Similarity Analysis on Unstructured Text using Dependency Trees in BioMedical Domain

by

Naga Venkata SivaPratap Palakurthi

B.E., VIT University - India, 2011

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Masters of Science

Department of Computer Science

2017

This thesis entitled:
Similarity Analysis on Unstructured Text using Dependency Trees in BioMedical Domain
written by Naga Venkata SivaPratap Palakurthi
has been approved for the Department of Computer Science

_____

Prof. Lawrence Hunter

_____

Robin Dowell

_____

Kevin Cohen

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Palakurthi, Naga Venkata SivaPratap (M.S., Computer Science)

Similarity Analysis on Unstructured Text using Dependency Trees in BioMedical Domain

Thesis directed by Prof. Lawrence Hunter

The published Biomedical scientific literature discusses most of the relationships between biomedical entities like drugs, genes, diseases and cellular processes. Relationships in the form of X (drug) inhibits Y (Gene), X (drug) treats Y (disease) and so forth are scattered in an unstructured format over millions of articles. Sentences like "X decreases Y", "Y is decreased by X" and "X reduces Y's effect" represents the same underlying relationship (decrease) between X and Y despite different sentence structures. Identifying such similarities in the relationships is critical to various applications in natural language processing and information retrieval.

Extracting these similar relationships between entities has various applications in question and answering [1], relationship analysis [2], and semantic search [3]. However, identifying these relationships from the vast corpus of unstructured data is a complex task which involves techniques like data mining, machine learning, and Natural language processing. We found that various methods like EBC [2] have inherent drawbacks in scaling to larger datasets and also in using full-text bodies for analysis. Inspired by this need, this thesis work focuses on scalable similarity analysis on the unstructured text of full-text bodies using entities from different ontologies.

We devised a new method - Mengsim, which is a dependency parse based similarity detection technique that finds similar relationships between semantic concepts from sentences like "X decreases Y", "Y is decreased by X" and "X reduces Y's effect". Mengsim relies on dependency grammar which gives syntactic connections between words in a sentence [4].

Mengsim's evaluation along with standard models showed its effectiveness in retrieving similar relationships. We also found that the proposed method can scale to larger datasets. We used concepts from three biomedical ontologies in our methods - diseases, drugs and genes which show the ability to scale to multiple ontologies.

Dedication

I dedicate this thesis to my dad who always encouraged me in pursuing my dreams and also to the rest of my family.

Acknowledgements

I would like to thank my advisor, Prof. Larry Hunter, who always amazes me with the knowledgeful discussions and also for introducing me to the fascinating field of BioMedical Informatics. Larry is always supportive of my research and has guided me in every step.

I would also like to thank Kevin Cohen, whose critical view of my work made me think deep into work and infact changed my perspectives towards life. I'm also greatly thankful to Robin Dowell for serving as my committee member and by providing the valuable feedback. I'm also thankful to Bethany Percha for helping me with the EBC's implementation.

I'm also thankful to my friends Suhas and Varshini for supporting me in tough times. Lastly, I would like to thank Grammarly application for helping me in producing quality content with its intelligent interfaces.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

This thesis proposes a dependency path based relational similarity method which applies to massive data sets of full-text articles like Pubmed. Scientific literature in the biomedical field is growing at an exponential rate for the human readers to keep up. This unprecedented increase in content has lead the researchers to develop computational methods to understand/analyze the literature automatically with/without supervision. One of the interesting problems in such analysis is relation extraction task which deals with identifying relations between nominals. Relational analysis by Swanson [5] lead to the discovery of a previously unknown connection between magnesium and migraines. Percha et al. [2] discovered five new drugs which are previously unknown to DrugBank [DrugBank is a manually curated database of drugs and its interactions].

Relations on a fundamental level, are the connections people perceive among concepts or nominals, i.e., how one nominal interact with another nominal. The relations in literature are in the form of text fragments which are usually described by the authors in a variety of ways. For example, x is increased by y, y increases x, y significantly increases y - all arguably describe a single generic relation 'increase'. Identifying the words (is increased by, increases, significantly increases) which represent relationships in the text often helps in mapping them to the standard relations like hypernyms and hyponyms (which are typically defined in ontologies and knowledge bases).

In practice, mapping the text fragments to standard relations (for example 'is increased by' to 'increase') on a large scale of data like PubMed is a complicated and computationally expensive process. Popular approaches use methods to identify similarities in the text fragments describing

the relations which help in grouping the similar text fragments and can be named to represent a standard relation.

Various methodologies were developed to find similarities in text fragments describing relationships. These methodologies fall in supervised, semi-supervised or unsupervised learning approaches. Further, the similarity techniques used in these approaches are based on the broad spectrum of methods like attributional similarities and relational similarities.

However, these methodologies often face problems in scaling to large datasets like Pubmed. Pubmed is a huge collection of 26+ million articles which is the primary reason for these scalability issues. This thesis focuses on a developing a PubMed scale relational similarity methodology for identifying text fragments representing similar relationships.

This thesis also focuses on using full bodies of literature rather than short abstracts which gives an opportunity to extract various relations not mentioned in abstracts.

Chapter 2

Relevant Background

2.1 Unstructured Text

Rich domain-specific unstructured knowledge resources such as Pubmed present large corpora of scientific journals for analysis. Most of the knowledge in the biomedical field is in the form of unstructured text in scientific journals [6]. The amount of data available to scientists in journals is overwhelming. As described by V Nastase et al., [5], "without automation in the form of knowledge discovery, connections between entities or phenomena may go unnoticed." The computational approaches for automation have gained popularity in the biomedical field from decades. These procedures need to extract data efficiently, aggregate, annotate and store information from these unstructured texts.

However, unstructured data presents significant challenges for computational methods. Authors represent results in natural language using different word choices and different sentence structures. Algorithms need to understand and link synonyms for named entities like drugs, genes, relationships, etc. Computational approaches need to overcome differences in sentence structures (like "X decreases Y", "Y is decreased by X") and word choices (like 'decrease', 'reduce') which leads to a formal representation of text for better analysis. Moreover, the scale of the unstructured data available for analysis is increasing rapidly [42] which lays the emphasis on scalability issues. Performance and scalability are important features for these computational approaches.

2.2 Concepts

The word 'Concept' is associated with many controversies related to its representation and relevance in the biomedical domain [7]. In our work, we refer concept as any ontologically defined term. Throughout the work, we use different terms like entities, nominals, and concepts which all mean the same.

Ontologies are a formal way of representing knowledge in which terms have a clear, unambiguous meaning which makes them suitable for representing semantics in an ambiguous unstructured text. Bio-ontologies represent concepts from life sciences and molecular biology. There are many Bio-ontologies in the biomedical domain like Chemical Entities of Biological Interest (CHEBI), Human Disease Ontology (DOID), Gene Ontology (GO), etc. There also exist controlled knowledge bases like OMIM, PharmGKB, DrugBank which provide more information about the ontological terms. These Bio-ontologies and knowledge bases play a central role in bioinformatics: they act as database integrators, shared vocabularies, and more [8]. Structured ontologies and knowledge bases are becoming increasingly central to the construction of sophisticated information retrieval systems.

Creation of Bio-ontologies and knowledge bases by humans is an intensive and a time-consuming task requiring a great deal of human effort. Moreover, these knowledge bases need an extensive maintenance, updating and deleting of relationships [2].

The ontological concepts are usually represented in a wide variety of ways in the scientific literature. Concepts can be expressed using synonyms, hyponyms or various morphological forms (Bollegala et al. [9]). Identifying such different types and mapping to an ontologically defined concept is a critical process in understanding the unstructured text and is performed using techniques like ontology term recognition. Table 2.1 shows various representations of concepts used in bio literature.

We used three different structured sources for analysis in our methods. 1) The Disease Ontology [10] is a standardized ontology which has concepts from human diseases, phenotype

Table 2.1: Various representations of semantic concepts

Semantic Concept	Representations
AIDS	acquired immunodeficiency syndrome, acquired immunodeficiency, AIDS
helminthiasis	helminth infections, helminth-infected, helminth infection, Helminthiasis, helminthiasis
Pneumocystis	pneumocystis pneumonia, pneumocystis, Pneumocystis

characteristics, and related medical vocabulary disease concepts. 2) ChEBI"(Chemical Entities of Biological Interest) ontology [11] provides concepts of molecular entities focused on 'small' chemical compounds like drugs, atoms, roles, etc. 3) Genes extracted from PharmGKB (Pharmacogenomics Knowledge base) [12] drug-gene associations. The examples of concepts used in our methods are in Table 2.2. Though we used concepts from 3 ontologies in our results, our method is designed to work with concepts from any ontology.

Table 2.2: Sample Concepts from 3 Ontologies used in our methods.

Ontologies	Sample Concepts
ChEB"- Drugs	Tamoxifen, codeine, metoprolol
PharmGKB - Genes	BCL2L1, BRAC1, CYP2D6, KIT
DOID - Diseases	Antiviral drug, fluconazole, chloroquine

2.2.1 Concept Extraction

Ontology term recognizers transform unstructured text to formal meaning representations. The output of such recognizers is a partial formal representation of the underlying text. As described by Funk et al. [7], concepts are difficult to recognize in text due to a disconnect between what is captured in an ontology and how the concepts are expressed in the text. Funk further describes the tremendous diversity in the possible forms of representations for these concepts.

Various automated concept mapping tools exist that recognize references to ontological terms in the text. Numerous advances in the availability of high-quality ontologies and the ability to accurately identify concepts in texts, and in language processing methods have made significant progress in concept analysis. Most widely used generic concept mapping tools are the National Library of Medicine's MetaMap, UIMA ConceptMapper and NBCOs Open Biomedical Annotator (NCBO Annotator) [7]. Other less familiar tools are Whatizit, KnowledgeMap, CONANN, IndexFinder, Terminizer, and Peregrin [7].

The system described in Funk et al. [7], uses UIMA ConceptMapper with best-performing parameters for each of the supported ontologies. Funk's concept mapper is used in our methods to identify the biomedical entities in sentences. It should also be noted that the concepts recognized by any of the automated concept recognition tools are not always accurate.

2.3 Relations

Relations are the connections people perceive between objects. As Tesniere [13] describes "The connection is indispensable to the expression of thought. Without the connection, we would not be able to express any continuous thought, and we could only list a succession of images and ideas isolated from each other and without any link between them."

Any piece of text describes a set of entities and the ways in which these entities connect. The connections usually come from 2 sources. One is through knowledge sources like ontologies or from general knowledge. A typical example of such knowledge is from GO ontology mitochondrion -¿ is -¿ intracellular organelle, where 'is' represents a relationship between mitochondrion and intracellular organelle. The second source of connections comes from text fragments. For example, in the text 'ibuprofen is prescribed to alleviate gingival pain', the connection between ibuprofen and gingival pain is 'alleviate.'

Depending on the granularity, relations can be unary (involves one object), binary (involves two objects), ternary (involves three objects) or a more generic n-ary (involves n objects). Unary relations describe properties of an object. Binary relations are most popular and extensively studied

Table 2.3: Demonstration of relationships between semantic concepts

Drug - Concepts	Gene - Concepts	Relation
chloroquine	TLR9	block
lapatinib	EGFR	can activate
capsaicin	TRPV1	binds
antidepressant	BDNF	activates
Oxaliplatin	TRPA1	abolishes
morphine	SON	affects
gefitinib	EGFR	antibody
metformin	STAT3	down regulates
morphine	CPP	induces
itraconazole	CYP3A4	inhibitor

in bioinformatics. Table 2.3 gives examples of binary relations between two class of objects Drugs and Genes. RelEx [14], JT Chang et al., [15], Percha et al., [2] have all concentrated on binary relations.

The relationships are often extracted using two methodologies. One methodology is the inventory approach, which is computationally suited to many computational models. This method needs to define an inventory of relations and later use it in the analysis. For example, Nastase and Szpakowicz [16] established 30 relations like cause, effect, product, part, whole, etc. and used them for relation extraction.

In the second approach, relations were not defined initially. The dataset of relations is built as the algorithms process and analyze data. EBC [2] uses this methodology to learn the structure of relationships. The inventory approach and exploratory approach each have their advantages and can be chosen based on the specifics of the problem.

The relations described in this thesis mainly describe the instance of relations like reduce, regulates, etc., rather than 'type' of relations (like causal, spatial) unless directly specified.

2.4 Syntactic Relations

Due to unstructured text's complexity and varied ways of representation of Information, we use underlying syntactic representations to capture interactions between concepts.

In the English language, sentence structures often follow subject-verb-object (SVO) format where subject comes first, verb comes second, and the object third. In the sentence "Paracetamol decreases fever", subject is paracetamol, verb is decreases and object is fever. There are other structures of language like SOV (Paracetamol fever decreases), VOS (decreases fever Paracetamol) etc., which are not valid in English. The sentence structures are governed using a set of rules, processes, and principles known as syntax. Parse trees (syntactic trees) are used to represent these structures in language according to a specified grammar. Parse trees will be discussed in detail in the following section.

2.4.1 Dependency Paths - Syntactic Relations

Dependency parse trees are a variation of parse trees that provide a representation of a Grammatical Relations (GR) between words in a sentence. These Grammatical Relations (GR), inspired by Dependency Grammar, offers a level of abstraction over specific syntactic analyses [17]. The importance of Grammatical Relations (GR) connecting ontology terms is that they may help in identifying the meaning of relationships.

Using dependency trees for identification of relationships is a popular approach discussed in various studies. Zelenko et al., [18] used tree kernels on dependency parse tree representation of sentences to extract relations between entities. Bunescu et al., [19] further hypothesized that the shortest path between entities in a dependency tree contains almost entire information needed for relationship extraction.

Each relationship in a dependency tree is an asymmetric binary relationship between a word called head and another word called modifier [13]. An example of a Dependency relationship is "dobj(decreases, fever)", where dobj is the "direct object" relationship between head - decreases and modifier - fever. The structure of a sentence can be represented by a set of such relationships through which a tree (directed graph) of a sentence can be created.

Regarding the implementation details, there exist multiple implementations of dependency parsers like Stanford Dependencies (SD) [17] and MINIPAR [20]. We used Stanford parser to

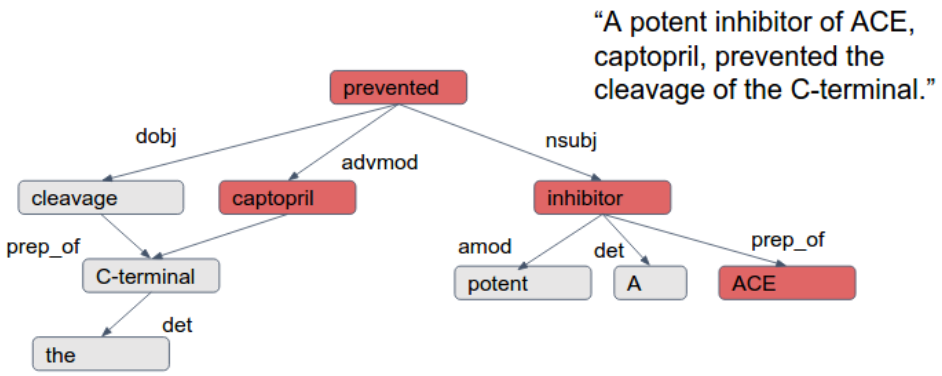
extract dependencies trees for sentences. Stanford Dependency (SD) representation is chosen as it showed promising results in relationship extraction tasks in the works of [2] [14] [21] [22]. Stanford Parser uses around 50 dependency relationships. Table 2.4 lists a subset of universal dependency relationships. SD provides good coverage of core grammatical relations, such as subject, object, internal noun phrase relations, and adverbial and subordinate clauses [23]. More details about the design principles and GR's used in Stanford parser can be found in [17].

Table 2.4: Subset of Universal Dependency Relations

Relation	Description	Example
det	determiner - Relation between the head of an NP and its determiner.	**The** man.
advmod	adverb modifier - An adverb or adverb-headed phrase that serves to modify the meaning of the word.	**Genetically** modified.
dobj	direct object - Direct object of a VP is the noun phrase which is the object of the verb.	She gave me a **raise**.
nsubj	nominal subject - is a noun phrase which is the syntactic subject of a clause.	**Clinton** defeated Dole.

The dependency tree for the sentence "A potent inhibitor of ACE, captopril, prevented the cleavage of the C-terminal." formed using SD is shown in Figure 2.1.

Figure 2.1: Graphical representation of the Dependency tree parsed using SD for the sentence: A potent inhibitor of ACE, captopril, prevented the cleavage of the C-terminal.



The links in Figure 2.1 represent dependency relationships. The direction of a link is always from the head to the modifier in the relationship. Labels associated with the relationship describe the type of dependency relationships.

The path between two words (which are not connected directly) is formed by concatenating the dependency relations and the words in between them. Previous work in [2] [14] [24] [19] have shown evidence that the information required to assert a relationship between two concepts in a sentence is typically captured by the shortest dependency path connecting those two concepts. The path between terms captopril and ACE from Figure 2.1 is

Captopril [advmod, prevented, nsubj, inhibitor, prep_of] **ACE**

2.5 Full text bodies

Scientific literature is published in the form of technical journals which contains abstract and body. Abstract contains a summarized version of the journal whereas body contains a long, and detailed version. Much of the previous relation centered methodologies like DIPRE, EBC, RelEx used only the short abstracts over the long and complicated full-text bodies.

According to Cohen et al. [25], full-text bodies differ structurally from abstracts in the scientific literature. The sentences are longer in full-text bodies. The longer sentences mean that

the number words per sentences will be high which might influence the performance of the parsers. Moreover, the sentences in full-text bodies use parenthesis significantly which further can inflict grave consequences on performance and efficiencies of information extraction tools.

Cohen further conducted a study on the number of mentions of biomedical entities in the full-text bodies vs. abstracts. The study revealed that on average article bodies have more mentions of drugs and disease semantic classes than abstracts. Drugs were mentioned on average 0.72 times in abstracts, whereas 13.6 times in bodies when a dataset of 97 articles is studied.

Thus factors like longer sentences, complex sentence structures, a significant number of entity mentions make full-text articles difficult to process than short text abstracts but offer a richer source of knowledge.

2.6 Learning methods on Pubmed Scale

Another good problem to solve in bioinformatics is the scale of data available for analysis. Most of the knowledge about biomedical entities like drugs, genes, proteins, etc., is published in the form of unstructured text in scientific journals. PubMed is a full-text collection of biomedical and life sciences journal literature at the U.S. National Institutes of Health's National Library of Medicine (NIH/NLM).

As of October 2016, Pubmed comprises more than 26 million journals with some records going back to 18th century. Adding to 26 million number, more than a million new records are added each year to Pubmed. Methods to analyze such a large scale data need specialized designs to process on time.

When the learning methods are considered, supervised methods need specific training data for each type of analysis. Creating training data involves tremendous manual effort and cannot be scaled easily even if there is a slight change in requirements. The same problem exists with semi-supervised methods, which often require a small seed set of training data. However unsupervised methods do not need any training data and can be scaled easily to different datasets.

2.7 Similarity Measures

Relation extraction tasks usually rely on similarity measures to group large number of relation instances which can later be named with hypernyms/hyponyms. In practice, a wide number of similarity measures are used for relation extraction. However, Turney broadly classifies the similarity measures into two types - Attributional Similarity and Relation Similarity.

Attributional similarity measures the similarity between two words based on their attributes. Two words are considered synonymous if they have a high degree of attributional similarity. Whereas, two words pairs are considered as analogous if they have high attributional similarity.

The second type of classification is Relational Similarity which corresponds to the similarities between relations. For example, two entities Acetaminophen: fever, ibuprofen: cold are relationally similar as the first word in each pair has a decreasing effect on the second word.

The relational similarity can be reduced to attributional similarity since Acetaminophen and ibuprofen are drugs whereas fever and cold are diseases. More about these methods are discussed in the methodology section.

Chapter 3

Methods

3.1 Introduction

In the previous chapters, we discussed the basic concepts like relations, scientific literature, huge datasets and learning methodologies. We also discussed the complexities of full-text bodies of scientific journals, realized the problems like scalability often faced in the relational analysis.

In this chapter, we will discuss in detail about a related research methodology - EBC and introduce a new method Mengsim to perform relationship analysis on a huge dataset. We will also demonstrate the method's applicability to various problems in the biomedical domain.

3.2 Related Work

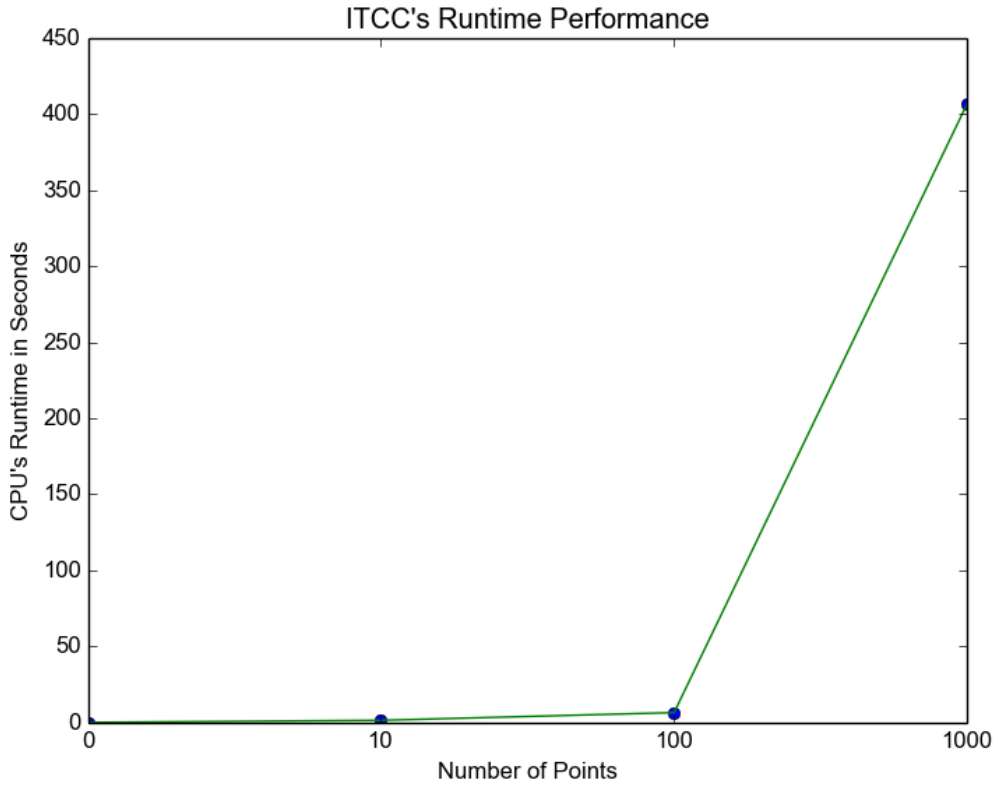
3.2.1 EBC

EBC (Ensemble Biclustering for Classification) [2] works on finding relationships between two class of entities like drugs and genes using a strong theoretic co-clustering algorithm called Information Theoretic CoClustering (ITCC) [26]. EBC relies on co-clustering frequencies, computed by ITCC to find similarities between dependency paths (relations) and also between concepts. EBC uses this similarity measure (clustered frequencies) to do further analysis of clustering and also ranks elements in clusters using seed sets to find new relationships.

EBC does not require to define a strict set of relations. It takes the naturally varying forms of relationships described in a text and attempts to learn the structure of relationships.

Though EBC is simple in implementation, our evaluation of EBC on huge text corpus showed its inabilities in scaling to large datasets. Each run of EBC's ITCC starts with random seed clusters, because of which it ends with different cluster assignments after each run. To get the right similarities, ITCC algorithm has to run for approx 2000 times. These enormous number of executions require large memory and also significant runtime. As can be seen in the figure, ITCC algorithm's runtime increases exponentially with the number of data points.

Figure 3.1: EBC's ITCC Performance



3.3 Mengsim Methodology

Methods like EBC relies on bootstrapping, which needs a small set of relation instances for extraction. Unfortunately, when working on various problems in the biomedical domain, one simply cannot specify seed samples for each of these relations (Turney et al. [27]). Thus we set us on a goal to develop methods which can scale to larger datasets like Pubmed without the need to specify seed sets.

Various types of learning methods rely fundamentally on similarity functions for analysis. Supervised relation extraction task relies on methods to find similarities between training data and testing data. Unsupervised methods use similarity functions to similar group relations.

As finding similarities between relation instances in free text is the fundamental task in relation extraction, we focus on building a scalable relational similarity method. To scale to larger datasets, the proposed method need to be fast and should be able to run in a distributed manner. The following sections outline the proposed approach.

3.3.1 Pre-Processing

Figure 3.2: Overview of methodology used in this thesis

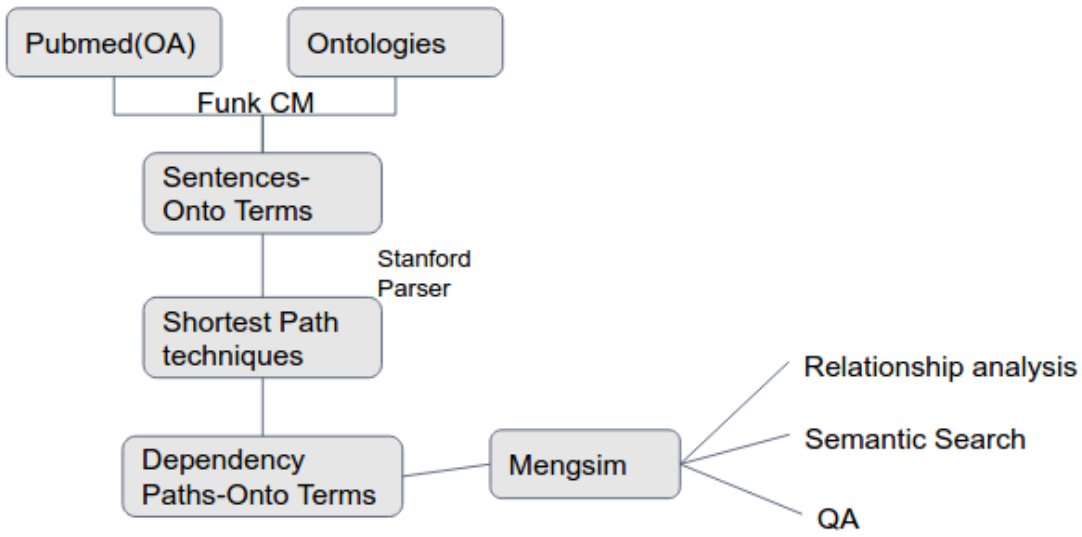


Figure 3.2 gives an overview of the system adopted in our methodology. The PubMed Central Open Access Corpus (November 2015 snapshot) has approximately a million articles which can be converted into text format from their original nxml format. Pubmed Sentences containing concepts from 3 different ontologies - ChEBI, DOID, PharmGKB Genes were extracted using state of the art Colorado Computational Pharmacology's concept mapper [7].

Many of the ontology terms are multi-worded, so we used a modified NLTK's sentence tokenizer which uses a dictionary of multi-word ontology terms. These tokenized sentences are then used to form the dependency trees using the Stanford Dependency parser. The dependency trees from the parser contain terms from the whole sentence. Whereas, a dependency path connecting the specified ontology terms gives the information needed to describe relationships between concepts. For this, we used the iGraph library to extract the dependency paths connecting specified ontology terms from the dependency trees. Each word in the dependency path is then represented using its morphological root.

Table 3.1 and Table 3.2 shows the datasets formed after pre-processing.

Table 3.1: Dataset 1 - ChEBI(drugs) and PharmGKB's Genes on whole pubmed

Total Number of sentences with these terms	35,322
Total Number of Unique Dependency Paths	31160
Total Number of Unique drug, gene pairs	2,115

Table 3.2: Dataset 2 - ChEBI(drugs) and Doid (infectious diseases) on whole pubmed

Total Number of sentences with these terms	2,716
Total Number of Unique Dependency Paths	2,521
Total Number of Unique drug, disease pairs	252

3.3.1.1 Filtering

Due to the usage of automatic processing methods, our datasets are susceptible to various errors. To eliminate errors, we filtered the dataset at different levels. Sentences which have more

than 500 characters are removed. Sentences of this length occur likely due to the errors in sentence segmentations due to the presence of characters like ';' and ':' as delimiters. Including such longer sentences will result in memory overloads in various stages like parsing and path extraction. We further removed all pairs of concepts with less than five co-occurrences, to eliminate relationships with less significance.

Some of the dependency paths obtained may not represent a relation between entities. These incorrect paths may reduce the performance of algorithms and also reduce efficiency. So these should be identified and filtered before proceeding to further steps. Paths containing conjunctions like 'and' and 'or' are one of such incorrect paths. The paths containing conjunctions usually connect two subsentences or two clauses and the presence of them between two entities signifies an error [Thomas et al. [28]]. The cause of such errors is due to the limitations in the conversions by Stanford parser [MC De Marneffe et al. [17]].

3.3.2 Similarity function - Dependency path similarity

From the previous steps of dataset preparation and filtering, a cleaned dataset of dependency paths connecting various entities is obtained. We now need a similarity function which can compute similarities between the huge number of dependency paths. The proposed method Mengsim computes the similarity score between pairs of dependency paths. The greater the score is, the greater the similarity between the relationships expressed.

The motivation for a new distance function like Mengsim is because of the inherent drawbacks with standard similarity models which are not designed to work with dependency paths. Dependency paths connecting biological concepts often consists of terms like prepositions, dependency terms (like nsub, amod) and other relevant terms describing underlying semantic relationships like decrease, inhibits and expedites. All these terms need to be handled separately due to the inherent differences, which makes non-weighted functions like cosine inapplicable.

Mengsim's distance function is inspired from the similarity detection method used in DIRT [24] (extended distributional hypothesis). Mengsim calculates similarities between two syntactic

expressions U and V based on the following intuitive principles.

- The similarity between relationships U and V is related to what they have in common. The more commonality the relationships have, the more similar they are.
- The difference in relationship lengths of U and V is inversely related to similarity between them.
- The commonality of verbs and actual words in relations like 'decrease', 'inhibitor' and so forth, have more weighting than commonality of dependency terms like 'advmod', 'nsubj', 'dobj' etc.
- The commonality of prepositions like 'of', 'in' etc., have lower weighting than commonality of dependency terms and actual words.

The distance is calculated using the formula

$$\text{mengism\_score (U, V, W)} = \frac{\lceil U.V \rceil * W}{(||U.W| - |V.W|| \,/\, Size(U)) + |(U+V).W|}$$

In the distance formula, U and V are vector representations of two dependency paths. W is a weighted vector which is used to weigh different terms in the relation.

Mengsim's score is symmetric which means

$$mengsim\_score(U, V, W) = mengsim\_score(V, U, W).$$

The order of weighted vector W is equal to the combined number of unique terms in two dependency paths U and V. Equal weights like [1 1 1 ... 1], gives equal importance to all terms in the dependency path which yields results similar to non-weighted methods like cosine.

We used simple weight functions with weights ranging from 0 to 1. The terms with weight:1 are most relevant in the similarity and terms with weight:0 are least relevant.

To determine the right weights, we looked closely at biological interactions between various types of entities. The following table describes the interactions between genes and drugs extracted from various journals in Pubmed.

Table 3.3: Demonstration of Drug - Gene relationships between concepts

Drug - Concepts	Gene - Concepts	Relation
chloroquine	TLR9	block
lapatinib	EGFR	can activate
capsaicin	TRPV1	binds
antidepressant	BDNF	activates
Oxaliplatin	TRPA1	abolishes
morphine	SON	affects
gefitinib	EGFR	antibody
metformin	STAT3	down regulates
morphine	CPP	induces
itraconazole	CYP3A4	inhibitor

Also, the possible relation types between drugs and diseases comprise relations like 'treating', 'cures', 'doesn't cure', 'stabilizes' etc., These relations shows that verbs describe the relation instance and prepositions add details to that relation. This work concentrates on finding hypernyms or phenomena of relationships between entities and doesn't deal with the finer details of a relationship. Thus for finding the similar relationships, we relatively down weigh the prepositions compared with other words like verbs.

The weights in our method were based on trial and error method. In future, these can be improved by using a machine learning model on a trained corpus of similar dependency paths.

To demonstrate similar dependency paths, we ran the proposed Mengsim metric on two datasets along with baseline methods - cosine and wminkowski. Cosine is the most popular non-weighted generic similarity function, whereas wminkowski is the popular weighted function. Though there exist many similarity functions designed for specific purposes, we choose the popular cosine and wminkowski as baseline models because of the complexities of manual evaluation mechanism used in this work. To demonstrate the results of each method, we extracted similar dependency paths using the following base sentence with the semantic concepts 'FAAH' gene and 'analgesic'

drug.

"This reasoning is supported by findings that the **analgesic** effects of a **FAAH** inhibitor persist after long-term administration and no apparent desensitization of CB1R function takes place after chronic FAAH inactivation."

The syntactic relation (dependency path) connecting these two concepts 'FAAH' and 'analgesic' is

[amod, effects, nmod:of, inhibitor, compound]

Table 3.4 shows the top 10 similar sentences retrieved when cosine similarities are used. Since it is a non-weighted approach, it treats all terms equally due to which terms with less significance like 'due' and shorter sentences get more similarity scores. Table 3.5 shows the top 10 similar sentences obtained from weighted minkowski distance function with different weights to propositions, dependency terms, and other words.

Table 3.4: Top 10 cosine similar sentences for base syntactic relation [amod, effects, nmod:of, inhibitor, compound] and concepts 'FAAH' and 'analgesic'

[u'amod', u'inhibitor', u'compound']	We report that coadministration of **ketoconazole**, a strong **CYP3A4** inhibitor, with 5 mg i.v. temsirolimus in healthy subjects had no effect on temsirolimus
[u'amod', u'inhibitor', u'nmod:of']	These different mutations are associated with different risk of relapse after resection of the primary tumor but also impact on the activity of **imatinib** (selective inhibitor of **KIT** and PDGFR-α) [4–6].
[u'compound', u'inhibitor', u'compound']	Furthermore, MCF-10A cells treated with **Nutlin-3**, a **MDM2** E3-ubiquitin ligase-specific inhibitor that arrests WT-p53 degradation, showed increased p53 stability, resulting in decreased endogenous Nox4 protein (Figure 4).
[u'amod', u'treatment', u'compound']	Combined with **dexamethasone**, **DES**treatment resulted in 50% PSA declines in 64–68% of patients in a small randomised study (Shamash), but venothromboembolic events occurred in 22% of patients in the combination arm.
[u'amod', u'Iressaxae', u'compound']	Study results demonstrated that the **EGFR** tyrosine kinase inhibitor **gefitinib** (Brand name Iressa®) achieved a response rate of more than 80% in mutant tumors, but was basically ineffective in wild-type tumors without mutations [5].
[u'amod', u'protein', u'compound']	Two additional genes whose functions may be associated with AMPA receptors in dopaminergic neurons are also highly dysregulated in PD, **NSF** (**N-ethylmaleimide** sensitive fusion protein; 202395_at, P = 1.18252E−05, DElog 2 P = 4.12902E−06, DElog 2 28];
[u'amod', u'inhibition', u'compound']	Compared to **ezetimibe**, **PCSK9** inhibition resulted in a mean LDL-C reduction of approximately 36 % (95 % confidence interval (CI), 33–39 %) [19].
[u'amod', u'resistant', u'amod', u'patients', u'nmod:in', u'expression', u'compound']	Also in clinical studies, an inverse correlation between **EGFR** and ERα expression in **tamoxifen** resistant patients has been reported [5,6,18-20]
[u'amod']	Involvement of EGF signalling in the pathogenesis of bone metastasis was implicated by the unexpected relief of bone pain in phase II clinical trials of **EGFR** inhibitor **gefitinib** in breast cancer patients (Albain ; von Minckwitz).

Table 3.5: Top 10 wminkwoski similar sentences for base syntactic relation [amod, effects, nmod:of, inhibitor, compound] and concepts 'FAAH' and 'analgesic'

[u'amod', u'inhibitor', u'compound']	We report that coadministration of **ketoconazole**, a strong **CYP3A4** inhibitor, with 5 mg i.v. temsirolimus in healthy subjects had no effect on temsirolimus
[u'amod', u'inhibitors', u'compound']	However, various anti-diabetic drugs, such as **metformin**, dipeptidyl peptidase 4 (**DPP4**) inhibitors, although it has been successfully work to blood glucose lowering in type 2 diabetes, have been repurposed from other clinical indications to treat renal injury.
[u'amod', u'Inhibitor', u'compound']	Renoprotection by benazepril and telmisartan in diabetes: A histopathological study Singh J 1 2 Department of Pharmacology Pathology Objective: To observe effects of **benazepril**{**ACE** Inhibitor} and telmisartan.
[u'amod', u'effects', u'compound']	In the TNF-transgenic mice, up-regulation of both TNFR mRNAs was detected after BDV-infection so that **anticonvulsantand** proconvulsive **TNF** effects might have been operative.
[u'nmod:of', u'effects', u'nsubj', u'inhibitor', u'compound']	Nishimura and Bailey also studied the effects of **captopril**, an **ACE** inhibitor, and proved that the effects involved Ang II generation.
[u'nmod:than', u'effective', u'nsubj', u'inhibitors', u'compound']	Four meta-analyses have suggested that angiotensin-converting enzyme (**ACE**) inhibitors are more effective than other **antihypertensive agents** in reducing LVH, for similar reductions in BP.
[u'nmod:of', u'inhibitor', u'compound']	Captopril, for instance, is a potent **ACE** inhibitor, and administration of **captopril** lowers the blood pressure.
[u'compound', u'effect', u'nmod:of']	However, in contrast to purported **antidepressant** effect of **BDNF** signaling, cytokine action is largely associated with depressive-like behaviors.
[u'compound', u'inhibitor', u'nmod:of']	In addition to regulating physiological vessel growth, the NRP1–ABL1 pathway promotes vascular pathology that can be inhibited by treatment with **Imatinib**, a small molecule inhibitor of **ABL1**, or through the genetic ablation of NRP1 in ECs; thus, both approaches significantly and similarly reduced vessel growth in a mouse model of human retinopathy.
[u'nmod:of', u'inhibitors', u'compound']	The majority related to hypotensive effects of varying combinations of **ACE** inhibitors, loop **diuretics** and calcium channel blockers.

Table 3.6: Top 10 Mengsim similar sentences for base syntactic relation [amod, effects, nmod:of, inhibitor, compound] and concepts 'FAAH' and 'analgesic'

[u'nmod:of', u'effects', u'nsubj', u'inhibitor', u'compound']	Nishimura and Bailey also studied the effects of **captopril**, an **ACE** inhibitor, and proved that the effects involved Ang II generation.
[u'amod', u'inhibitor', u'compound']	We report that coadministration of **ketoconazole**, a strong **CYP3A4** inhibitor, with 5 mg i.v. temsirolimus in healthy subjects had no effect on temsirolimus
[u'amod', u'inhibitors', u'compound']	However, various anti-diabetic drugs, such as **metformin**, dipeptidyl peptidase 4 (**DPP4**) inhibitors, although it has been successfully work to blood glucose lowering in type 2 diabetes, have been repurposed from other clinical indications to treat renal injury.
[u'amod', u'Inhibitor', u'compound']	239 Renoprotection by benazepril and telmisartan in diabetes: A histopathological study Singh J 1 2 Department of Pharmacology Pathology Objective: To observe effects of **benazepril**{**ACE** Inhibitor} and telmisartan.
[u'amod', u'effects', u'compound']	In the TNF-transgenic mice, up-regulation of both TNFR mRNAs was detected after BDV-infection so that **anticonvulsant** and proconvulsive **TNF** effects might have been operative.
[u'amod', u'inhibitor', u'nmod:of', u'receptor', u'compound']	The first to receive FDA approval was **gefitinib**, a small molecule inhibitor of the epidermal growth factor receptor (**EGFR**) tyrosine kinase, a transmembrane receptor whose activation leads to intracellular signaling involved in cancer cell proliferation and survival.
[u'amod', u'effects', u'nmod:of', u'antagonists', u'compound']	Accordingly, first case reports show **analgesic** effects of **TNF** antagonists in patients with treatment-refractory pain caused by bone metastases [26].
[u'amod', u'inhibitor', u'nmod:of', u'binding', u'compound']	We found that **mithramycin**, an inhibitor of **SP1** binding, could synergize with paclitaxel in some TNBC (basal-like) cell lines, MDA-MB-231, MDA-MB-468, and HDQ P1.
[u'amod', u'inhibitor', u'nmod:of', u'AUC', u'compound']	Furthermore, when zibotentan was administered in combination with **itraconazole**, a potent inhibitor of **CYP3A4**, AUC increased by 28% [17].
[u'amod', u'inhibitor', u'nmod:of', u'PDGFRA', u'compound']	By the introduction of the tyrosine kinase inhibitor (TKI) **imatinib**, a selective inhibitor of **KIT**, PDGFRA and a few other kinases, a new era in the management of GIST began [4,5].

Table 3.6 shows top 10 similarities using Mengsim. Similar dependency paths extracted with Mengsim have balanced number of terms, unlike cosine method which extracts shorter dependency paths. A detailed assessment and evaluation of Mengsim along with baseline methods will be described in chapter 4.

Chapter 4

Evaluation

Various similarity measures are developed and studied in many text applications. These models were developed to measure similarities at different levels like sentence level, relation level, phrase level, entity level. Achananuparp et al. [29] summarize 14 similarity metrics which are used to measure sentence similarity. SemEval has a dedicated challenge for measuring relational similarity. The usual relation extraction task in SemEval is to find the degree to which the semantic relations between A and B are similar to those between C and D from two pairs of words, A:B and C:D. The training and test datasets for the SemEval task are manually prepared by a limited number of people where they vote on the similarity between two sentences/entities.

We now discuss in detail the evaluation mechanisms adopted by some of the related methodologies. Zhang et al. [30] proposed an unsupervised learning method using dependency trees to extract relations between named entities like persons and locations. The method is targeted towards relations in general news data. They used Precision, Recall, and F-measure to evaluate the similarity function. Hasegawa et al. [31] proposed a similar method but using contextual words instead of dependency trees and used similar evaluation metrics. As the methods employed in this research are closer to the these, we follow a similar approach for evaluation. For preparations of test data, we rely on manual processes as used in SemEval tasks.

The evaluation of the similarity function used in our methods is performed in two phases.

4.1 Phase 1

In the first phase, we prepared a dataset of sentences containing at least one occurrence of (drug, gene) or one occurrence of (drug, disease). A collection of 6 base sentences is chosen in random from the dataset. On each of these sentences, we ran our proposed similarity metric - Mengism along with standard relation metrics (cosine and wmin) on all other sentences containing same entities as the base sentence and extracted ten sentences with highest similarity score for each metric.

We then use Mechanical turk (MTurk) to evaluate the relational similarity between the results of each of these metrics. MTurk is a popular choice in computational linguistics for gathering large numbers of human responses to scientific questions (SemEval task [32]; Mohammad and Turney [27]). Mturk contains the major elements required to conduct research: an integrated participant compensation system; a large participant pool; and a streamlined process of study design, participant recruitment, and data collection [33].

In our evaluations, each HIT (Human Intelligence Task) in MTurk is presented with two sentences (base sentence and one from top similar sentence retrieved) along with a different pair of entities. In order to complete a hit, an MTurk worker has to choose one option among five options presented - (Perfectly Similar, Almost Similar, Opposite, Not Sure, Not Similar). Few examples are presented to the worker in the instructions along with explanations. An example task looks the following way,

Sentence1: When it contacts with epidermis, capsaicin activates the TRPV1 receptor, which elicits a rapid response via the release of neuropeptides (such as calcitonin gene related peptide, substance P, and tachykinins) and monoamimes (histamine and serotonin) [32] [34] [35]. Sentence1-Term-1: capsaicin Sentence1-Term-2: TRPV1

Sentence2: Also included were two transient vanilloid receptors, TRPV1 (the capsaicin receptor) and TRPV4 (another heat receptor). Sentence2-Term-1: capsaicin Sentence2-Term-2: TRPV1 Options:

- Perfectly Similar
- Almost Similar
- Not Sure
- Opposite
- Not Similar

Each sentence pair is presented to 3 workers, and the option with a majority vote is chosen. For simplicity purposes, we combined Perfectly Similar and Almost Similar categories as Similar. Opposite and Not Similar options are combined into Not Similar.

We used the precision metric to measure the efficiency of the similarity methods as it is consistent with related methodologies (Zhang et al. [30], Achananuparp et al. [29]). Due to the manual nature of evaluation by MTurk workers, we did not calculate recall as it involves evaluating the similarity between entities in hundreds of sentence pairs.

The following graph and the table show the results of Phase 1 evaluation. Mengsim's precision metrics are higher than others for the relational similarity task. Thus Mengsim performed better than the standard methods in both categories Perfectly similar and Perfectly Similar + Mostly Similar.

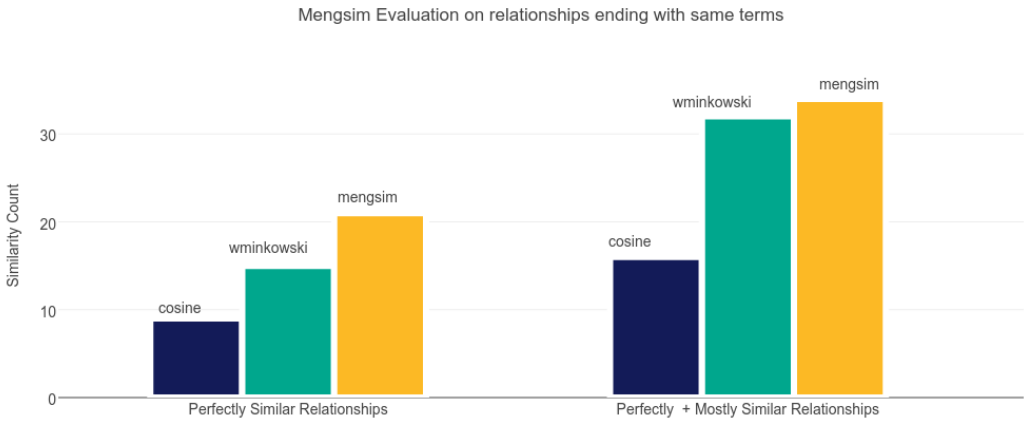
Table 4.1: Phase 1: Similarity evaluation results on three methods using fixed concepts

Evaluation type	cosine	wmin	mengsim
Perfectly Similar	0.15	0.25	0.35
Perfectly + Almost Similar	0.27	0.54	0.57

4.2 Phase 2

In phase two, a random six sentences are chosen with at least one occurrence of (drug-gene) or one occurrence of (drug disease). On each of these sentences, we ran our proposed similarity

Figure 4.1: Evaluation - Relational Similarity with fixed concepts



metric - mengism along with standard relation metrics (cosine and wmin) on all other sentences containing same entities or different entities as the base sentence and extracted seven sentences with highest relational similarity score for each metric.

The following figure shows the results of phase 2 evaluation. The results indicate that Mengsim performs better than the standard similarity metrics in both categories tested - Perfect Similar and Perfectly + Mostly Similar relationships .

Figure 4.2: Evaluation - Relational Similarity with all concepts

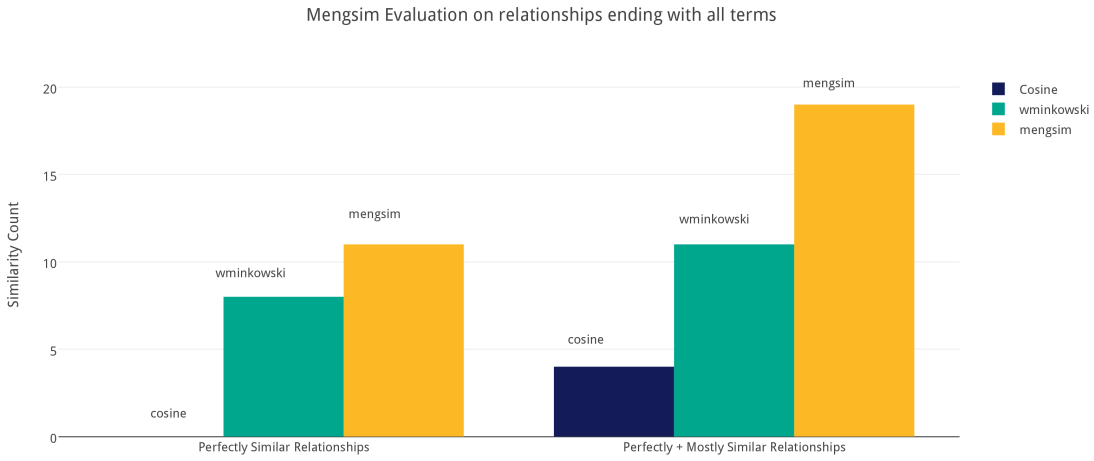


Table 4.2: Phase 2: Similarity Evaluation Results on three methods using all concepts

Evaluation type	cosine	wmin	mengsim
Perfectly Similar	0	0.19	0.26
Perfectly + Almost Similar	0.09	0.26	0.45

The results from phase 1 and phase 2 indicates that Mengsim performs far better than other methodologies in finding sentences with better relational similarity.

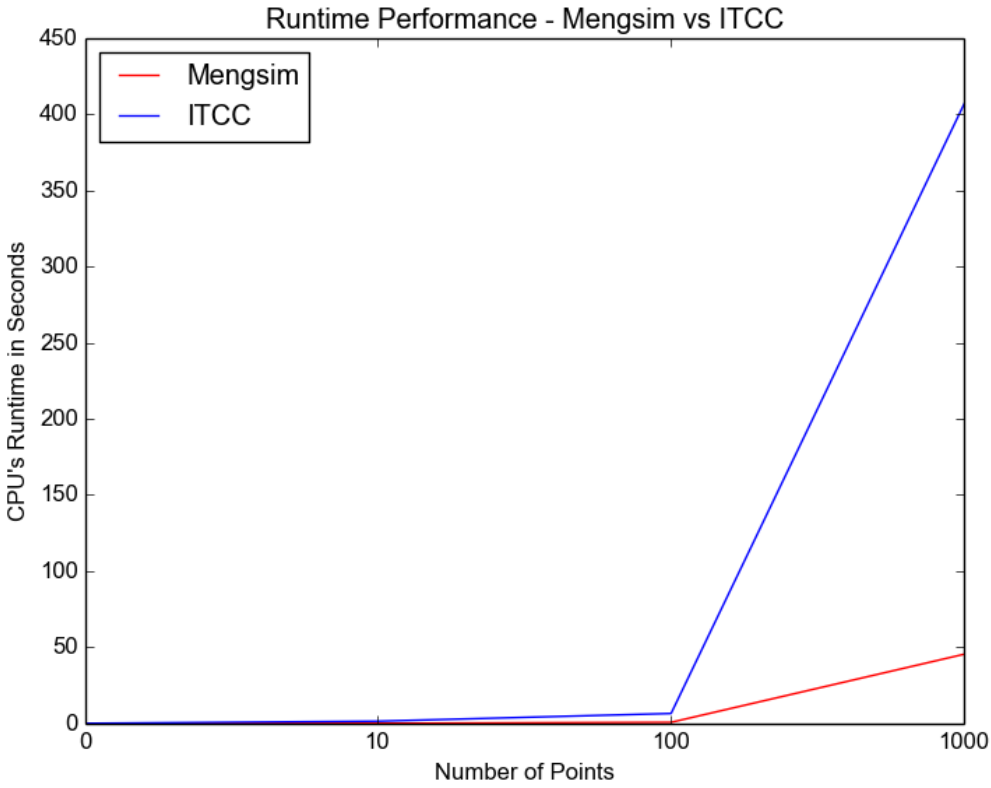
4.3 Performance

Apart from measuring the relational similarity between entities, the key motivation for developing a new system is to scale on big datasets like Pubmed. This work started while working on the EBC methodology (Percha et al. [2]) and realized its potential limitations when applied to huge datasets like Pubmed.

EBC uses ITCC to measure relationship similarity between entities like drugs and genes. The relationship similarity is based on clustering frequencies which are obtained by running the ITCC algorithm on the data for N number of times. As percha mentioned, a typical N number of iterations is over 1000. The runtime complexity for each run of ITCC is O(m*n) (Dhillon IS et al. 2003), where m = number of rows and n = number of columns. If ITCC needs to be executed for 1000 rounds, the runtime becomes O(1000*m*n). This enormous number of executions makes ITCC computationally expensive and becomes very hard to run on huge datasets. Adding to that, after each run of ITCC, clustering frequencies needs to be updated which further increases runtime of algorithm linearly. The memory requirements for this approach is O(m*m).

Whereas with Mengsim, computation of distance function has a runtime complexity of O(N*N) if processed sequentially, N represents the number of sentences. The runtime can further to reduced to O(N) by distributing the computation to N processes. This makes it easier to find similar relationships on huge datasets. The following figure shows the runtime evaluations of EBC and Mengsim.

Figure 4.3: ITCC vs Mengsim - Runtime Performance



Chapter 5

Discussion

The similarity results from chapters 3 and 4 demonstrated the Mengsim's ability in retrieving the sentences with similar relationships between entities. The usage of dependency path connecting the entities for relationship analysis reiterated the argument of Bunescu et al. [19] which is "the information relevant to relation extraction is almost entirely concentrated in the shortest path in the dependency tree, leading to an even smaller representation." The evaluation results in Chapter 4 showed that the proposed method Mengsim performs better than the standard relation metrics in extracting similar sentences.

Chapter 4 also demonstrated the Mengsims ability in scaling to larger datasets with ease. The methods can be distributed to various processes to reduce the time of executions.

5.0.1 Limits

The relational similarity approach proposed is dependent on various pre-processing steps - entity detection, sentence segmentation, parse tree construction and path extraction. All of these data processing stages uses off the shelf software components, and we do not calculate the performance of these steps.

The proposed Mengsim methodology does not need all the data to be in memory to compute distance for relational similarity. This independence on memory requirements makes the methods applicable to datasets of any size as long as the processes run in distributed fashion.

5.0.2 Error Analysis

The Stanford Parser [17] which was used in our system to extract dependency trees is trained on newswire data. Relying on newswire data favors performance on a particular type of linguistic data: formal text, written in a carefully constructed language and thoroughly revised. The performance of the parser on other kinds of data then suffers due to this bias. Web data is different from Bio literature. So often the output of parsers is subjected to errors in capturing dependencies between words which are far apart and also in accurately mapping dependency relations [17] [22]. For example, from the Figure 2.1, 'captopril' which is the name of a drug is incorrectly recognized as an adverb. We do not attempt to correct such dependency errors in our work.

Chapter 6

Conclusion

This thesis work started by evaluating a related work EBC to extract relationships from unstructured data. Inspired by the EBC's limitations regarding performance, we began work on alternative models for similarity analysis. The fundamental task in relation extraction methods is a relational similarity measure. This similarity measure is common to any learning methods like supervised or unsupervised.

We devised a new similarity measure called Mengsim which demonstrated better results than standard models in retrieving similarity relationships. Mengism can be used on sentence level by restricting dependency paths connecting any one concept to perform meaningful relationship analysis. Similarity analysis on two datasets formed from different sets of ontologies drugs-genes, drugs-diseases shows that Mengsim can be applied to terms of a diverse set of ontologies.

In future, we will use machine learning models to determine accurate weights for Mengsim. We will also focus on better visual mechanisms for displaying similar relationships. We will examine newer vector models like GloVe [36] in determining similarity measures in the biomedical analysis. We'll work on better evaluation mechanisms for Mengsim, as the evaluation method using a crowd-sourced system like mTurk is subjected to certain problems as mentioned in "Amazon mechanical turk: Gold mine or coal mine?" [37].

Bibliography

[1] David Martinez, Andrew MacKinlay, Diego Mollá Aliod, Lawrence Cavedon, and Karin Verspoor. Simple similarity-based question answering strategies for biomedical text. In CLEF (Online Working Notes/Labs/Workshop), 2012.

[2] Bethany Percha and Russ B Altman. Learning the structure of biomedical relationships from unstructured text. PLoS Comput Biol, 11(7):e1004216, 2015.

[3] Sebastian Köhler, Marcel H Schulz, Peter Krawitz, Sebastian Bauer, Sandra Dölken, Claus E Ott, Christine Mundlos, Denise Horn, Stefan Mundlos, and Peter N Robinson. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. The American Journal of Human Genetics, 85(4):457–464, 2009.

[4] Sebastian Padó and Mirella Lapata. Dependency-based construction of semantic space models. Computational Linguistics, 33(2):161–199, 2007.

[5] Vivi Nastase, Preslav Nakov, Diarmuid O Seaghdha, and Stan Szpakowicz. Semantic relations between nominals. Synthesis Lectures on Human Language Technologies, 6(1):1–119, 2013.

[6] Elizabeth K White. Pattern-based extraction of argumentation from the scientific literature. University of Colorado at Boulder, 2010.

[7] Christopher Funk, William Baumgartner, Benjamin Garcia, Christophe Roeder, Michael Bada, K Bretonnel Cohen, Lawrence E Hunter, and Karin Verspoor. Large-scale biomedical concept recognition: an evaluation of current automatic annotators and their parameters. BMC bioinformatics, 15(1):59, 2014.

[8] Mikel Egaña Aranguren, Erick Antezana, Martin Kuiper, and Robert Stevens. Ontology design patterns for bio-ontologies: a case study on the cell cycle ontology. BMC bioinformatics, 9(5):S1, 2008.

[9] Danushka Tarupathi Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. Relational duality: Unsupervised extraction of semantic relations between entities on the web. In Proceedings of the 19th international conference on World wide web, pages 151–160. ACM, 2010.

[10] Lynn Marie Schriml, Cesar Arze, Suvarna Nadendla, Yu-Wei Wayne Chang, Mark Mazaitis, Victor Felix, Gang Feng, and Warren Alden Kibbe. Disease ontology: a backbone for disease semantic integration. Nucleic acids research, 40(D1):D940–D946, 2012.

[11] Kirill Degtyarenko, Paula De Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcántara, Michael Darsow, Mickaël Guedj, and Michael Ashburner. Chebi: a database and ontology for chemical entities of biological interest. Nucleic acids research, 36(suppl 1):D344–D350, 2008.

[12] Micheal Hewett, Diane E Oliver, Daniel L Rubin, Katrina L Easton, Joshua M Stuart, Russ B Altman, and Teri E Klein. Pharmgkb: the pharmacogenetics knowledge base. Nucleic acids research, 30(1):163–165, 2002.

[13] Lucien Tesnière. Eléments de syntaxe structurale. Librairie C. Klincksieck, 1959.

[14] Katrin Fundel, Robert Küffner, and Ralf Zimmer. Relexrelation extraction using dependency parse trees. Bioinformatics, 23(3):365–371, 2007.

[15] Jeffrey T Chang and Russ B Altman. Extracting and characterizing gene–drug relationships from the literature. Pharmacogenetics and Genomics, 14(9):577–586, 2004.

[16] Vivi Nastase and Stan Szpakowicz. Exploring noun-modifier semantic relations. In Fifth international workshop on computational semantics (IWCS-5), pages 285–301, 2003.

[17] Marie-Catherine De Marneffe and Christopher D Manning. Stanford typed dependencies manual. Technical report, Technical report, Stanford University, 2008.

[18] Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella. Kernel methods for relation extraction. Journal of machine learning research, 3(Feb):1083–1106, 2003.

[19] Razvan C Bunescu and Raymond J Mooney. A shortest path dependency kernel for relation extraction. In Proceedings of the conference on human language technology and empirical methods in natural language processing, pages 724–731. Association for Computational Linguistics, 2005.

[20] Dekang Lin. Dependency-based evaluation of minipar. In Treebanks, pages 317–329. Springer, 2003.

[21] Michael A Covington. A fundamental algorithm for dependency parsing. In Proceedings of the 39th annual ACM southeast conference, pages 95–102. Citeseer, 2001.

[22] Michael Kaisser and Bonnie Webber. Question answering based on semantic roles. In Proceedings of the Workshop on Deep Linguistic Processing, pages 41–48. Association for Computational Linguistics, 2007.

[23] Marie-Catherine de Marneffe, Miriam Connor, Natalia Silveira, Samuel R Bowman, Timothy Dozat, and Christopher D Manning. More constructions, more genres: Extending stanford dependencies. Proceedings of DepLing, 2013.

[24] Dekang Lin and Patrick Pantel. Dirt@ sbt@ discovery of inference rules from text. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pages 323–328. ACM, 2001.

[25] K Bretonnel Cohen, Helen L Johnson, Karin Verspoor, Christophe Roeder, and Lawrence E Hunter. The structural and content aspects of abstracts versus bodies of full text journal articles are different. BMC bioinformatics, 11(1):492, 2010.

[26] Inderjit S Dhillon, Subramanyam Mallela, and Dharmendra S Modha. Information-theoretic co-clustering. In Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 89–98. ACM, 2003.

[27] David A Jurgens, Peter D Turney, Saif M Mohammad, and Keith J Holyoak. Semeval-2012 task 2: Measuring degrees of relational similarity. In Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pages 356–364. Association for Computational Linguistics, 2012.

[28] Philippe Thomas, Stefan Pietschmann, Illés Solt, Domonkos Tikk, and Ulf Leser. Not all links are equal: Exploiting dependency types for the extraction of protein-protein interactions from text. In Proceedings of BioNLP 2011 Workshop, pages 1–9. Association for Computational Linguistics, 2011.

[29] Palakorn Achananuparp, Xiaohua Hu, and Xiajiong Shen. The evaluation of sentence similarity measures. In International Conference on Data Warehousing and Knowledge Discovery, pages 305–316. Springer, 2008.

[30] Min Zhang, Jian Su, Danmei Wang, Guodong Zhou, and Chew Lim Tan. Discovering relations between named entities from a large raw corpus using tree similarity-based clustering. In International Conference on Natural Language Processing, pages 378–389. Springer, 2005.

[31] Takaaki Hasegawa, Satoshi Sekine, and Ralph Grishman. Discovering relations among named entities from large corpora. In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, page 415. Association for Computational Linguistics, 2004.

[32] Md Arafat Sultan, Steven Bethard, and Tamara Sumner. Dls@ cu: Sentence similarity from word alignment. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 241–246, 2014.

[33] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? Perspectives on psychological science, 6(1):3–5, 2011.

[34] Kevin Bretonnel Cohen and Dina Demner-Fushman. Biomedical natural language processing, volume 11. John Benjamins Publishing Company, 2014.

[35] Natalia Silveira, Timothy Dozat, Marie-Catherine De Marneffe, Samuel R Bowman, Miriam Connor, John Bauer, and Christopher D Manning. A gold standard dependency corpus for english. In LREC, pages 2897–2904. Citeseer, 2014.

[36] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In EMNLP, volume 14, pages 1532–1543, 2014.

[37] Karën Fort, Gilles Adda, and K Bretonnel Cohen. Amazon mechanical turk: Gold mine or coal mine? Computational Linguistics, 37(2):413–420, 2011.

[38] Raoul Frijters, Marianne Van Vugt, Ruben Smeets, René Van Schaik, Jacob De Vlieg, and Wynand Alkema. Literature mining for the discovery of hidden connections between drugs, genes and diseases. PLoS Comput Biol, 6(9):e1000943, 2010.

[39] Trevor Cohen and Dominic Widdows. Empirical distributional semantics: methods and biomedical applications. Journal of biomedical informatics, 42(2):390–405, 2009.

[40] Lawrence Hunter, Zhiyong Lu, James Firby, William A Baumgartner, Helen L Johnson, Philip V Ogren, and K Bretonnel Cohen. Opendmap: an open source, ontology-driven concept analysis engine, with applications to capturing knowledge regarding protein transport, protein interactions and cell-type-specific gene expression. BMC bioinformatics, 9(1):78, 2008.

[41] 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(6):462–477, 2003.

[42] Daniel Müllner et al. fastcluster: Fast hierarchical, agglomerative clustering routines for r and python. Journal of Statistical Software, 53(9):1–18, 2013.

[43] James Richard Curran. From distributional to semantic similarity. 2004.

[44] 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.

[45] David Sánchez, Albert Solé-Ribalta, Montserrat Batet, and Francesc Serratosa. Enabling semantic similarity estimation across multiple ontologies: An evaluation in the biomedical domain. Journal of biomedical informatics, 45(1):141–155, 2012.

[46] Francisco M Couto, Mário J Silva, and Pedro M Coutinho. Measuring semantic similarity between gene ontology terms. Data & knowledge engineering, 61(1):137–152, 2007.

[47] Judith Blake. Bio-ontologiesfast and furious. Nature biotechnology, 22(6):773–774, 2004.

[48] Han Rauwerda, Marco Roos, Bob O Hertzberger, and Timo M Breit. The promise of a virtual lab in drug discovery. Drug Discovery Today, 11(5):228–236, 2006.

[49] Timo Jarvinen and Pasi Tapanainen. Towards an implementable dependency grammar. In CoLing-ACL'98 workshop'Processing of Dependency-Based Grammars', Kahane and Polguere (eds), pages 1–10, 1998.

[50] Rion Snow, Daniel Jurafsky, Andrew Y Ng, et al. Learning syntactic patterns for automatic hypernym discovery. In NIPS, volume 17, pages 1297–1304, 2004.

[51] Michael A Tanenblatt, Anni Coden, and Igor L Sominsky. The conceptmapper approach to named entity recognition. In LREC, 2010.

[52] Fei Wu and Daniel S Weld. Open information extraction using wikipedia. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 118–127. Association for Computational Linguistics, 2010.