6-19-2019

# Multi-Ontology Refined Embeddings (MORE): A Hybrid Multi-Ontology and Corpus-based Semantic Representation for Biomedical Concepts

Steven Jiang
*Dartmouth College*

# Multi-Ontology Refined Embeddings (MORE): A Hybrid Multi-Ontology and Corpus-based Semantic Representation for Biomedical Concepts

Steven Jiang [1] [2]

Dartmouth Computer Science Technical Report TR2019-873

June 19, 2019

[1] Saeed Hassanpour, Department of Biomedical Data Science, Geisel School of Medicine
[2] Department of Computer Science, Dartmouth College

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Abstract

### Objective

Currently, a major limitation for natural language processing (NLP) analyses in clinical applications is that a concept can be referenced in various forms across different texts. This paper introduces Multi-Ontology Refined Embeddings (MORE), a novel hybrid framework for incorporating domain knowledge from various ontologies into a distributional semantic model, learned from a corpus of clinical text. This approach generates word embeddings that are more accurate and extensible for computing the semantic similarity of biomedical concepts than previous methods.

### Materials and Methods

We use the RadCore and MIMIC-III free-text datasets for the corpus-based component of MORE. For the ontology-based component, we use the Medical Subject Headings (MeSH) ontology and two state-of-the-art ontology-based similarity measures. In our approach, we propose a new learning objective, modified from the Sigmoid cross-entropy objective function, to incorporate domain knowledge into the process for generating the word embeddings.

### Results and Discussion

We evaluate the quality of the generated word embeddings using an established dataset [6] of semantic similarities among biomedical concept pairs. We show that the similarity scores produced by MORE have the highest average correlation (60.2%), with the similarity scores being established by multiple physicians and domain experts, which is 4.3% higher than that of the word2vec baseline model and 6.8% higher than that of the best ontology-based similarity measure.

## Conclusion

MORE incorporates knowledge from biomedical ontologies into an existing distributional semantics model (i.e. word2vec), improving both the flexibility and accuracy of the learned word embeddings. We demonstrate that MORE outperforms the baseline word2vec model, as well as the individual UMLS-Similarity ontology similarity measures.

# Chapter 2

# Introduction

## 2.1 Problem Statement

With the increasing availability of health-related textual data, such as Electronic Health Records (EHR), novel applications of Natural Language Processing (NLP) in the field of medical informatics is a growing topic of interest [7, 8, 9]. Currently, a major limitation of NLP analysis techniques for clinical text is that due to the free-text format of these records and notes, the same concept can be referenced in various forms across different texts (e.g. "kidney failure" and "renal failure"). Particularly, different physicians and institutions may use unique terminologies for the same concepts for reporting in EHRs. In order to address this issue, researchers use semantic similarity measures to identify similar biomedical concepts in these free-text records and notes. A semantic similarity measure takes as input two concepts and returns a numeric score that quantifies how alike they are in meaning [6].

A hybrid biomedical semantic similarity measure would allow for more accurate clustering of concepts across a wider range of domains. Being able to accurately cluster groups of semantically similar biomedical concepts can improve patient-care and clinical outcomes. For example, patient health records can be analyzed to identify subjects with similar conditions or pathologies. With this information, data-mining techniques can be used to extract useful information about previous care processes, evolution of certain diseases, social trends, etc [10]. Semantic similarity measures can also assist in identifying patients for clinical studies and clustering "symptoms and disorders found in the text of clinical reports for postmarketing medication safety surveillance" [5]. Furthermore, semantic similarity measures can be used to integrate heterogeneous clinical data, which can improve interoperability between medical sources and allow hospitals to share patient health information more effectively [10]. Finally, in the fields of medical information retrieval and literature mining, where large amounts of electronic text data are available, keyword-based search engines "can be improved by extending user's queries to conceptually equivalent formulations using semantically similar terms" [10]. All in all, semantic similarity measures can improve the statistical power of NLP analyses, making it easier to identify associations between conditions and clinical outcomes in health records and improve information retrieval from scientific journals and clinical reports [5].

## 2.2 Overview of Existing Methods

A variety of semantic similarity measures have been developed to describe the strength of the relationships between concepts in biomedicine. These existing semantic similarity measures mostly fall into two common categories: ontology-based or corpus-based semantic similarities. Ontology-based semantic similarities typically rely on different graph-based features [11, 10], such as the shortest path length between concepts, the depth of the concepts in a hierarchy, and the position of their lowest common ancestors, to capture semantic similarity. As these ontology-based approaches are sensitive to the completeness and quality of the underlying ontologies [11], curating and maintaining domain ontologies is critical to guarantee the accuracy and robustness of ontology-based semantic similarities. Although there have been major efforts, such as the ongoing support by National Library of Medicine (NLM), to curate and maintain biomedical ontologies as valuable sources of domain knowledge, it is a labor-intensive and elaborate task. Furthermore, due to the heterogeneity of biomedical domains and their corresponding concepts, there is no single top-performing ontology-based similarity measure across all domains and applications [11, 12].

As an alternative to ontology-based semantic similarity, corpus-based semantic similarities are based on distributional semantics and co-occurrences of terms in free text [13, 14]. These corpus-based models rely on the linguistic principle that the meaning of a word (i.e. semantics) can be inferred based on its surrounding words (i.e. context). With recent advances in deep learning and the widespread use of distributional semantics to construct word embeddings for word representation in deep-neural networks, these corpus-based models have gained vast popularity. The word2vec [15] distributional semantics model is the most common method for generating such word embeddings. Intuitively, the word2vec model is a neural network that maps words with similar context to nearby points in a vector space. The cosine similarity between resulting word representations is commonly considered to be a corpus-based semantic similarity in various settings [16, 17, 18]. Although corpus-based semantic similarities are generated by unsupervised models, the lack of human curation and the availability of large, relevant biomedical corpora limit their accuracy and usability in biomedical applications [18, 10].

## 2.3 Proposed Solution

In this paper, we propose Multi-Ontology Refined Embeddings (MORE) semantic similarity to effectively integrate ontological knowledge and corpus-based context into a novel semantic similarity measure. MORE uses existing ontology-based semantic similarity measures from the Unified Medical Language System (UMLS) to modify the objective function of the word2vec skip-gram model, a popular distributional semantic model. In our approach, we propose a mathematical framework for vector representation refinement that relies on a collection of the most established and reliable ontology-based measures, rather than a single ontology-based similarity, to maximize the utility of our measure in a broad domain. In other words, MORE uses multiple ontology-based semantic similarities as the overall indicator of ontological similarity to refine the distributional semantic representations. Of note, our implementation is based on the official TensorFlow implementation of word2vec and we have made it available for public use [1].

Our model is benchmarked against existing state-of-the-art semantic similarities using an established evaluation dataset for semantic similarity between biomedical concepts. We find that MORE outperforms the baseline corpus-based semantic similarity model, as well as the individual ontology-based semantic similarities, in terms of correlation with Physician and Expert similarity scores in the evaluation dataset. The main contributions of this paper are two-fold: 1) we present a generalizable and extensible framework for incorporating domain-specific knowledge into a distributional semantic model and 2) we show that this hybrid framework outperforms the baseline word2vec model and ontology similarity measures on an established benchmark. In the remainder of this paper, we provide context for corpus-based, ontology-based, and hybrid semantic similarity measures in the biomedical domain. We also discuss the following components: the corpora used to train the corpus-based component of the model (i.e. RadCore and MIMIC-III), the UMLS-Similarity ontology measures used to modify the objective function, the mathematical framework for modifying the cross-entropy objective function, and the benchmark dataset against which the proposed method is evaluated. We also discuss the results from evaluating the proposed measure against state-of-the-art benchmarks, present a conclusion, and propose a direction for future research.

[1] https://github.com/BMIRDS/MORE

# Chapter 3

# Literature Review

## 3.1 Corpus-Based Methods

With recent advances in deep learning and the widespread use of distributional semantics to construct word embeddings for word representation, corpus-based models have become more useful for a variety of language modeling tasks. For instance, distributed representations of words in a vector space help learning algorithms to achieve better performance in natural language processing tasks by grouping similar words together [15]. These corpus-based methods are based on cooccurrences of terms in free text and rely on the linguistic principle that a word's meaning can be inferred from its surrounding words. The generated word embeddings are able to capture relationships between the words; for example, Mikolov et. al. found that "vec("Madrid") - vec("Spain") + vec("France") is closer to vec("Paris") than to any other word vector" [15]. This section of the paper discusses word2vec and GloVe, two state-of-the-art methods for generating word embeddings.

### 3.1.1 word2vec

Intuitively, the word2vec model is a neural network that maps words with similar context to nearby points in a vector space. It was initially developed as a way to learn high-quality word vectors from extremely large corpora, billions of words in length and millions of words in vocabulary size [1]. The original word2vec paper presented two model architectures for learning distributed representations of words that try to minimize computational complexity [1]. The first of these model architectures is the Continuous Bag-of-Words model (CBOW). As shown in Figure 1, the model uses both previous and future words as the input to correctly classify the current word. More specifically, the context words are used as input to a log-linear classifier and the goal is to correctly classify the current word. The model uses a continuous distributed representation of the context and, similar to standard the bag-of-words model, the order of the context words does not matter [1]. The second model architecture is the skip-gram model, which is also shown in Figure 1. The architecture is similar to that of the CBOW model; however, rather than predicting the current word from the context words, it predicts context words using the current word. More precisely, each current word is used as input to a log-linear classifier with continuous projection layer to predict words within a certain range before and after the current word [1].

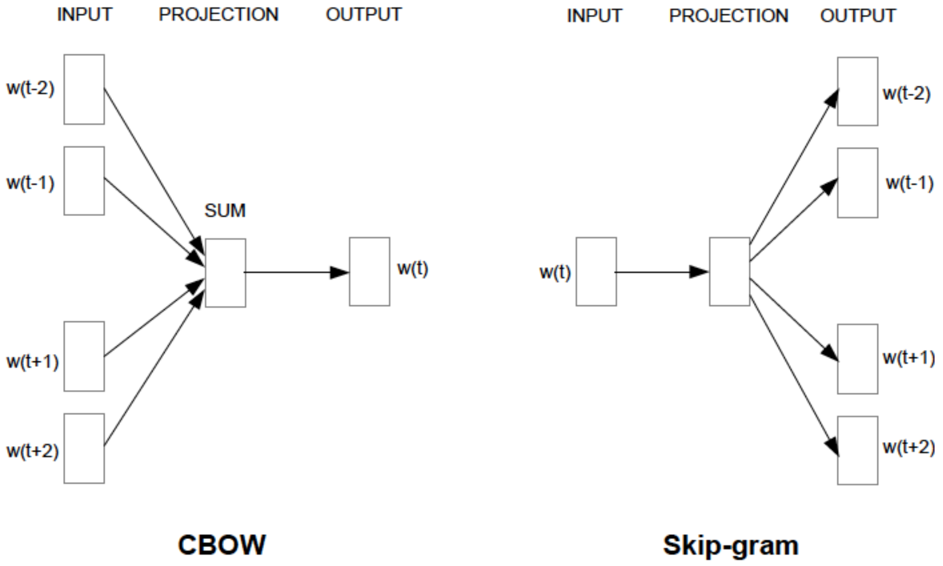


Figure 3.1: CBOW predicts the current word based on the context and Skip-gram predicts surrounding words given the current word [1]

The main benefit of these model architectures is that, unlike global matrix factorization methods, which rely on term co-occurrence matrices, the training processes for the word2vec models do not involve dense matrix multiplications. This makes the training extremely efficient, in that an optimized single-machine implementation can train on more than 100 billion words in one day [15].

### 3.1.2 GloVe

Introduced in 2014 by a group of researchers at Stanford, GloVe is another model for generating distributed representations of words in vector space. GloVe attempts to combine two major model families, global matrix factorization (i.e. latent semantic analysis) and local context window methods (i.e. word2vec). GloVe achieves this by efficiently incorporating global statistical information by training only on the nonzero elements in a term co-occurrence matrix, rather than on the entire sparse matrix or on individual context windows in a large corpus [19]. As a result, unlike the word2vec model, GloVe is able to make use of the global word occurrence statistics in a corpus as the primary source of information in learning word representations [19].

GloVe relies on a contextual co-occurrence matrix X, whose entries $X_{ij}$ represent the number of times word $j$ occurs in the context of word $i$. The overall complexity of the model depends on the number of nonzero elements in the matrix $X$. Since this number must be less than the total number of items in the matrix, the model scales no worse than $O(|V|^2)$, where $V$ is the size of the vocabulary [19]. With certain assumptions about the distributions of word co-occurrences in corpora, the authors argue that the complexity of the model is actually far better than $O(|V|^2)$. In fact, for the corpora studied in the article, they observed the complexity of the model was closer to $O(|C|^{0.8})$, where $C$ is the size of the corpus [19].

## 3.2 Ontology-Based Methods

Ontology-based methods rely on graph features of ontologies to compute semantic similarity between concepts. Most ontology-based methods calculate similarity between two concepts by using the location of the concepts in the ontology and the paths among them. Some of common methods rely on "edge counting, shortest path, and ontological depth, while others add the least common subsumer (LCS) to capture the granularity of a concept in the ontology." [20]. We are primarily interested in biomedical domain ontologies, which represent knowledge of medical concepts, such as the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) and Medical Subject Headings (MeSH). This project mainly relies on the UMLS-Similarity Perl package, which contains five semantic similarity measures: Rada, et. al. [14], Wu & Palmer (wup) [21], Leacock & Chodorow (lch) [22], and Nguyen & Al-Mubaid (nam) [23], and the Path measure (path).

### 3.2.1 Rada et al. (cdist)

Rada et al. [14] presented the Conceptual Distance (cdist) measure, which calculates semantic similarity between two concepts by counting the number of edges between them. The similarity scores outputted by this measure are in the range between zero and twice the depth of the taxonomy [5].

$$sim_{cdist}(c_1, c_2) = |shortest\_path(c_1, c_2)|$$

### 3.2.2 Wu and Palmer (wup)

Wu and Palmer [21] is a similarity measure that is based on the most specific concept that subsumes both of the concepts being measured [6]. In other words, it considers the position of concepts $c_1$ and $c_2$ in the ontology relative to the position of the most specific common concept $c$. Since there can be multiple parents for each concept, two concepts can share parents by multiple paths. The most specific common concept $c$ is the common parent related with the minimum number of "is-a" links with concepts $c_1$ and $c_2$ [24]. The similarity between two concepts is calculated as "twice the depth of the two concepts least common subsumer (LCS) divided by the product of the depths of the individual concepts" [5]. The similarity scores outputted by this measure fall in the range between zero and one [5].

$$sim_{wup}(c_1, c_2) = \frac{2 * depth(LCS)}{depth(c_1) + depth(c_2)}$$

### 3.2.3 Leacock and Chodorow (lch)

Leacock and Chodorow [22] is a path length based measure of semantic similarity. This measure is based on finding the length of the shortest path between two concepts in an ontology, dividing it by two times the maximum depth of the hierarchy, and taking the negative logarithm to get the resulting score. Its range is unbounded [5].

$$sim_{lch}(c_1, c_2) = -log(\frac{shortest\_path(c_1, c_2)}{2 * depth(ontology)})$$

### 3.2.4 Al-Mubaid and Nguyen (nam)

Al-Mubaid and Nguyen [23] is a similarity measure that is based on that the depth of the concept nodes and distance (path length) between them [23]. To compute the semantic similarity distance between two concepts, this measure takes the depth of their least common subsumer (LCS), and the distance of shortest path of between them. As a result, the method is able to assign "higher similarity when the two concepts are in a lower level of the hierarchy" [23]. The similarity between two concepts is calculated as "the log of two plus the product of the shortest distance between the two concepts minus one and the depth of the taxonomy minus the depth of the concepts LCS" [5]. The range of similarity scores outputted by this measure depends on the depth of the taxonomy [5].

$$sim_{nam}(c_1, c_2) = log((|shortest\_path(c_1, c_2)| - 1) * (depth(ontology) - depth(lcs(c_1, c_2))) + 2)$$

### 3.2.5 Path measure (path)

The Path measure (path) calculates the similarity between two concepts as the “reciprocal of the number of nodes between two concepts and its range is between zero and one” [5].

$$sim_{path}(c_1, c_2) = \frac{1}{|shortest\_path(c_1, c_2)|}$$

## 3.3 Hybrid Methods

As the name suggests, hybrid methods for computing semantic similarity combine elements from both corpus-based methods and ontology-based methods. There have been previous efforts to combine ontology-based and corpus-based similarities to better capture semantic similarities; however, there doesn’t currently exist a framework for incorporating ontological knowledge into the process of generating word embeddings for semantic similarity the biomedical domain.

### 3.3.1 Yu and Dredze (2014)

Yu and Dredze [25] introduced a general method for learning word embeddings by incorporating prior information. Their model extends the objective function of word2vec to include prior knowledge about synonyms from semantic resources. The authors also define a Relation Constrained Model (RCM) which is trained solely from the semantic resources. Finally, they define their combined objective function as a linear combination of the word2vec CBOW objective function and the RCM objective function. They show that the word embeddings, generated from the combined objective function and trained on a general corpus, outperforms the baseline word embeddings in three tasks: language modeling, measuring word similarity, and predicting human judgement on word pairs [25].

### 3.3.2 Xu et al. (2014)

Xu et al. [26] introduce RC-NET, a general framework for incorporating knowledge into word embeddings. First, The authors define two models, R-NET and C-NET, which use different objective functions to capture relational knowledge and categorical knowledge, respectively. Relational knowledge builds ”the global structure of the learned word representations by utilizing the relationship between different words” and categorical knowledge improves ”the local structure of the learned word representations by clustering similar words together” [26]. Relational knowledge and categorical knowledge complement each other because the absence relational knowledge can be made up for with the presence of categorical knowledge and vice versa. As a result, they combine these two separate objective functions with the original objective function of the skip-gram model to build the objective function used by RC-NET [26].

$$J = \alpha E_r + \beta E_c - L$$

where $E_r$ is the objective function R-NET, $E_c$ is the objective function of C-NET, and $L$ is objective function of the skip-gram model. They show that RC-NET, trained on a general corpus, outperforms R-NET, C-NET, and the baseline skip-gram model in the word similarity and topic prediction tasks.

### 3.3.3 Faruqui et al. (2014)

Faruqui et al. [27] propose a method for augmenting vector space representations of words using relational information from semantic lexicons [27]. The main contribution of their proposed method, retrofitting, is that it makes no assumptions about how the input vectors were constructed. In other words, retrofitting is applied as a post-processing step, allowing it to be used on any pre-trained word vectors [27]. The authors evaluate their proposed method on word embeddings generated from the five following models: Glove, skip-gram, Global Context, and Multilingual. Additionally, they use three following semantic ontologies: PPDB, WordNet, and FrameNet. They show that using retrofitting as a post-processing step improves performance on a variety of tasks, including word similarity, syntactic relations, synonym selection, and sentiment analysis [27].

### 3.3.4 Pivovarov and Elhadad (2012)

Pivovarov and Elhadad [20] present a hybrid score that uses a filtration and weighted average of ontology-based measures and corpus-based measures to calculate semantic similarity. Their method consists of three complementary similarity measures; one of the measures is corpus-based and relies on distributional semantics and the other two are ontology-based and rely on concept definitions and their relationships in the SNOMED-CT [20]. The pipeline works as follows. First, the corpus is preprocessed to extract concepts. Next, a filtration process prunes out the extracted concepts and keeps a homogeneous set of concepts to be aggregated. Then, the corpus-based similarity measure ranks all pairs of concepts. Finally, the top-k pairs with the highest context-based similarity are reordered using the two ontology-based similarity measures [20]. However, their proposed hybrid method is limited in that it doesn't incorporate ontological knowledge in the generation process of the word embeddings; rather, it uses corpus-based and ontology-based measures in a pipeline to produce a final semantic similarity score.

# Chapter 4

# Methods

## 4.1 Utilized Corpora

In this work, we use the RadCore and MIMIC-III corpora to train the corpus-based component of our proposed model. Assembled at Stanford in 2007, RadCore exists an effort to construct a large multi-institutional radiology report corpus for NLP [28]. The reports in the RadCore corpus range from 1995 to 2006 and were de-identified by their source organizations before submission to RadCore. In its entirety, RadCore contains 1,899,482 reports from three major healthcare organizations: Mayo Clinic (812 reports), MD Anderson Cancer Center (5000 reports), and Medical College of Wisconsin (1,893,670 reports) [28]. Additionally, all of the radiology reports are in free text format and do not contain any metadata about the type and nature of the imaging exams [28]. Medical Information Mart for Intensive Care (MIMIC-III) is a database containing information gathered from patients that were admitted to critical care units at a large hospital [29]. In this study, we use MIMIC-III's gold standard corpus of 2,434 ICU nursing notes that were "gathered simultaneously with the signals, trends, laboratory reports, discharge summaries and other data in the MIMIC-III databases" [30]. The corpus was thoroughly de-identified; all detected instances of Protected Health Information (PHI) were replaced by realistic surrogate data [30]. The final training corpus, which is a combination of the RadCore and MIMIC-III corpora, contains 195,101,383 total words, 145,274 unique words, and 43,232 unique frequent words with at least 5 occurrences in the corpora.

## 4.2 UMLS-Similarity

One of the challenges with having numerous medical domain ontologies is that these ontologies are typically developed independently of each other and rely on different standards, programming languages and interfaces to ontological resources [5]. The UMLS framework addresses this by creating a standard for medical ontologies. UMLS includes over 100 controlled medical ontologies, such as the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) and Medical Subject Headings (MeSH) [5]. It integrates these ontologies into the UMLS Metathesaurus by labeling concepts with Concept Unique Identifiers (CUIs) [5]. A CUI can refer to multiple concepts from the individual ontologies. In order to distinguish between concepts with the same CUI, each concept is also labeled with an Atomic Unique Identifier (AUI). For instance, the ”AUI Cold Temperature [A15588749] from MeSH and the AUI Low Temperature [A3292554] from SNOMED-CT are mapped to the CUI Cold Temperature [C0009264]” [5]. UMLS contains information about over ”1 million biomedical concepts and 5 million concept names from more than 100 incorporated controlled vocabularies and classifications (some in multiple languages) systems” [31].

UMLS-Interface is a Perl package that provides an API to a local installation of the UMLS in a MySQL database, allowing users to interactively explore the UMLS [5]. The corresponding Perl package, UMLS-Similarity, is used in conjunction with UMLS-Interface to provide an API to obtain the semantic similarity between CUIs in the UMLS. UMLS-Similarity contains five semantic similarity measures proposed by Rada, et. al. [14], Wu & Palmer [21], Leacock & Chodorow [22], and Nguyen & Al-Mubaid [23], and the Path measure [5].

# Chapter 5

# Proposed Method: MORE

Multi-Ontology Refined Embeddings (MORE) is a hybrid semantic similarity measure that effectively integrates ontological knowledge and corpus-based context in a novel semantic similarity measure. MORE uses a mathematical framework for vector representation refinement that relies on a collection of the most established and reliable ontology-based measures, rather than a single ontology-based similarity, to maximize our measure's utility in a broad domain. For the ontology-based component, MORE uses the MeSH ontology and various ontology-based semantic similarities measures within the UMLS-Similarity Perl package. Specifically, MORE uses two notable UMLS-based semantic similarities within the MORE framework: (1) Wu and Palmer [21] and (2) Leacock and Chodorow [22]. These ontology-based similarity measures are used to modify the objective function of the word2vec skip-gram model in order to refine a context-based similarity measure according to the domain ontologies. This framework is extensible because any number of ontology-based semantic similarly measures can be incorporated in the proposed semantic similarity framework.

## 5.1 Corpus Model: Skip-gram

### 5.1.1 Overview

In 2013, Mikolov et al. introduced the word2vec distributional semantics model, a neural network that maps words with similar context to nearby points in a vector space. The article introduced two model architectures: Continuous Bag-of-Words (CBOW) and skip-gram. While these models are algorithmically similar, CBOW predicts target words from context words, whereas skip-gram "does the inverse and predicts source context-words from the target words" [32] (See Figure 5.1). Statistically, CBOW smooths over a lot of the distributional information because it treats the entire context as one observation, which works better for smaller datasets. However, skip-gram treats each “context-target pair as a new observation, and this tends to do better when we have larger datasets” [32]. As a result, the skip-gram model is slower to train, but generalizes better to infrequent words [1].

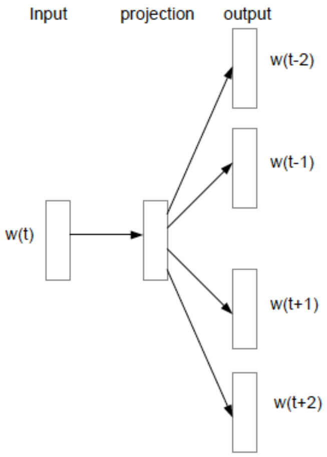
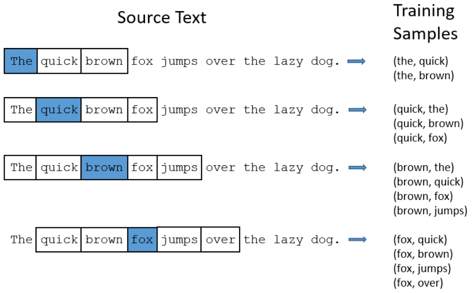


Figure 5.1: The objective of the skip-gram model is to learn word vector representations for predicting context words

The objective of the skip-gram model is to learn word vector representations for predicting context words (See Figure 5.2). Given an input word represented as a one-hot vector, the model looks at the context (i.e. nearby) words within a certain window size and picks one at random. The output of the network is a single vector containing, for every word in our vocabulary, “the probability that a randomly selected nearby word is that vocabulary word” [2]. For example, if the input word is “pancreatic”, we would expect a much higher output probability for words like “cancer” than other words like “brain” or “fracture.” Figure 5.3 depicts the skip-gram model architecture.



The quick brown fox jumps over the lazy dog. → (the, quick)<br>(the, brown)

The quick brown fox jumps over the lazy dog. → (quick, the)<br>(quick, brown)<br>(quick, fox)

The quick brown fox jumps over the lazy dog. → (brown, the)<br>(brown, quick)<br>(brown, fox)<br>(brown, jumps)

The quick brown fox jumps over the lazy dog. → (fox, quick)<br>(fox, brown)<br>(fox, jumps)<br>(fox, over)

Figure 5.2: Skip-gram training example with context window size of 2 [2].

Intuitively, if two words have similar contexts, then the network should output similar probability vectors for them. The model accomplishes this by learning similar vectors for similar words [2]. Thus, the cosine similarity between resulting word vectors is commonly considered to be a corpus-based semantic similarity measure.

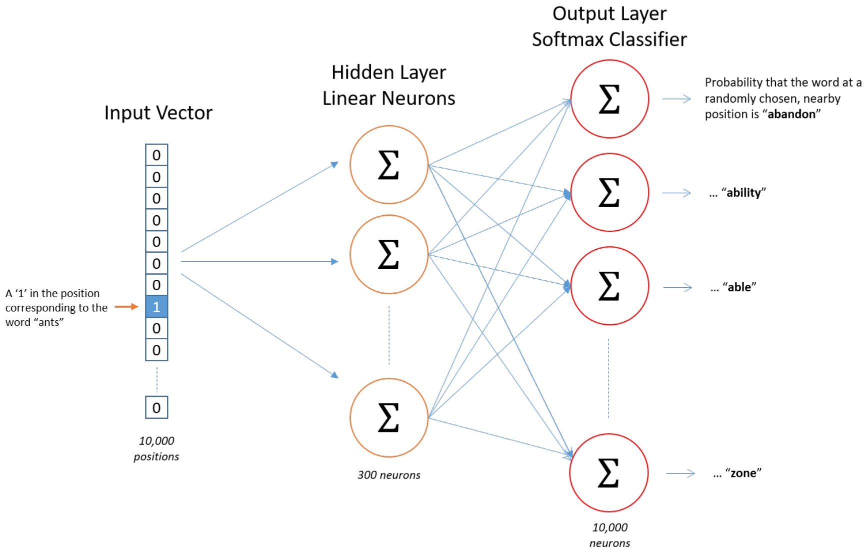


Figure 5.3: Skip-gram model architecture [2].

### 5.1.2 Subsampling of Frequent Words

In very large corpora, the most frequent words can occur "hundreds of millions of times (e.g., 'in', 'the', and 'a')" [15]. Since these words appear so often, they typically provide less information value than the less frequently occuring words. Intuitively, the skip-gram model benefits much more from observing co-occurrences of "France" and "Paris" than observing the frequent co-occurrences of "France" and "the", since there are far more other words that co-occur frequently with "the" [15]. As a result, the authors propose a subsampling method to address this issue. Through subsampling, for each word that is encountered in our training text, there is a chance that it will effectively be deleted it from the text [2]. The probability that any given word is deleted is proportional to the word's frequency. A word $w_i$ is discarded from the training set with the probability computed by the formula:

$$P(w_i) = 1 - \sqrt{\frac{t}{f(w_i)}}$$

where $f(w_i)$ is the frequency of word $w_i$ and $t$ is a selected threshold, typically around 10-5 [15]. The authors found that, even though this subsampling formula was chosen heuristically, "it accelerates learning and even significantly improves the accuracy of the learned vectors of the rare words" [15].

### 5.1.3 Negative Sampling

In the first iteration of the skip-gram model, training the neural network meant taking a training example backpropogating the loss to adjust all of the weights slightly, so that it predicts that training sample more accurately [2]. In other words, each training sample will tweak all of the weights in the neural network. However, with a large enough corpus, the size of the vocabulary would make the skip-gram neural network have a tremendous number of weights and, thus, incredibly inefficient to train. In the next iteration of the skip-gram model, Mikolov et al. [15] introduced negative sampling as a method to address this issue.

Negative sampling addresses this issue by having each training sample only “modify a small percentage of the weights, rather than all of them” [2]. When training the skip-gram model on a word pair, the label or target word is a one-hot vector. In other words, given an input word, the output neuron corresponding to the label word should output a 1. For all of the other words in the vocabulary, the corresponding output neurons should output a 0. With negative sampling, we update the weight for the neuron corresponding to the label word; however, instead of backpropogating the loss on all of the other words in the vocabulary, we randomly select small number of “negative” words, train the corresponding neurons to output a 0, and update the weights for only those neurons [2]. The negatively sampled words are selected using a unigram distribution, meaning more frequent words are more likely to be selected as negative samples. Through experimentation, Mikolov et al. found the following equation, for the probability of negatively sampling a word $w_i$, to perform best:

$$P(w_i) = \frac{f(w_i)^{\frac{3}{4}}}{\sum_{j=0}^{n}(f(w_j)^{\frac{3}{4}})}$$

where $f(w_i)$ is the frequence of word $w_i$ in the corpus [2].

## 5.2 Ontology Measures

As mention previously, the UMLS-Similarity contains five semantic similarity measures proposed by Rada, et. al. [14], Wu & Palmer (wup) [21], Leacock & Chodorow (lch) [22], and Nguyen & Al-Mubaid (nam) [23], and the Path measure (path). In this study, we use the Wu & Palmer (wup) and Leacock & Chodorow (lch) semantic similarity measures on concepts in the MeSH ontology. In this study, we use the Wu & Palmer (wup) and Leacock & Chodorow (lch) semantic similarity measures on concepts in the MeSH ontology:

$$sim_{wup}(c_1, c_2) = \frac{2 * depth(LCS)}{depth(c_1) + depth(c_2)} \tag{5.1}$$

$$sim_{lch}(c_1, c_2) = -log(\frac{shortest\_path(c_1, c_2)}{2 * depth(ontology)}) \tag{5.2}$$

In order to use the ontology similarities to modify the objective function of the skip-gram model, we first identified the set of words that appear the intersection of the set of words in the corpus vocabulary and the set of words that exist in the MeSH ontology. Using this intersection set, we generate a similarity matrix containing all pair-wise similarities of the intersection terms, normalizing each measure to be in the range of 0 to 1. It's important to note that, given the relative positions of the words in the ontology, not every pair of words in the ontology has a similarity score as defined by the ontology similarity measures. For each word pair, if multiple ontology similarity measures produce scores, we compute the median of the similarity scores for the final matrix similarity score. If a single ontology measure produces a score, we use it as the final matrix similarity score. And, if there doesn't exist a similarity score as defined by any of the ontology similarity measures, we use a placeholder value of $-1$ to denote that only the skip-gram model output will be used in the training process. The final similarity matrix contains 4,878 unique words and 11,945,574 pair-wise similarity scores.

## 5.3 Modifying the Objective Function

We can extend the basic formulation of the word2vec model using knowledge from ontologies by adjusting the similarities outputted by the neural network. In determining the conditional probability of context words given the input word, the neural network relies solely on the co-occurrences of the corresponding terms in the corpus. Given the vocabulary size V, we are about to learn word embedding vectors of size N. The model learns to predict a context word using a target word (See Figure 5.4). Both the input word $w_i$ and the output word $w_j$ are one-hot encoded into binary vectors $x$ and $y$ of size $V$. Multiplying the vector $x$ and the word embedding matrix $W$ gives the embedding vector of the input word $w_i$. This embedding vector of dimension $N$ becomes part of the hidden layer. Next, multiplying the hidden layer and the word context matrix $W'$ produces the output one-hot encoded vector $y$. It's important to note that the output context matrix W' encodes the meanings of words as context, which is different from the embedding matrix W [3].

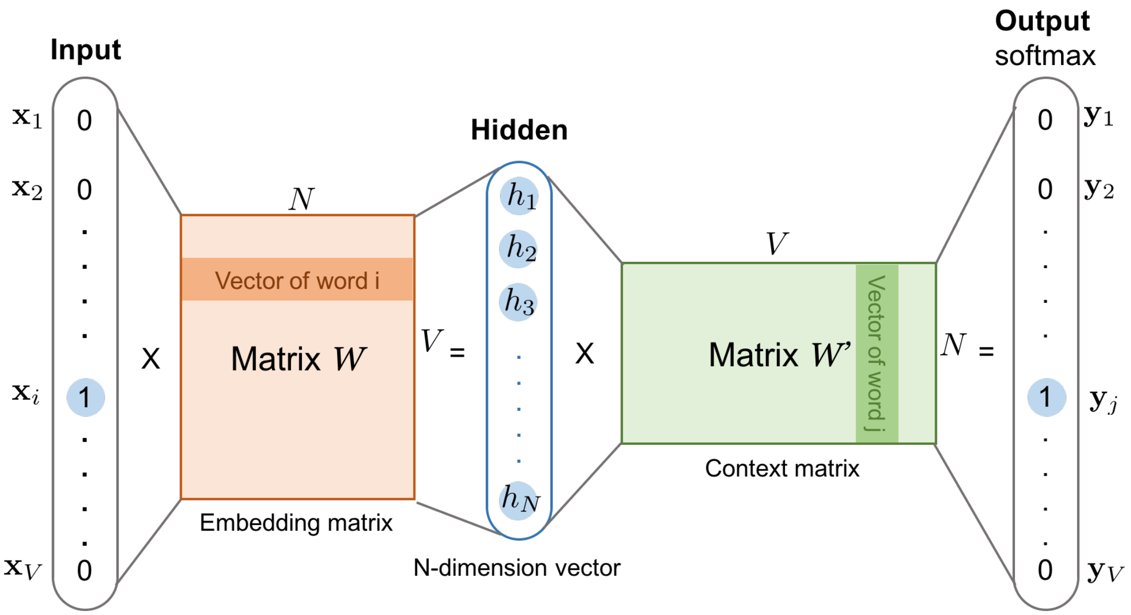


Figure 5.4: The input vector $x$ and the output vector $y$ are one-hot encoded word representations. The hidden layer is the word embedding of size N [3].

Traditionally, the objective function of the skip-gram model is a full softmax function. However, the specific implementation of the skip-gram model used for this project relies on a simplified variant of Noise Contrastive Estimation (NCE) [33] for training the skip-gram model that "results in faster training and better vector representations for frequent words, compared to more complex hierarchical softmax that was used in the prior work" [1]. The loss function $L_\theta$ is the average sigmoid cross entropy loss, which incorporates both the loss computed from the context words $L_{POS}$ and the loss computed from the negatively sampled words $L_{NEG}$, over the batch size:

$$L_\theta = \frac{\sum L_{POS} + \sum L_{NEG}}{BatchSize} \tag{5.3}$$

$$L_{POS} = -log(S(logit(w_C|w_I))) \tag{5.4}$$

$$L_{NEG} = -log(1 - S(logit(w_{NEG}|w_I))) \tag{5.5}$$

where $S$ is the Sigmoid function, $w_I$ is the input word, $w_C$ is a context word, $w_{NEG}$ is a negatively sampled word, and $logit(w_i|w_I)$ is the log odds of the conditional probability of the label word ($w_C$ or $w_{NEG}$) given the input word, as predicted by the model.

### 5.3.1 Modified Cross-Entropy Loss

where $S$ is the sigmoid function, $w_I$ is the input word, $w_C$ is a context word, $w_{NEG}$ is a negatively sampled word, and $logit(w_i|w_I)$ is the log odds of the conditional probability of the label word ($w_C$ or $w_{NEG}$) given the input word, as predicted by the model.

In computing the loss for the context words and negatively sampled words, we modify the binary labels used in the traditional cross-entropy loss function to incorporate the ontology similarities. For the context words, rather than multiplying the negative log of the sigmoid of the model output by one, we multiply it by the average of 1 and the ontology similarity score (Equation 5.6). Similarly, for the negatively sampled words words, rather than multiplying the negative log of the sigmoid of the model output by one minus zero, we multiply it by one minus the average of zero and the ontology similarity score (Equation 5.7).

$$L_{POS*} = \frac{1 + sim_{ont}(w_C, w_I)}{2} * -log(S(logit(w_C|w_I))) \tag{5.6}$$

$$L_{NEG*} = (1 - \frac{sim_{ont}(w_{NEG}, w_I)}{2}) * -log(1 - S(logit(w_{NEG}|w_I))) \tag{5.7}$$

By averaging the binary labels (i.e. 1 and 0) with the similarity scores outputted by the model, the loss function is adjusted to incorporate relational knowledge from the ontologies. For instance, in the case of the computing the loss for context words ($L_{POS*}$), if the word pair has high ontology similarity score, the loss will be higher. Conversely, if the word pair has a low ontology similarity score, the loss will be lower. As a result, in order to minimize the loss, the neural network will adjust the neurons' weights in the direction suggested by the ontological knowledge, encouraging the model to output higher probabilities for word pairs with high ontology similarity scores and lower probabilities for word pairs with low ontology similarity scores.

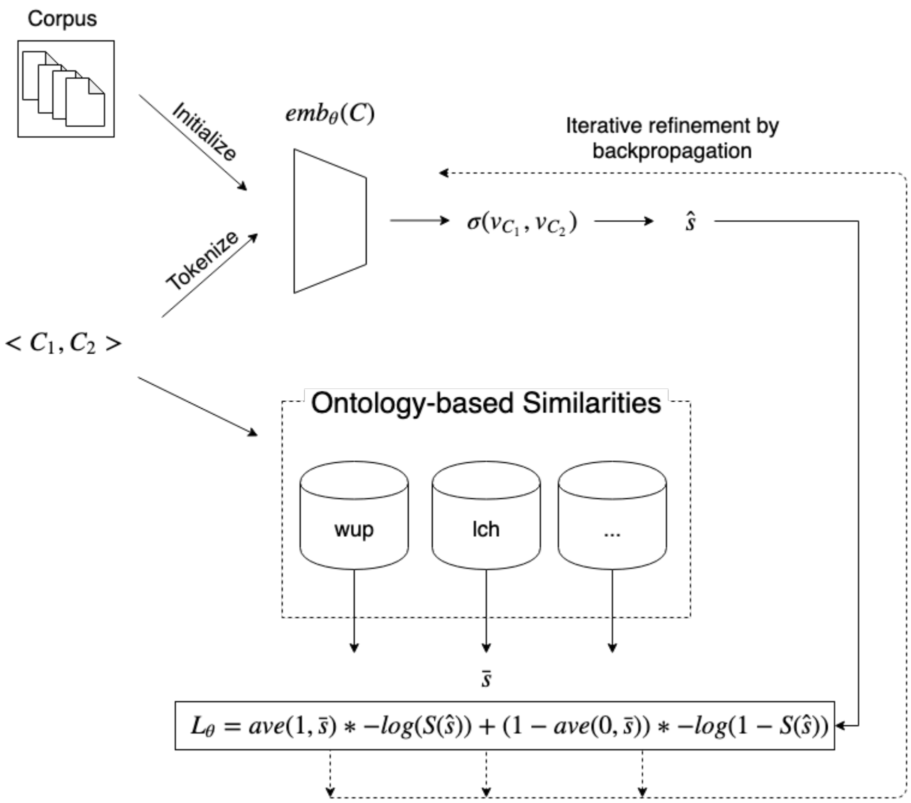

Corpus
Initialize
Tokenize
$emb_\theta(C)$
Iterative refinement by backpropagation
$\sigma(v_{C_1}, v_{C_2})$
$\hat{s}$
$< C_1, C_2 >$
Ontology-based Similarities
wup
lch
...
$\bar{s}$
$L_\theta = ave(1, \bar{s}) * -log(S(\hat{s})) + (1 - ave(0, \bar{s})) * -log(1 - S(\hat{s}))$

Figure 5.5: Overview of Multi-Ontology Refined Embeddings (MORE) framework. In training, the similarity between two concepts $(C_1, C_2)$ is measured in different ways: by cosine similarity $\sigma(v_{C_1}, vC_2)$ in a vector space, which gives the skip-gram model's output for the word pair $\hat{s}$, and through ontology-based similarity scores $\bar{s}$. $\bar{s}$ is the median of different ontology-based similarities. The network optimizes the parameters of the skip-gram model by minimizing the modified loss function and backpropagating the loss to refine the embedding layer. The semantic similarity scores are computed as the cosine similarity of the resulting word embeddings.

# Chapter 6

# Evaluation and Results

## 6.1 Evaluation

In 2007, Pedersen et al. [6] introduced a test set of word pairs for the evaluation of measures of semantic similarity and relatedness in the biomedical domain. The 30 concept pairs of medical terms (See Table 1) were scored by multiple physicians and domain experts on a 4-point scale, according to their relatedness: “practically synonymous (4.0), related (3.0), marginally related (2.0) and unrelated (1.0)” [6]. The average correlation between physicians was 0.68, the average correlation between experts was 0.78, and the correlation across groups was 0.85 [6]. In this study, term pair 5, “Delusion — Schizophrenia”, has been excluded from the final evaluation dataset because one of the terms did not appear a minimum of five times in our combined corpora. As a result, the resulting test set consists of 29 of the 30 original pairs. To evaluate the different measures, we calculate the correlation between the similarity scores outputted by the measure and the Physician/Expert similarity scores.

| **Concept 1** | **Concept 2** | **Phys.** | **Expert** |
|---|---|---|---|
| Renal failure | Kidney failure | 4.0000 | 4.0000 |
| Heart | Myocardium | 3.3333 | 3.0000 |
| Stroke | Infarct | 3.0000 | 2.7778 |
| Abortion | Miscarriage | 3.0000 | 3.3333 |
| Delusion | Schizophrenia | 3.0000 | 2.2222 |
| Congestive heart failure | Pulmonary edema | 3.0000 | 1.4444 |
| Metastasis | Adenocarcinoma | 2.6667 | 1.7778 |
| Calcification | Stenosis | 2.6667 | 2.0000 |
| Diarrhea | Stomach cramps | 2.3333 | 1.3333 |
| Mitral stenosis | Atrial fibrillation | 2.3333 | 1.3333 |

Table 6.1: First 10 pairs of evaluation dataset [4]

| Medical Term Pair | | CUIs | | SNOMED-CT | | | | MeSH | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Path | | LCH | | Path | | LCH | | WUP | | NAM | |
| Term 1 | Term 2 | CUI1 | CUI2 | score | rank | score | rank | score | rank | score | rank | score | rank | score | rank |
| Renal failure | Kidney failure | C0035078 | C0035078 | 1.00 | 1 | 4.22 | 1 | 1.00 | 1 | 3.64 | 1 | 1.00 | 1 | 1.00 | 1 |
| Abortion | Miscarriage | C0156543 | C0000786 | 0.50 | 2 | 3.53 | 2 | 0.10 | 20 | 1.34 | 20 | 0.47 | 20 | 7.10 | 20 |
| Heart | Myocardium | C0018787 | C0027061 | 0.25 | 3 | 2.83 | 3 | 0.50 | 2 | 2.94 | 2 | 0.93 | 2 | 3.81 | 2 |
| Metastasis | Adenocarcinoma | C0027627 | C0001418 | 0.25 | 3 | 2.83 | 3 | 0.14 | 7 | 1.69 | 7 | 0.63 | 9 | 6.43 | 8 |
| Pulmonary fibrosis | Lung cancer | C0034069 | C0242379 | 0.25 | 3 | 2.83 | 3 | 0.33 | 3 | 2.54 | 3 | 0.89 | 3 | 4.70 | 3 |
| Brain tumor | Intracranial hemorrhage | C0006118 | C0151699 | 0.25 | 3 | 2.83 | 3 | 0.25 | 4 | 2.25 | 4 | 0.88 | 4 | 5.13 | 4 |
| Rheumatoid arthritis | Lupus | C0003873 | C0409974 | 0.20 | 7 | 2.61 | 7 | 0.11 | 12 | 1.44 | 12 | 0.56 | 13 | 6.83 | 13 |
| Pulmonary embolus | Myocardial infarction | C0034065 | C0027051 | 0.17 | 8 | 2.43 | 8 | | | | | | | | |
| Antibiotic | Allergy | C0003232 | C0020517 | 0.17 | 8 | 2.43 | 8 | 0.10 | 20 | 1.34 | 20 | 0.47 | 20 | 7.10 | 20 |
| Depression | Cellulitis | C0011581 | C0007642 | 0.17 | 8 | 2.43 | 8 | 0.10 | 20 | 1.34 | 20 | 0.47 | 20 | 7.10 | 20 |
| Diarrhea | Stomach cramps | C0011991 | C0344375 | 0.14 | 11 | 2.27 | 11 | | | | | | | | |
| Multiple sclerosis | Psychosis | C0026769 | C0033975 | 0.14 | 11 | 2.27 | 11 | 0.10 | 20 | 1.34 | 20 | 0.47 | 20 | 7.10 | 20 |
| Mitral stenosis | Atrial fibrillation | C0026269 | C0004238 | 0.14 | 11 | 2.27 | 11 | 0.20 | 5 | 2.03 | 5 | 0.78 | 5 | 5.64 | 5 |
| Congestive heart failure | Pulmonary edema | C0018802 | C0034063 | 0.14 | 11 | 2.27 | 11 | 0.14 | 7 | 1.69 | 7 | 0.67 | 6 | 6.32 | 7 |
| Lymphoid hyperplasia | Laryngeal cancer | C0333997 | C0007107 | 0.13 | 15 | 2.14 | 15 | 0.14 | 7 | 1.69 | 7 | 0.67 | 6 | 6.43 | 8 |
| Diabetes mellitus | Hypertension | C0011849 | C0020538 | 0.13 | 15 | 2.14 | 15 | 0.17 | 6 | 1.85 | 6 | 0.67 | 6 | 6.17 | 6 |
| Carpal tunnel syndrome | Osteoarthritis | C0007286 | C0029408 | 0.13 | 15 | 2.14 | 15 | 0.11 | 12 | 1.44 | 12 | 0.56 | 13 | 6.83 | 13 |
| Xerostomia | Alcoholic cirrhosis | C0043352 | C0023891 | 0.11 | 18 | 2.02 | 18 | 0.11 | 12 | 1.44 | 12 | 0.59 | 12 | 6.73 | 12 |
| Peptic ulcer disease | Myopia | C0030920 | C0027092 | 0.11 | 18 | 2.02 | 18 | 0.14 | 7 | 1.69 | 7 | 0.63 | 9 | 6.43 | 8 |
| Appendicitis | Osteoporosis | C0003615 | C0029456 | 0.11 | 18 | 2.02 | 18 | 0.11 | 12 | 1.44 | 12 | 0.56 | 13 | 6.83 | 13 |
| Hyperlipidemia | Metastasis | C0020473 | C0027627 | 0.11 | 18 | 2.02 | 18 | 0.11 | 12 | 1.44 | 12 | 0.56 | 13 | 6.83 | 13 |
| Cortisone | Total knee replacement | C0010137 | C0086511 | 0.09 | 22 | 1.82 | 22 | 0.08 | 24 | 1.15 | 24 | 0.42 | 24 | 7.38 | 24 |
| Acne | Syringe | C0702166 | C0039142 | 0.08 | 23 | 1.65 | 23 | 0.11 | 12 | 1.44 | 12 | 0.50 | 18 | 6.93 | 18 |
| Stroke | Infarct | C0038454 | C0021308 | 0.07 | 24 | 1.58 | 24 | 0.11 | 12 | 1.44 | 12 | 0.56 | 13 | 6.83 | 13 |
| Varicose vein | Entire knee meniscus | C0042345 | C0224701 | 0.07 | 24 | 1.58 | 24 | | | | | | | | |
| Rectal polyp | Aorta | C0034887 | C0003483 | 0.07 | 24 | 1.58 | 24 | | | | | | | | |
| Delusion | Schizophrenia | C0011253 | C0036341 | 0.07 | 27 | 1.51 | 27 | 0.13 | 11 | 1.56 | 11 | 0.63 | 9 | 6.64 | 11 |
| Cholangiocarcinoma | Colonoscopy | C0206698 | C0009378 | 0.07 | 27 | 1.51 | 27 | 0.07 | 25 | 1.00 | 25 | 0.38 | 25 | 7.62 | 25 |
| Calcification | Stenosis | C0175895 | C0009814 | 0.00 | 29 | 0.00 | 29 | 0.11 | 12 | 1.44 | 12 | 0.50 | 18 | 6.93 | 18 |

Table 6.2: UMLS-Similarity Results [5].

## 6.2 Results

In this section, we compare the proposed model against four established ontology similarity measures from UMLS-Similarity and the baseline skip-gram model. The correlation values for the UMLS-Similarity ontology measures are from McInnes et al. [5]. The goal of the experiments is to demonstrate the value of using the MORE framework to learn semantic embeddings with information from ontology similarity measures. In each experiment, we compare the baseline embeddings trained with skip-gram against the embeddings trained using the MORE framework. We quantify the evaluation task of measuring semantic similarity using the correlation between the similarity scores generated by the embeddings and the similarity scores produced by the Physicians and Experts.

In training the baseline skip-gram model and the proposed model, we used the following default parameters of the tensorflow implementation of the skip-gram model: embedding size of 300, window size of 10, minimum word count of 5, and a subsampling threshold of 0.001. In order to expedite the training process, we used a learning rate of 0.3 and a batch size of 1024. We trained each model for 10 epochs at a time, warm starting each model with the previous model as a checkpoint, for a total of 150 epochs. Table 6.3 shows a comparison of the best results achieved by all of the models and ontology measures. Figures 6.1 and 6.2 show the correlations between the similarity scores outputted by the models and Physician and Expert similarity scores, respectively, over 150 training epochs.

| Measure | Phys. | Rank | Coder | Rank | Ave. Corr. | Rank |
|---|---|---|---|---|---|---|
| path | 0.486 | 3 | **0.581** | **1** | 0.534 | 3 |
| lch* | 0.486 | 3 | **0.581** | **1** | 0.534 | 3 |
| wup* | 0.453 | 5 | 0.535 | 5 | 0.494 | 6 |
| nam | 0.448 | 6 | 0.551 | 3 | 0.500 | 5 |
| Baseline | 0.612 | 2 | 0.506 | 6 | 0.559 | 2 |
| MORE | **0.662** | **1** | 0.542 | 4 | **0.602** | **1** |

Table 6.3: Similarity correlations of ontology-based measures, baseline model, and MORE. The asterisk denotes that the ontology-based measure was used in the proposed models

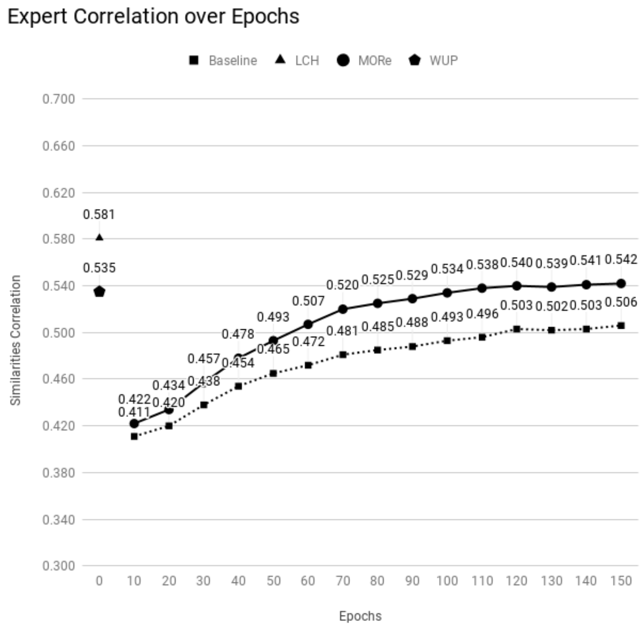


Figure 6.1: Comparison of Baseline and MORE correlations with Expert similarities

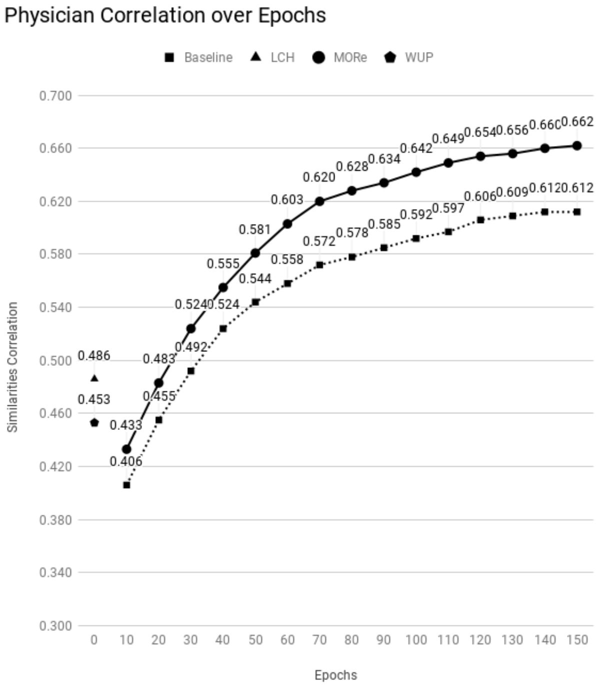


Figure 6.2: Comparison of Baseline and MORE correlations with Physician similarities

## 6.3 Discussion

We find that, under identical training conditions, MORE consistently outperforms the baseline skip-gram model in terms of correlation with Physician similarity scores and correlation with Expert similarity scores. Figures 6.1 and 6.2 show that, after 90 to 100 training epochs, the correlations with the Physician and Expert similarity scores begin to plateau and the differential between the baseline model and MORE remains relatively constant. Table 6.3 illustrates that, after training both models for 150 epochs, MORE has a 5% higher correlation with the Physician similarity scores and a 3.6% higher correlation with the Expert similarity scores than the baseline model. Additionally, MORE has a 17.6% higher correlation with the Physician similarity scores than the best ontology similarity measures (path and lch). However, we also find that MORE has a 3.9% lower correlation with the Expert similarity scores than the best ontology measures (path and lch). It's possible that since our corpora comprised of physician/nurse notes, the model had a higher correlation with the Physician similarity scores than with the Expert similarity scores. The fourth column in Table 6.3 displays the average correlation, which is simply an average of the correlation with Physician similarity scores and the correlation with the Expert similarity scores. We use the average of the correlations as a proxy for generalizablilty in our comparison. We show that MORE has the highest average correlation of 60.2%, which is 4.3% higher than the average correlation of the baseline model and 6.8% higher than the best ontology measures (path and lch).

As mentioned in the Introduction section, due to the heterogeneity of biomedical concepts, there is no single top-performing corpus-based or ontology-based semantic similarity measure across all applications and domains. However, by modifying the objective function of the skip-gram model with knowledge from the MeSH ontology and multiple UMLS similarity measures, we can generate embeddings from the RadCore and MIMIC-III corpora that incorporate knowledge beyond the scope of the corpora and maximize the measure's utility in a broad domain. MORE outperforms the baseline skip-gram model in every case, as well as the ontology similarity measures in most cases. As a result, we have demonstrated that the embeddings generated using the MORE framework are more effective at capturing semantic similarity for biomedical concepts, in a broader domain, than any of MORE's individual components.

Despite MORE's promising performance in our evaluation, we recognize that this study has several limitations. First, aside from the increased learning rate and batch size used to expedite training, we used the default training parameters, as suggested by the TensorFlow implementation of the skip-gram model, to train both the baseline skip-gram model and the proposed model. While these parameters have been optimized for training the baseline model, we did not experiment with tuning hyperparameters to optimize the training of the proposed model. However, this suggests that, under equal but potentially sub-optimal training conditions, MORE outperforms the baseline skip-gram model. Second, we have only incorporated two ontology similarity measures (lch and wup) from one ontology (MeSH) into our novel framework. With a broader range of similarity measures and more ontologies, such as SNOMED-CT, it's possible that MORE could generate embeddings that are more generalizable and accurate than those produced by the present work. Finally, in this study, we only evaluate the quality of the generated word embeddings with a semantic similarity task on relatively a small dataset.

To address these limitations, in future work, we plan to experiment by tuning different training parameters (e.g. learning rate, number of training epochs, and batch size). Furthermore, we plan to extend the model by incorporating more ontologies, such as SNOMED-CT, and other ontology-based similarity measures. Finally, we expect that the proposed framework has further implications beyond semantic similarity. Accordingly, in future work, we plan to evaluate the quality of the MORE embeddings on other semantic tasks, such as analogical reasoning, text classification, synonym selection, and topic modeling.

# Chapter 7

# Conclusion

Learning high-quality word embeddings for semantic similarity in the biomedical domain is valuable for improving the statistical power of NLP analyses, thus making it easier to identify associations between conditions and clinical outcomes in health records and improve information retrieval from scientific journals and clinical reports. To address existing limitations of biomedical semantic similarity measures, we propose a new modified objective function that incorporates domain knowledge into the process for generating word embeddings. In this paper, we presented a novel framework for incorporating knowledge from biomedical ontologies into an existing distributional semantic model (i.e. skip-gram) to improve both the flexibility and accuracy of the learned word embeddings. Our implementation is based on the official TensorFlow implementation of word2vec and we have made it available for public use. We demonstrate that MORE generally outperforms the baseline skip-gram model, as well as the individual UMLS ontology similarity measures, in computing semantic similarity scores for biomedical word pairs using a benchmark evaluation dataset.

# Bibliography

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

[2] Chris McCormick. Word2vec tutorial-the skip-gram model, 2016.

[3] Lilian Weng. Learning word embedding, Oct 2017.

[4] Ted Pedersen, Serguei Pakhomov, and Siddharth Patwardhan. Measures of semantic similarity and relatedness in the medical domain. *University of Minnesota digital technology center research report DTC*, 12:2005, 2005.

[5] Bridget T McInnes, Ted Pedersen, and Serguei VS Pakhomov. Umls-interface and umls-similarity: open source software for measuring paths and semantic similarity. In *AMIA Annual Symposium Proceedings*, volume 2009, page 431. American Medical Informatics Association, 2009.

[6] Ted Pedersen, Serguei VS Pakhomov, Siddharth Patwardhan, and Christopher G Chute. Measures of semantic similarity and relatedness in the biomedical domain. *Journal of biomedical informatics*, 40(3):288–299, 2007.

[7] W Katherine Tan, Saeed Hassanpour, Patrick J Heagerty, Sean D Rundell, Pradeep Suri, Hannu T Huhdanpaa, Kathryn James, David S Carrell, Curtis P Langlotz, Nancy L Organ, et al. Comparison of natural language processing rules-based and machine-learning systems to identify lumbar spine imaging findings related to low back pain. *Academic radiology*, 25(11):1422–1432, 2018.

[8] Hannu T Huhdanpaa, W Katherine Tan, Sean D Rundell, Pradeep Suri, Falgun H Chokshi, Bryan A Comstock, Patrick J Heagerty, Kathryn T James, Andrew L Avins, Srdjan S Nedeljkovic, et al. Using natural language processing of free-text radiology reports to identify type 1 modic endplate changes. *Journal of digital imaging*, 31(1):84–90, 2018.

[9] Saeed Hassanpour, Graham Bay, and Curtis P Langlotz. Characterization of change and significance for clinical findings in radiology reports through natural language processing. *Journal of digital imaging*, 30(3):314–322, 2017.

[10] David Sanchez and Montserrat Batet. Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective. *Journal of biomedical informatics*, 44(5):749–759, 2011.

[11] Montserrat Batet, David Sánchez, and Aida Valls. An ontology-based measure to compute semantic similarity in biomedicine. *Journal of biomedical informatics*, 44(1):118–125, 2011.

[12] Hisham Al-Mubaid and Hoa A Nguyen. A cluster-based approach for semantic similarity in the biomedical domain. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 2713–2717. IEEE, 2006.

[13] Catia Pesquita, Daniel Faria, Andre O Falcao, Phillip Lord, and Francisco M Couto. Semantic similarity in biomedical ontologies. *PLoS computational biology*, 5(7):e1000443, 2009.

[14] Roy Rada, Hafedh Mili, Ellen Bicknell, and Maria Blettner. Development and application of a metric on semantic nets. *IEEE transactions on systems, man, and cybernetics*, 19(1):17–30, 1989.

[15] 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.

[16] Lance De Vine, Guido Zuccon, Bevan Koopman, Laurianne Sitbon, and Peter Bruza. Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1819–1822. ACM, 2014.

[17] SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. Distributional semantics resources for biomedical text processing. *Proceedings of LBM*, pages 39–44, 2013.

[18] MUNEEB TH, Sunil Sahu, and Ashish Anand. Evaluating distributed word representations for capturing semantics of biomedical concepts. In *Proceedings of BioNLP 15*, pages 158–163, Beijing, China, July 2015. Association for Computational Linguistics.

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

[20] Rimma Pivovarov and Noémie Elhadad. A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. *Journal of biomedical informatics*, 45(3):471–481, 2012.

[21] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[22] Claudia Leacock and Martin Chodorow. Combining local context and wordnet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283, 1998.

[23] Hoa A Nguyen and Hoa Al-Mubaid. New ontology-based semantic similarity measure for the biomedical domain. In *2006 IEEE International Conference on Granular Computing*, pages 623–628. IEEE, 2006.

[24] Angelos Hliaoutakis. Semantic similarity measures in mesh ontology and their application to information retrieval on medline. *Master's thesis*, 2005.

[25] Mo Yu and Mark Dredze. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 545–550, 2014.

[26] Chang Xu, Yalong Bai, Jiang Bian, Bin Gao, Gang Wang, Xiaoguang Liu, and Tie-Yan Liu. Rc-net: A general framework for incorporating knowledge into word representations. In *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, pages 1219–1228. ACM, 2014.

[27] Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*, 2014.

[28] Saeed Hassanpour and Curtis P Langlotz. Unsupervised topic modeling in a large free text radiology report repository. *Journal of digital imaging*, 29(1):59–62, 2016.

[29] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G

Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[30] AL Goldberger, LAN Amaral, L Glass, JM Hausdorff, P Ch Ivanov, RG Mark, JE Mietus, GB Moody, CK Peng, and HE Stanley. Components of a new research resource for complex physiologic signals. *PhysioBank, PhysioToolkit, and Physionet.*

[31] Thabet Slimani. Description and evaluation of semantic similarity. 2013.

[32] Vector representations of words — tensorflow core — tensorflow.

[33] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 13(Feb):307–361, 2012.