semantic relations*, BMC bioinformatics 16 (2015), no. 1, 6.

[26] Betsy L Humphreys and DA Lindberg, *The umls project: making the conceptual connection between users and the information they need.*, Bulletin of the Medical Library Association 81 (1993), no. 2, 170.

[27] Z. Jiang, L. Li, D. Huang, and Liuke Jin, *Training word embeddings for deep learning in biomedical text mining tasks*, 2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (Los Alamitos, CA, USA), IEEE Computer Society, nov 2015, pp. 625–628.

[28] Siddhartha Reddy Jonnalagadda, Guilherme Del Fiol, Richard Medlin, Charlene Weir, Marcelo Fiszman, Javed Mostafa, and Hongfang Liu, *Automatically extracting sentences from medline citations to support clinicians' information needs*, Journal of the American Medical Informatics Association 20 (2012), no. 5, 995–1000.

[29] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom, *A convolutional neural network for modelling sentences*, arXiv preprint arXiv:1404.2188 (2014).

[30] Halil Kilicoglu, Graciela Rosemblat, Marcelo Fiszman, and Thomas C Rindflesch, *Constructing a semantic predication gold standard from the biomedical literature*, BMC bioinformatics 12 (2011), no. 1, 486.

[31] Halil Kilicoglu, Dongwook Shin, Marcelo Fiszman, Graciela Rosemblat, and Thomas C Rindflesch, *Semmeddb: a pubmed-scale repository of biomedical semantic predications*, Bioinformatics 28 (2012), no. 23, 3158–3160.

[32] Yoon Kim, *Convolutional neural networks for sentence classification*, arXiv preprint arXiv:1408.5882 (2014).

[33] Tetsuya Kobayashi and Chi-Ren Shyu, *Representing clinical questions by semantic type for better classification*, AMIA Annual Symposium Proceedings, vol. 2006, American Medical Informatics Association, 2006, p. 987.

[34] Alberto Lavelli, Fabrizio Sebastiani, and Roberto Zanoli, *Distributional term representations: an experimental comparison*, Proceedings of the thirteenth ACM international conference on Information and knowledge management, ACM, 2004, pp. 615–624.

[35] Omer Levy and Yoav Goldberg, *Linguistic regularities in sparse and explicit word representations*, Proceedings of the Eighteenth Conference on Computational Natural Language Learning (Ann Arbor, Michigan), Association for Computational Linguistics, June 2014, pp. 171–180.

[36] Omer Levy and Yoav Goldberg, *Neural word embedding as implicit matrix factorization*, Advances in neural information processing systems, 2014, pp. 2177–2185.

[37] Xin Li and Dan Roth, *Learning question classifiers*, Proceedings of the 19th inter-

national conference on Computational linguistics-Volume 1, Association for Computational Linguistics, 2002, pp. 1–7.

[38] Alexa T McCray, *The umls semantic network.*, Proceedings. Symposium on Computer Applications in Medical Care, American Medical Informatics Association, 1989, pp. 503–507.

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

[40] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean, *Distributed representations of words and phrases and their compositionality*, Advances in neural information processing systems, 2013, pp. 3111–3119.

[41] Tomas Mikolov, Scott Wen-tau Yih, and Geoffrey Zweig, *Linguistic regularities in continuous space word representations*, Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013), Association for Computational Linguistics, May 2013.

[42] George A Miller, *Wordnet: a lexical database for english*, Communications of the ACM 38 (1995), no. 11, 39–41.

[43] SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou, *Distributional semantics resources for biomedical text processing*, Proceedings of the 5th International Symposium on Languages in Biology and Medicine, Tokyo, Japan, 2013, pp. 39–43.

[44] Mariana Neves and Milena Kraus, *Biomedlat corpus: Annotation of the lexical answer type for biomedical questions*, Proceedings of the Open Knowledge Base and Question Answering Workshop (OKBQA 2016), 2016, pp. 49–58.

[45] Arvid Österlund, David Ödling, and Magnus Sahlgren, *Factorization of latent variables in distributional semantic models*, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 227–231.

[46] Serguei Pakhomov, Bridget McInnes, Terrence Adam, Ying Liu, Ted Pedersen, and Genevieve B Melton, *Semantic similarity and relatedness between clinical terms: an*

experimental study, AMIA annual symposium proceedings, vol. 2010, American Medical Informatics Association, 2010, p. 572.

[47] Jeffrey Pennington, Richard Socher, and Christopher Manning, *Glove: Global vectors for word representation*, Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[48] Thomas C Rindflesch and Marcelo Fiszman, *The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text*, Journal of biomedical informatics 36 (2003), no. 6, 462–477.

[49] Alexandre Salle, Marco Idiart, and Aline Villavicencio, *Matrix factorization using window sampling and negative sampling for improved word representations*, arXiv preprint arXiv:1606.00819 (2016).

[50] Abeed Sarker and Graciela Gonzalez, *Portable automatic text classification for adverse drug reaction detection via multi-corpus training*, Journal of biomedical informatics 53 (2015), 196–207.

[51] Steffen Schulze-Kremer, Barry Smith, and Anand Kumar, *Revising the umls semantic network*, (2004).

[52] Arshad Shaik and Wei Jin, *Biomedical semantic embeddings: Using hybrid sentences to construct biomedical word embeddings and its applications*, 2019 IEEE International Conference on Healthcare Informatics (ICHI), 2019.

[53] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil, *Learning semantic representations using convolutional neural networks for web search*, Proceedings of the 23rd International Conference on World Wide Web, ACM, 2014, pp. 373–374.

[54] Richard Socher, Christopher D. Manning, and Andrew Y. Ng, *Learning continuous phrase representations and syntactic parsing with recursive neural networks*, In Proceedings of the NIPS-2010 Deep Learning and Unsupervised Feature Learning Workshop, 2010.

[55] Pontus Stenetorp, Hubert Soyer, Sampo Pyysalo, Sophia Ananiadou, and Takashi

Chikayama, *Size (and domain) matters: Evaluating semantic word space representations for biomedical text*, Proceedings of SMBM 12 (2012).

[56] Muneeb Th, Sunil Kumar Sahu, and Ashish Anand, *Evaluating distributed word representations for capturing semantics of biomedical concepts*, BioNLP@IJCNLP, 2015.

[57] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al., *An overview of the bioasq large-scale biomedical semantic indexing and question answering competition*, BMC bioinformatics 16 (2015), no. 1, 138.

[58] Joseph Turian, Lev Ratinov, and Yoshua Bengio, *Word representations: a simple and general method for semi-supervised learning*, Proceedings of the 48th annual meeting of the association for computational linguistics, Association for Computational Linguistics, 2010, pp. 384–394.

[59] Wang Weiming, Dawei Hu, Min Feng, and Liu Wenyin, *Automatic clinical question answering based on umls relations*, Third International Conference on Semantics, Knowledge and Grid (SKG 2007), IEEE, 2007, pp. 495–498.

[60] Zi Yang, Niloy Gupta, Xiangyu Sun, Di Xu, Chi Zhang, and Eric Nyberg, *Learning to answer biomedical factoid & list questions: Oaqa at bioasq 3b.*, CLEF (Working Notes), 2015.

[61] Wen-tau Yih, Xiaodong He, and Christopher Meek, *Semantic parsing for single-relation question answering*, Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), vol. 2, 2014, pp. 643–648.

[62] Zhiguo Yu, Trevor Cohen, Byron Wallace, Elmer Bernstam, and Todd Johnson, *Retrofitting word vectors of mesh terms to improve semantic similarity measures*, Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis, 2016, pp. 43–51.