ANAPHORA RESOLUTION BASED ON SEMANTIC RELATEDNESS IN THE BIOMEDICAL DOMAIN

Mengqian Wang

A thesis submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Arts in the Department of Linguistics in the School of Arts and Science.

Chapel Hill
2018

Approved by:

Katya Pertsova

Elliott Moreton

Jules Michael Terry

© 2018
Mengqian Wang
ALL RIGHTS RESERVED

ABSTRACT

Mengqian Wang: Anaphora Resolution Based on Semantic Relatedness in the Biomedical Domain
(Under the direction of Katya Perstova)

In Linguistics, an anaphor is an expression whose interpretation depends upon another expression in context, namely an antecedent expression. Anaphora resolution is a task of identifying the anaphorical relation between the anaphor and its antecedent. Anaphora resolution is used in many high-level tasks of Natural Language Processing. Traditionally, the rule-based approaches to anaphora resolution rely on the syntactic structures and discourse features. In my study, I implement two semantic approaches on biomedical texts, ontology-dependent method and ontology-independent vector semantic method. The ontology-dependent method will be used to locate the antecedent for noun phrases with determiners while the ontology-independent method will be implemented on pronouns. The results show that the semantic approaches are promising directions in investigating resolutions for anaphora problems in the future.

To my mentor and friend, I couldn't have done this without you.
Thank you for all of your support along the way.

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Dr. Katya Pertsova of the Department of Linguistics at the University of North Carolina at Chapel Hill. The door to Prof. Pertsova office was always open whenever I ran into a trouble spot or had a question about my research or writing. She consistently allowed this paper to be my own work, but steered me in the right the direction whenever she thought I needed it.

I would also like to acknowledge Dr. Elliott Moreton and Dr. Jules Michael Terry of Department of Linguistics at the University of North Carolina at Chapel Hill as the committee members of this thesis, and I am gratefully indebted to them for their very valuable comments on this thesis.

Finally, I must express my very profound gratitude to my parents and to my friends for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. Introduction

In Linguistics, an anaphor(Huddleston, 1984) is an expression whose interpretation depends upon another expression in context, namely an antecedent expression. For example, in the sentence, *Sarah had made the decision, but everyone thinks she was wrong*, the pronoun *she* is the anaphor which refers back to its antecedent *Sarah*. The terminology “anaphor” in this paper is used slightly differently from the one used in traditional Linguistics. The interpretation of some noun phrases with definite or demonstrative determiners also depends on their antecedents. And practically, the resolution for these phrases are also important. Thus, in this paper, I will also take these noun phrases into account. For convenience, the terminology "anaphor" will be used to indicate pronouns and some noun phrases with determiners.

Anaphora is a common device in daily conversation to avoid repetition and help with communication. While identifying the antecedent of an anaphor is intuitive for human, it could be challenging for machines. The anaphora resolution(Mitkov, 2014) is the task of automatically finding the antecedent for an anaphor. This task is basic and used in many high-level tasks of Computational Linguistics and Natural Language Processing. Without anaphora resolution, the efficiency of information extraction could be largely impaired. We take SemRep as a concrete example.

SemRep (Rindflesch & Fiszman 2003) is a program that extracts semantic predications (subject–relation–object triples) from a biomedical free text. Here is an example which is given on the SemRep official website:

(1).
1. We used hemofiltration to treat a patient with digoxin overdose that was complicated by refractory hyperkalemia.
2. Hemofiltration-TREATS-Patients
Digoxin overdose-PROCESS_OF-Patients
hyperkalemia-COMPLICATES-Digoxin overdose
Hemofiltration-TREATS(INFER)-Digoxin overdose

From the sentence in (1)1, SemRep extracts the predications in (1)2. Previous studies have found that the failure of detecting the anaphorical relationship may lead to a loss in information extraction. And the failure will thus affect the performance of SemRep. An example from Halil et al (2016) is given in (2).

(2).
There are currently 3 classes of drugs approved for the treatment of PAH: prostacyclin analogues, endothelin receptor antagonists, and phosphodiesterase type 5 inhibitors. …
Although definitive evidence will require randomized and properly controlled long-term trials, the current evidence supports the long-term use of these drugs for the treatment of patients with PAH.

The concept corresponding to these drugs in the UMLS Metathesaurus is "drugs". Thus, a non-informative relation will be extracted from the concluding sentence, as shown in (3).

(3). Drugs-TREATS-PAH

However, with a co-reference resolution, these drugs could be substituted by Prostacyclin analogues, endothelin receptor antagonists, and phosphodiesterase type 5 inhibitors. The relations in (4) will be extracted.

(4).
Prostacyclin analogues-TREATS-PAH
Endothelin receptor antagonists-TREATS-PAH
Phosphodiesterase type 5 inhibitors-TREATS-PAH

Therefore, implementing anaphora resolution will potentially expand the scope of information extracted by SemRep.

Traditionally, the approaches in anaphora resolution rely heavily on syntactic features. In my study, I would investigate the anaphora resolution from a perspective of semantics. A real human will consider not only the syntactic structure but also the semantic meaning of the antecedent and the context of the anaphor when he/she is trying to figure out an anaphora relationship. For example, in the sentence (5), "it" potentially refers to "a dog" or "an apple". Human beings can easily deduce the true antecedent of "it" is "dog" since apples do not die. That means, "died" is more semantically related to "dog" than to "apple". We may expect to see "dog" and "died" co-occur more frequently in texts.

(5). *A dog ate an apple. It died.*

Now let's change sentence (5) to sentence (5)'.

(5)' . *A dog ate an apple. This poor animal died.*

The anaphor in (5)' is "this poor animal". We may find some extra information from the anaphor itself in this sentence. Since a dog is a type of animal, we would expect the antecedent of the anaphor is "a dog". Thus, it would make sense to try to find the anaphora relationship from a perspective of semantics. In my study, two types of semantic approaches will be implemented – ontology-dependent method and ontology-independent vector cosine similarity method. Using which one of the approaches depends on the type of anaphor.

In the following parts of the introduction section, I will introduce the types of anaphora that I am going to investigate in my study and the semantic approaches that I will implement in detail. Then, I will briefly review the previous studies on anaphora resolution and I will state the reasons why I think semantics is worth to try. After that, I will give detailed descriptions of the database and tools that I will use in the experiment.

The second section will be the body of the experimental procedure. By reading through this section, you should be able to see how I did my experiment and tested my hypothesis. The results and evaluations will be included in section three. I will generally discuss the results and the future direction in the last part of the paper.

The anaphora discussed in this paper include two types, pronouns and some noun phrases with determiners. Personal pronouns in English are given in table 1. Demonstrative pronouns including *this, that, these* and *those* are not shown in table 1 but will be included in my study. These four pronouns can also be used as determiners in noun phrases. For example, in the phrase *this dog,* "this" functions as a determiner. I will also consider the NPs with the definite determiners as anaphora. For example, *the black cat* is a noun phrase composed of a determiner *the* and a noun phrase *black*

cat. A special case is *which*. If *which* shows up by itself, it will be treated as a pronoun while if it shows up in a noun phrase, it will be treated as a demonstrated determiner. To sum up, the anaphora I will consider include pronouns and some noun phrases with definite and demonstrative determiners. *Which* as a special case will be treated as a demonstrative pronoun or demonstrative determiner.

Person	**Number**	**Case**		
		Subject	**Object**	**Possessive**
First	**Singular**	I	me	mine
	Plural	we	us	ours
Second	**Singular**	you		yours
	Plural			yours
Third	**Singular**	he	him	his
		she	her	hers
		it		its
	Plural	they	them	theirs

Table 1 pronouns in English

II. Previous Studies

Charniak (1972), Winograd (1972) and Hobbs (1974) introduced first NLP approaches to anaphora resolution which are heavily based on commonsense knowledge. The chief value of Charniak's work has been to show how difficult the pronoun resolution problem is. He showed in his 1972 paper a large amount of difficult cases in understanding children's stories. For deducing the correct antecedent of a pronoun in a child's story, arbitrarily detailed world knowledge could be required. Winograd was the first to write procedures for locating antecedents. He rated all possible referents on the basis of syntactic position. A subject is favored over an object as an antecedent, while both of them are favored over a complement of a preposition. The rating was very similar to the one proposed by Hobbs (1974) in his Naïve Algorithm which is the most representative syntax-based algorithm of pronoun resolution. Interestingly, in the 1978 paper, Hobbs suggested that this approach was very limited and he proposed that semantics could be a possible solution in the future. More recent studies focus on statistical and machine learning approaches. ARPA's Message Understanding Conference (MUC, 1992-1997), the first big initiative in Information Extraction, changed NLP by producing the first annotated data for tasks including "name entity extraction" and "co-reference". "Coreference chain" is a terminology which was created and firstly used by MUC and it indicates a set of mentions referring to an entity. Anaphora is a special case of co-reference. They are closely related but not exactly the same. While co-reference describes the situation

that different entities refer to a same concept, anaphora means that an anaphor is referring to its antecedent. An anaphor is semantically empty, which means it is practically useless if it fails to refer to an antecedent. These two terms are more or less used alternatively in recent years, which causes confusion. Some of the researches I take reference in my study use "co-reference" to refer to "anaphora". I will be consistent with the authors for their usage of these two terms. Without extra illustration, both of "co-reference" and "anaphora" refer to "anaphora" in my study.

More attention has been attracted to anaphora and co-reference resolutions since MUC. The earliest works on the co-reference resolution based on machine learning are McCarthy & Lehnert (2000) and Soon et al. Haghighi and Klein proposed a deterministic co-reference system in their 2009 paper. This system was driven entirely by syntactic and semantic compatibility while most of the co-reference systems in the same period of time heavily relied on discourse constraints. They parsed all sentences with Stanford parser and extracted rich syntactic features that are representative for co-reference relations. To get semantic knowledge between mentions, they applied the syntactic features to other datasets including WIKI (25k English articles from Wikipedia) and BLIPP (1.8 million sentences of newswire parsed with the Charniak (2000) parser) and acquired the pairs of words that have comparative meanings. These pairs of words were later used for the purpose of disambiguation. Their experiment was dealing with the coreference resolution, which means they were not only finding the antecedents for anaphora but also finding the coreference relationships for entities. In coreference resolution, they were given a document which consists of a set of mentions; each mention is a phrase in the document and they were asked to cluster mentions according to the underlying

referent entity(Charniak, 1972). The combination of syntactic features with semantic knowledge, although simple than most of the contemporary approaches, made their approach (precision 87.2%, recall 77.3%, F-1 Score 81.9) outperform the best unsupervised approach (precision 83.0%, recall 75.8, F-1 Score 79.2) of the day and be comparative with the state-of-art supervised approach (precision 89.7%, recall 55.1%, F-1 Score 68.3) at that time. However, their study on semantic knowledge was on the string level and the semantic knowledge was used only as a tool of disambiguation. By looking into the concept and with the help of vector semantics, semantic knowledge could play a more important role in the co-reference and anaphora resolution.

Not only in general literature, the co-reference resolution has also been an important task in biomedical natural language processing. In the BioNLP shared task 2011, co-reference was launched as a supporting task. The goal was to find the gene/protein co-reference relations. 6 teams submitted their results while the champion team successfully grabbed 22.18% relations with a precision of 73.26%. Their approaches, as well as an approach developed by a team after the shared task, was mostly based on the syntactic structure and discourse salience. However, with the rich biomedical ontology system and a large number of biomedical texts, implementing semantic method would be potentially improving the performance. In my study, I will implement both ontology-dependent and ontology-independent methods on the data provided by the BioNLP shared task 2011. The data was well-annotated and released to the public.

III. The Semantic Approaches

There are two types of semantic approaches going to be used in my study, ontology-dependent method and ontology-independent vector semantic method. The ontology-dependent method relies on the biomedical ontology relations stored in the Unified Medical Language System (UMLS) (see section 1.4.1) and the ontology-independent method is based on vector semantics which relies on the words co-occurrence frequency in the biomedical texts (see section 1.3.1).

Also, as mentioned before, there are two types of anaphora I am going to examine in my study, pronouns and noun phrases with determiners. Since pronouns are semantically empty, looking at their semantic relations with other ontology is meaningless. Instead, I would acquire the semantic information from the context of them using vector semantic method. The vector semantic method will be introduced in 1.3.1. I will talk about the ontology-dependent method which is used to deal with the cases of the noun phrases with determiners in 1.3.2.

3.1 Vector Semantics

As Firth (1957) said in Studies in Linguistic Analysis, you shall know a word by the company it keeps. Words that occur in similar contexts tend to have similar meanings. It is reasonable to assume that an entity is semantically related to its context. We shall also assume that the antecedent of the anaphor has semantic

relatedness with the context of the anaphor since an anaphor is semantically empty and its interpretation depends upon its antecedent.

For example, in (5), "it" potentially refers to "dog" or "apple". Human beings can easily deduce the true antecedent of "it" is "dog" since apples do not die. That means, "died" is more semantically related to "dog" than to "apple". We may expect to see "dog" and "died" co-occur more frequently in texts.

(5). *A dog ate an apple. It died.*

In the sentence above, "dog" will be assigned as the true antecedent of the anaphor because "dog" has a higher semantic relatedness with the context of the anaphor. In example (5), the context contains one word, "died".

The way I calculated the semantic relatedness is called vector semantics. Vector semantics works with distributional methods, in which the meaning of a word is computed from the distribution of words around it. These words are generally represented as a vector or array of numbers related in some way to counts (Jurafsky & Martin, 2016). Table 2 shows a word-word matrix. Each row represents the vector of the word that we examine while each column represents the frequency of a word co-occurring with the examined word.

	aardvark	...	computer	data	pinch	result	sugar	...
apricot	0	...	0	0	1	0	1	
pineapple	0	...	0	0	1	0	1	
digital	0	...	2	1	0	1	0	
information	0	...	1	6	0	4	0	

Table 2 an example word-word matrix from Jurafsky & Martin (2016)

Each row of numbers will be represented as a vector in the vector space, the smaller the angle between the vectors is, the similar the two vectors are. That is to say,

the words represented by the two vectors that have a small angle are more likely to show up in the same context, namely, they are more semantically similar. The way we decide the angle between the vectors is by calculating the cosine score. The higher the cosine score is, the smaller the angle is. Figure 1 shows a simplified two-dimensional vector space for words "digital" and "information".

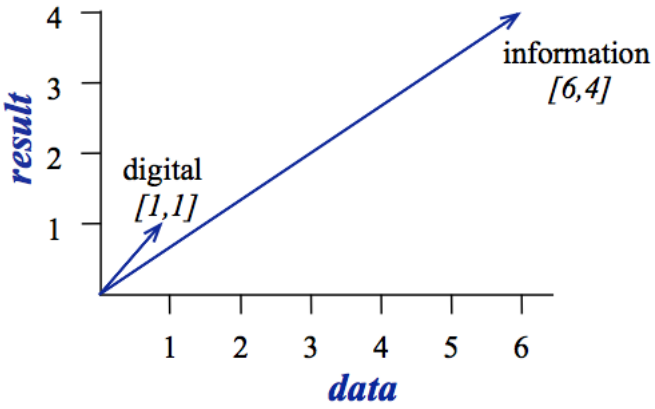

Figure 1 an example vector space from Jurafsky & Martin (2016)

The vector semantic method has been integrated into an open source software UMLS::Similarity (Mclnnes, Pedersen & Pakhomov, 2009). The true antecedent will be the one that has the highest cosine score with the context of the anaphor, which means the true antecedent will be semantically most related to the context of the anaphor.

3.2 The Ontology-Dependent Semantic Approach

For noun phrases (NPs) with determiners, the ontology-dependent method will be used since the N' under the NP gives us extra information. For example, in sentence (5)', which is a modified version of sentence (5), the anaphor is the NP "this poor animal" instead of the pronoun "it", we shall know the antecedent of the anaphor is

"dog" rather than "apple" because "apple" is not a type of "animal". To achieve this goal, I will take advantage of the Semantic Network in the UMLS. An illustration for the Semantic Network could be found in section 1.4.1.

(5)'. *A dog ate an apple. This poor animal died.*

IV. Database and Tools

In this section, I will introduce the detail of the database and tools and the ways I am going to implement them.

Generally, I am trying to find the true antecedent of an anaphor from a semantic perspective. First of all, If the anaphor is a personal pronoun or a demonstrative pronoun, I would calculate the semantic relatedness between the potential antecedents and the context of the anaphor by vector cosine similarity method. The method has been integrated into an open source software called UMLS::Similarity.

Secondly, if the anaphor is in some of the noun phrases with demonstrative or definite determiners, I will look up the concept relationship between the potential antecedents and the anaphor in the UMLS Semantic Network. They will be assigned as co-referenced if they have an *is-a* or a *parent-child* relationship. These relationships will be introduced in section 1.4.1.

The data used for evaluation are 83 annotated abstracts from the co-reference supporting task of BioNLP's shared task 2011. The anaphora are marked in the data. The potential antecedents and the entities in the context of the anaphor will be extracted by MetaMap, a supporting program that takes a plain text as input and returns a set of biomedical entities. The data source that backs up the MetaMap is the

UMLS Metathethurus. The UMLS Metathethurus will be introduced in section 1.4.1 while the MetaMap will be illustrated in section 1.4.2.

4.1 UMLS

According to the official website, the purpose of the Unified Medical Language System(UMLS) is to facilitate the development of computer systems that behave as if they "understand" the meaning of the language of biomedicine and health. There are three knowledge sources that mainly support the functioning of the UMLS, the Metathesaurus, the Semantic Network and the SPECIALIST Lexicon & Lexical Tools.

The Metathesaurus is a large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health related concepts, their various names, and the relationships among them. Concept Unique Identifier (CUI) is used to identify concept in the database.

The Semantic Network consists of semantic types and semantic relationships. There are 133 semantic types and 54 semantic relationships in the Semantic Network. The primary link between most semantic types is the *is-a* relationship. The *is-a* relationship is a type of semantic relationship in the Semantic Network of the UMLS. It can be illustrated as “is an instance of”. Part of the Semantic Network is shown in figure 1.

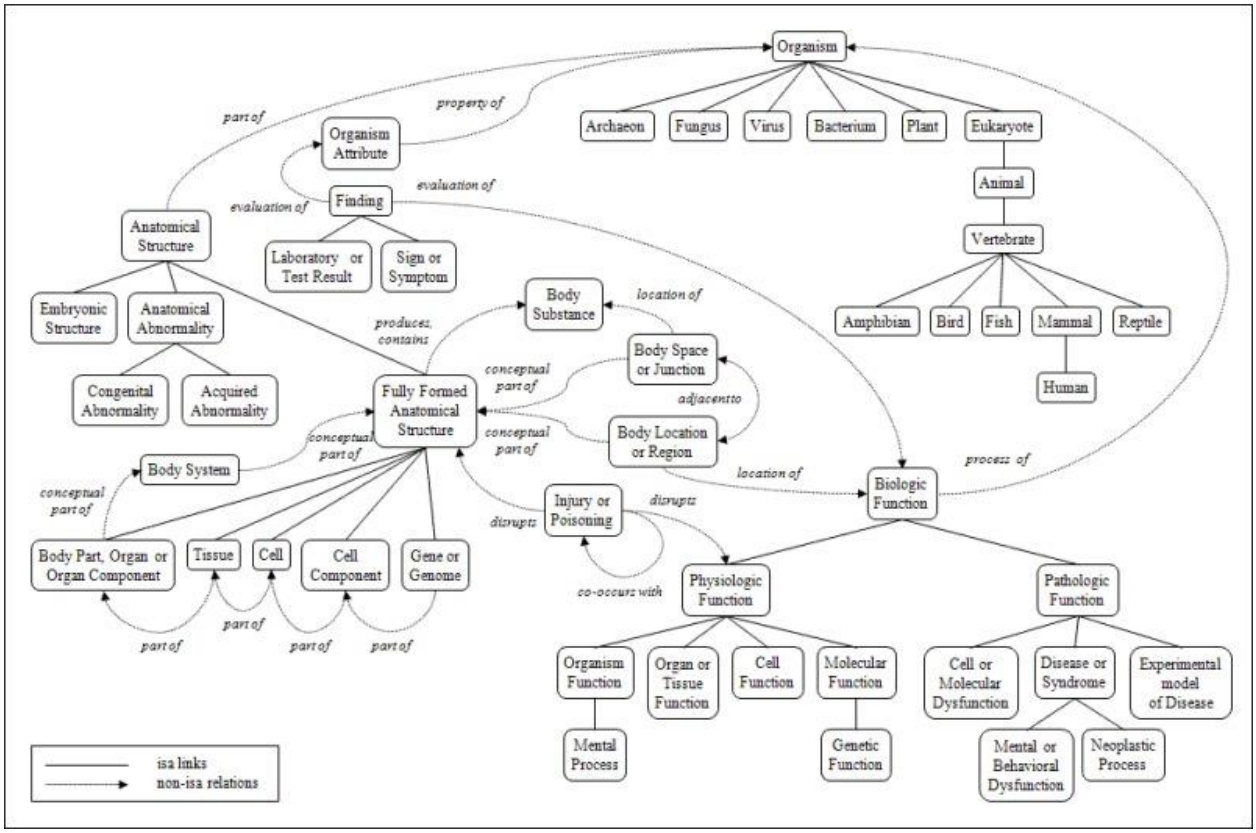

Figure 2 Part of the Semantic Network (from the UMLS basic tutorials)

The *is-a* relationship establishes the hierarchy of types within the Semantic Network and is used for deciding on the most specific semantic type available for assignment to a Metathesaurus concept. For example, in figure 2 (Liu et al, 2012) *doctor* and *physician* have an *is-a* relationship, i. e. *a physician* is an instance of a *doctor*.

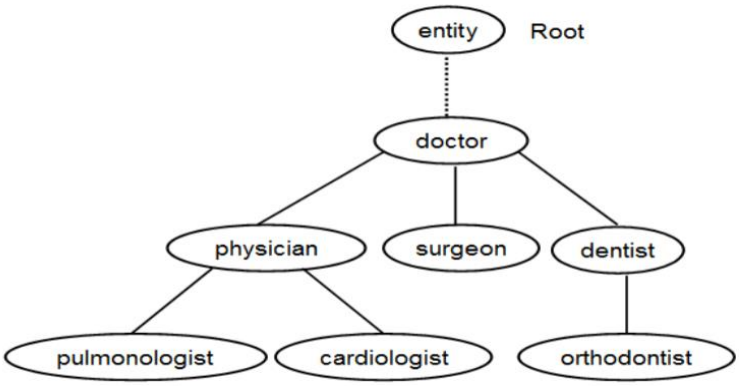

Figure 3 A sample hierarchical structure in the Semantic Network (from Liu et al, 2012)

Parent-child (broader/narrower) is another important relationship in the Semantic Network. Child (narrower) concept can be seen as a subtype of the Parent (broader) concept. Here is an example from the UMLS basic tutorials. The semantic type *Biologic Function* is the parent of, or broader than, the semantic type *of Physiologic Function*.

One thing to be noted is that a parent-child relationship only works for the concepts that are immediately connected in the Semantic Network. For example, in figure 2, *"doctor*" is the parent of *"physician*" while *"physician*" has two child concepts – *pulmonologist* and *cardiologist*. We do not say *"doctor*" is the parent of *pulmonologist* and *cardiologist*. However, *pulmonologist* and *cardiologist* each has an is-a relationship with *"doctor*", i.e. *pulmonologist/cardiologist* is an instance of *"doctor*". The UMLS Metathesaurus and the Semantic Network are accessible through a wrapper of Python.

The SPECIALIST Lexicon & Lexical Tools are in another part of UMLS which allows users to develop Natural Language Processing programs. However, the main parts of the UMLS used in this experiment are the Metathesaurus and the Semantic Network.

The UMLS is accessed through a Python wrapper called PyMedTermino which is an open source software developed by Lamy, Venot and Duclos (2015).

4.2 MetaMap

MetaMap is a supporting program for mapping the terms to the concepts in the UMLS. It takes raw texts as input and returns a set of concepts and the terms of which

the corresponding concepts have been found. MetaMap is a program built on Java. It is accessible by command-line and also has Java API and wrapper of Python developed by Anthony Rios (https://github.com/AnthonyMRios/pyMetaMap).

4.3 Semantic Relatedness and UMLS::Similarity

The ontology-independent method implemented in this study is vector cosine similarity which has been introduced in section 1.3.1.

This method is implemented and available in UMLS::Similarity (Mclnnes, Pedersen & Pakhomov, 2009), an open source software package written in Perl which integrates several well-tested methods of computing semantic similarity and relatedness between concepts in the UMLS. UMLS::Similarity needs UMLS::interface as a foundation package. The latter provides an API to a local installation of the UMLS in a MySQL database, as well as command line programs to allow a user to interactively explore the UMLS.

V. The Procedure of the Experiment

The data used for testing comes from BioNLP's shared task 2011. Annotations of co-reference relations are available under their co-reference supporting tasks. Figure 5 shows an annotated example in its training set. Arrows indicate the co-reference relation.

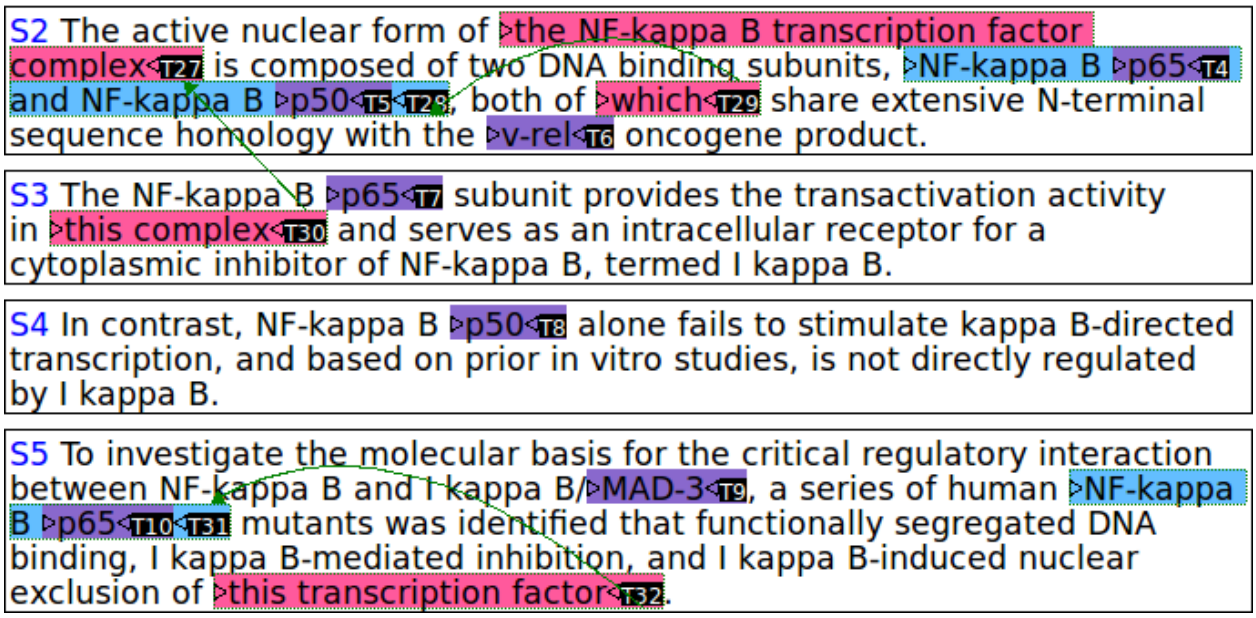

Figure 4 Coreference annotation from BioNLP's shared task 2011

A detailed illustration of the procedure of my experiment is given below. Figure 5 shows a visual illustration of the procedure of the experiment. A close examination of the example in figure 4 could be found in Appendix I. The example in Appendix I shows that the antecedent of the anaphora has semantic relatedness with the context of the anaphora. And calculating the semantic relatedness by the method of second-order cosine similarity is a feasible way to figure out the anaphorical relationship.

As a reminder, MetaMap is a supporting program which takes the plain text as input and returns a set of terms that could be mapped to the concept of the UMLS. I will use the term “extraction” in the following section to indicate this process of MetaMap.

The data in my experiment comes from the training data of the co-reference supporting task of the BioNLP’s shared task 2011. The anaphora and the co-reference relationship between anaphora and antecedents were annotated by a group of experts from the shared task. The annotations of the anaphora and the coreference relationships were produced based on the GENIA-MedCo coreference corpus, which is a product of collaboration between GENIA project and MedCo Annotation Project (from http://2011.bionlp-st.org/home/protein-gene-coreference-task). I did not use the testing data from the shared task for the following reasons. First of all, the goal of the shared task was to find the co-reference and anaphorical relations for gene and protein name entities, which means in the test data, only gene and protein relations were marked. However, since I was working on the general anaphorical relations in the biomedical text, the annotations in the testing data were not appropriate for me. Secondly, the main purpose of my study is to figure out whether an anaphorical relationship could be decided by semantic relatedness between an anaphor and its antecedent. An additional task of recognizing anaphor would be a distraction and would be beyond the scope of this study.

I divided the documents in the training data of the shared task into training and testing set. I manually examined the data in the training set and decided the rules that are implemented in steps a to c. The steps marked in red in figure 5 require human intervention And the steps from a to c can be done automatically. After

implementing my method and acquiring the results, I calculated the precision for evaluation and investigated the results for further analysis.

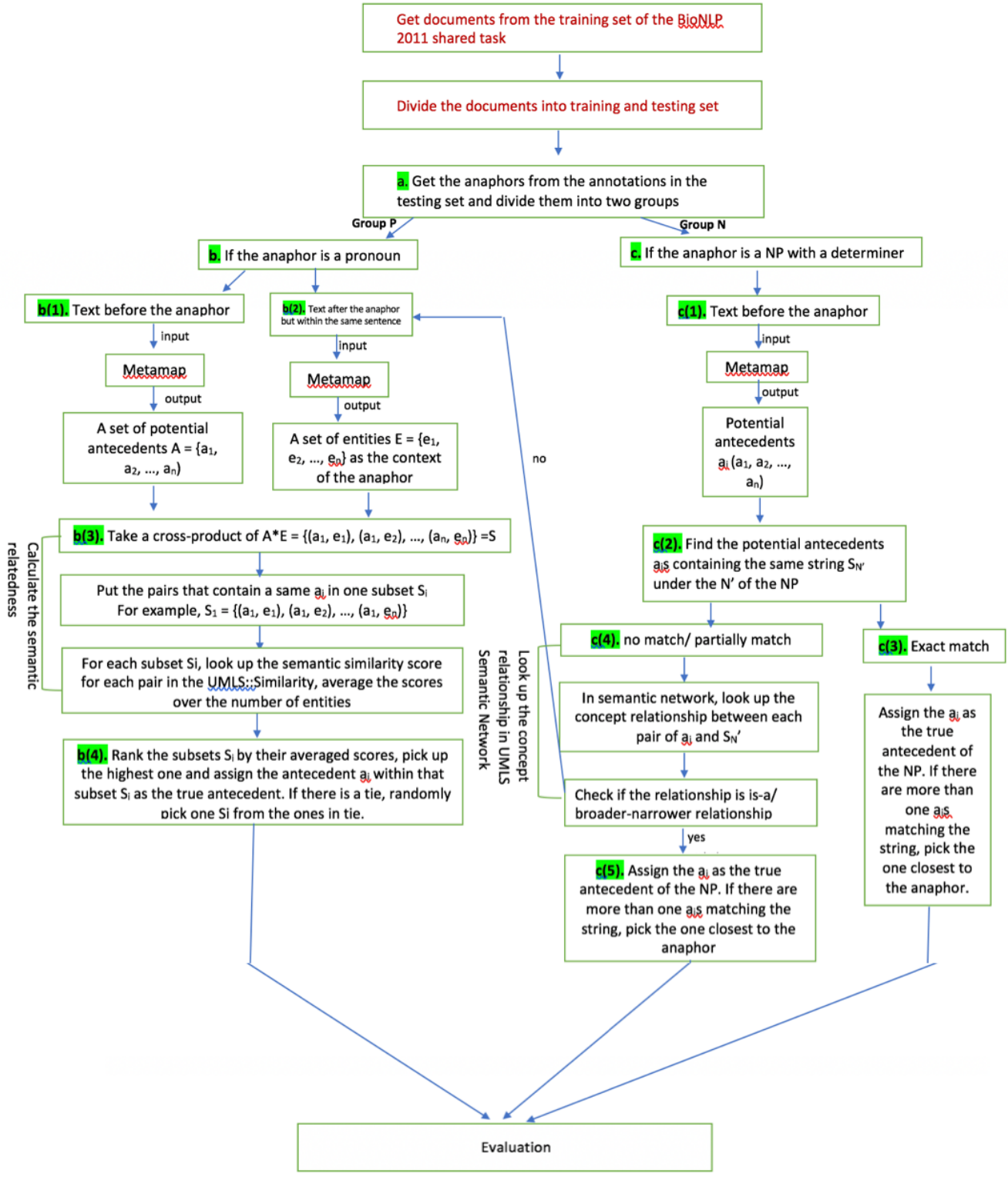

Figure 5 Visual illustration for the procedure of the experiment

First of all, my system pre-processed the text. In the pre-processing step, the anaphora which are marked in the dataset were divided into two groups. Group P included pronouns and group N included noun phrases with determiners. This step was done with the help of MetaMap. The anaphor would be considered as a noun phrase with a determiner if it had one or more corresponding concepts in the UMLS that could be extracted by MetaMap. Otherwise, the anaphor would be considered as a pronoun because a single pronoun does not have a corresponding concept stored in the UMLS. Some of the noun phrases in my dataset did not have corresponding concepts in the UMLS. These noun phrases were added to group P by my algorithm and treated as pronouns. The pre-processing step corresponding to the step *a* in figure 5 was done in Python automatically.

After the pre-processing, the two semantic approaches were implemented separately according to the type of anaphor. The steps under *c* correspond to the ontology-dependent method which was applied to group N (noun phrases with determiners). The steps under *b* were, on the contrary, illustrating the ontology-independent vector semantic method, which was applied to group P (pronouns). In practice, the steps under *c* were less time-consuming than steps under *b,* thus, I worked on the ontology-dependent part first in my experiment. I will keep the same order when I illustrate in the paper.

Group N:

A noun phrase is not semantically empty and at least one concept would be extracted from the anaphor by MetaMap. To make the result accurate, I got rid of the complement of the noun phrase. Everything in the anaphor that was after a proposition or a complementizer "that"was considered as a complement. “That” would

be seen as a complementizer if it did not show up at the beginning of the phrase. Then for the rest of the anaphor, I took the concept that was extracted from the rightmost of the phrase as the main concept because, in English, a noun phrase is usually right-headed if there is no complement.

I used the ontology-dependent method for these anaphora in group N. Firstly, I considered the text before the anaphor as the source of potential antecedents. I input the source text into MetaMap. MetaMap returned me a set of concepts and the phrases in the text that triggered the extraction. I checked the phrases to see if any of them has exactly the same string of the anaphor. If an exact match had been found, the phrase would be assigned as the true antecedent. Otherwise, through the semantic network in the UMLS, I would check if any the phrases have concepts that are *parents or children* of the concept of the anaphor. If no such a relation existed, I would expand the scope to *is-a* relation which allows a longer distance between the two concepts. If I found that more than one phrase had the *parent-child* or *is-a* relation with the anaphor, I would search for the phrase that partially matched the anaphor or randomly chose one if I failed to find a partial match. The anaphor that failed to get an antecedent from the above steps will be added to group P. The steps for noun phrases with determiners that correspond to the steps under *c* were done in Python automatically.

Group P:

The group P included personal pronouns, demonstrative pronouns and the noun phrases with determiners which I failed to find their antecedents by an ontology-dependent method. The ontology-independent vector semantic method was implemented on Group P. The first step of getting potential antecedents was as the

same as the ontology-dependent method. However, instead of getting the concept of the anaphor, I extracted the concepts from the context of the anaphor. The window of the context was the text after the anaphor but within the same sentence. By MetaMap, a set of concepts were extracted from the text. Also, MetaMap returned the entities that trigger the extraction. Till this point, I had a set of potential antecedents and a set of context concepts and entities for each anaphor. Next, I got cross products for the antecedents and the entities. The cross products were stored in a set S of pairs, for example, (potential antecedent [a1], context entity [e1]), (potential antecedent [a1], context entity [e2]), etc. I output the pairs in the format of the script which was readable by Perl since the open software UMLS::Similarity is a program written in Perl. The similarity scores were calculated in Perl and were output in .txt files again so that the results could be input back and analyzed by Python. At this point, I had the scores between the potential antecedents (a_1, a_2, ...a_n) and the context entities (e_1, e_2...e_n). After that, I put the pairs that contained the same a_i in one subset Si, For example, S_1 = {(a_1, e_1), (a_1, e_2), ..., (a_1, e_n)}. Then, for each subset S_i, I averaged the scores over the number of entities e_i. The reason why I did the average was to make sure that each entity in the context had an even contribution. The potential antecedent that achieved the highest score of similarity was assigned as the true antecedent.

VI. Results and Evaluation

The data of my experiment came from the data of BioNLP's Shared Task 2011. The anaphora were annotated and my task was finding the true antecedents for those anaphora. The results are shown in table3 through table5. In this experiment, all the cases (anaphor-antecedent relations) were positive and I calculated the precision (the percentage of true positive) for evaluation.

	Exact string matches	Partial string matches with the is_a relationship	No string match but with the is_a relationship	Number in total
Number of all members	11	21	10	42
Number of true positive	5	18	7	30
Percentage of precision	45%	86%	70%	71%

Table 3 Results for the ontology-dependent method on Group N

There were 83 abstracts among which 270 anaphora were annotated. 42 of them were noun phrases with determiners, 30 antecedents were correctly found in my experiment. The precision value was 71%. Among all the cases, 11 of them were found by exact string match and 31 of them were found by looking up the semantic relationship. Most of the true positive antecedents were the ones that are partially matched and semantically related to the anaphora.

Anaphor	Number of anaphora	Number of antecedents that are correctly assigned	Percentage of precision
pronouns	228	42	18%

Table 4 Results for ontology-independent vector semantic method on Group P

anaphor	Number of the true antecedents that achieve the top 5% scores among all potential antecedents	percentage
pronouns	176	77%

Table 5 Results for Further Analysis on Group P

However, in the rest 228 pronouns, only 42 of the assigned antecedents matched the true antecedents. The precision was as low as 18%. The antecedents were assigned according to the similarity score with the context of the anaphor. To get a further insight of my results, for each anaphor, I tried to compare the similarity score of the annotated antecedent and the scores of other potential antecedents. The results showed that 77% of the scores of the annotated antecedents fell in the top 5% of all the potential antecedents ($p < 0.05$), which means that, although not the top one, most of the true antecedents were in the top range. Just to clarify, in my dataset, the amount of the potential antecedents for one anaphor is usually in a range of 30 to 50. The results are shown in table 4.

VII. Discussion and Future Study

While the results of the study did not perfectly support my hypothesis, they do show ontology-dependent and ontology-independent vector semantics are possible solutions for anaphora resolution.

The ontology-dependent method which looks at string match and the semantic relationship between the concept of the anaphor and the concept of its antecedent is reliable. The anaphor with a Noun Phrase is usually in the same semantic category of its antecedent. Moreover, I also found that the concept extracted from the anaphor is usually broader than the one extracted from its antecedent, which means the antecedent is usually the hyponym of the anaphor. This could be a reason why most of the true antecedents were not the exact string matches with the anaphor. If the two phrases are identical, they should correspond to the same concept in the UMLS. Also, one of the important usage of anaphora is to avoid repetition. It is unlikely for people to use exact the same string of words as an anaphor of a phrase.

As in Haghighi and Klein's (2009), the coreference relationships were located by applying rich syntactic features with semantic knowledge for disambiguation. On the contrary, in my study, I tried to figure out the antecedents of the anaphora by mainly looking at the semantic relatedness. Unfortunately, my system does not perform as good as Haghighi and Klein's (2009) in terms of precision. One possible way to improve the performance of the ontology dependent method would be doing a syntactic parsing for the anaphor before extracting the concept. The structure of

complement is more complex than simply proceeding by a proposition or a complementizer “that”.

Vector semantics has been largely used in document classification. I am not aware of any anaphora resolution in the biomedical literature based on vector semantics. The ontology-independent method was imperfect by itself. However, the further analysis showed that the true antecedent has a high semantic relatedness with the context of the anaphor although not always the highest. The semantic relatedness is definitely a useful parameter that should be taken into account in the future study of anaphora resolution, especially in a domain with specific domain language. The domain language and terminologies are usually defined by human beings and are with less ambiguity compared to general language. For example, in the domain of biomedicine, the domain knowledge provided by the well-maintained dictionaries could be an advantage.

In my experiment, for each potential antecedent of an anaphor, I averaged the scores between that potential antecedent and entities in the context of the anaphor to make sure that every entity contributes evenly to the meaning of the context. However, in fact, some of the entities should be more contributive than the others. It was unfair to consider them as the same. Finding a way to weight the entities is a possible direction for my further study. Nevertheless, the window of the source text from which the context entities were extracted should be tackled. I originally planned to use the smallest main clause that contains the pronoun because the smallest main clause represents the smallest integrate semantic domain. The limitation of this study does not allow a syntactic parsing so that the main clause was not tagged.

The results are not possible to compare with the other teams' results in the shared task because I did not use the testing data for evaluation. The reasons have been mentioned in section II. First of all, the goal of the shared task was to find the co-reference and anaphorical relations for gene and protein name entities, which means in the test data, only gene and protein relations were marked. However, since I was working on the general anaphorical relations in the biomedical text, the annotations in the testing data were not appropriate for me. Secondly, the main purpose of my study is to figure out whether an anaphorical relationship could be decided by semantic relatedness between an anaphor and its antecedent. An additional task of recognizing anaphora would be a distraction and would be above the scope of this study. As mentioned in the overview of the shared task, locating the real anaphor is challenging especially for the definite noun phrases. Most of the teams gave up on finding those noun phrases even though those phrases took the largest part of the anaphora. Although on different datasets, if we look at the precision, my approach(71% precision of ontology-dependent method and 18% precision of ontology-independent vector semantic method) is still too simple and too coarse to outperform the teams(best at 73.26% precision) in the shared task.

My study shows that, for getting the full information of a word or a text, the string itself is not the only thing to investigate. We should take into account the words or texts surrounding it as well as the knowledge behind it. The ontology-dependent semantics and the ontology-independent vector semantics are promising approaches in dealing with anaphora problems and other topics under Linguistics and Natural Language Processing.

Appendix I

The example abstract is quoted in A.I.(1) with four sentences marked as 1, 2, 3 and 4.

A.I.(1).

1. The active nuclear form of the NF-Kappa B transcription factor complex is composed of two DNA binding subunits, NF-kappa B p65 and NF-kappa B p50, both of which share extensive N-terminal sequence homology with the v-rel oncogene product.
2. The NF-Kappa B p65 subunit provides the transactivation activity in this complex and serves as an intracellular receptor for a cytoplasmic inhibitor of NF-kappa B, termed I kappa B.
3. In contrast, NF-kappa B p50 alone fails to stimulate kappa B-directed transcription, and based on prior in vitro studies, is not directly regulated by I kappa B.
4. To investigate the molecular basis for the critical regulatory interaction between NF-kappa B and I kappa B/MAD-3 a series of human NF-kappa B p65 mutants was identified that functionally segregated DNA binding, I kappa B-mediated inhibition, and I kappa B-induced nuclear exclusion of this transcription factor.

The potential antecedents are found by inputting the raw text into MetaMap. The function of MetaMap has been illustrated in section 1.4.2. The result is shown in table 3. The leftmost column shows the term and the middle column shows the concept to which the term corresponds. In the rightmost column, the semantic type is given.

The original term in the text	**The corresponding term in the UMLS**	**The Semantic Type of the UMLS**
nuclear transcription factor complex	nuclear transcription factor complex	Cell component
NF-kappa B	NF-kappa B	Amino Acid, Peptide, or Protein,Immunologic Factor
NF-kappa B p65	Transcription Factor RelA	Amino Acid, Peptide, or Protein,Biologically Active Substance
NF-kappa B p50	Transcription Factor RelA	Amino Acid, Peptide, or Protein,Biologically Active Substance
DNA Binding	DNA Binding	dna binding
Sequence homology	Homology, sequence	Quantitative concept
v-rel Oncogene	REL gene	Gene or Genome

Protein Subunit	Protein Subunits	Amino Acid, Peptide, or Protein
NF-kappa B p65 Subunit	Transcription Factor RelA	Amino Acid, Peptide, or Protein,Biologically Active Substance
Transactivation	Trans-Activation, Genetic	Genetic Function
intracellular	Protoplasm	Cell Component
Receptor	receptor	Amino Acid, Peptide, or Protein,Receptor
NF-Kappa B Inhibitor	I-kappa B Proteins	Amino Acid, Peptide, or Protein,Immunologic Factor
Inhibitor of Kappa B	I-kappa B Proteins	Amino Acid, Peptide, or Protein,Immunologic Factor
Cytoplasmic	Cytoplasm	Cell Component
Term	Term Birth	Organism Function
IKappaB	I-kappa B Proteins	Amino Acid, Peptide, or Protein,Immunologic Factor
Transcription	Transcription, Genetic	Genetic Function
IKappaB/MAD-3	NFKBIA protein, human	Amino Acid, Peptide, or Protein,Biologically Active Substance
Human	Homo sapiens	Human
mutants	mutant	Cell or Molecular Dysfunction
segregations	Racial Segregation	Social Behavior
Inhibition	Metabolic Inhibition	Molecular Function
I-	Iodides	Inorganic Chemical
Nuclear Factor kappa B	NF-kappa B	Amino Acid, Peptide, or Protein,Immunologic Factor
Nuclear Factor I/B	NFIB gene	Amino Acid, Peptide, or Protein,Immunologic Factor

Table 6 The entities that map to the concepts in UMLS

The next step is distinguishing pronouns and NPs with determiners. There are three anaphora found in the sample text as underlined in (7), *which* in sentence 1, *this complex* in sentence 2 and *this transcription factor* in sentence 3.

Which

Which works as a pronoun while the other two are noun phrases with determiners. For *which* in the sentence 1, every entity within the same main clause have been extracted as *eis*. Thus, we have *e1* v-rel Oncogene and e2 sequence homology. In this experiment, every entity before the pronoun is considered to be potentially co-referenced with the pronoun. I calculated the semantic relatedness between each of the potential antecedents and entities. I show the result in table 4.

Entities before the mention *which* (Potential co-referenced entities)	**The score of relatedness with e1 v-rel Oncogene**	**The score of relatedness with e2 sequence homology**	**Average**
Nuclear transcription factor complex	0.836	0.2399	0.538
NF-kappa B	0.8108	0.3437	0.5772
NF-kappa B p65	0.6493	0.3255	0.4874
NF-kappa B p50	0.6732	0.2924	0.4828
DNA Binding	0.2099	0.2624	0.4743

Table 7 The score of semantic relatedness for mention >which< in the first sentence

NF-kappa B achieved the top one. It is considered as the true antecedent of *which*. However, this is not in accordance with figure 5. The true antecedent should be *NF-kappa B P65 and NF-kappa B p50* which is a combination of two entities. If we read the sentence 1 carefully we may see that the whole pronoun should be *both of which* instead of *which* by itself. However, the dataset fails to give this information. For the scope limit of this study, we will not consider the condition of antecedents containing more than one entities in the method of measuring semantic relatedness. The result still makes sense since NF-kappa B p65 and NF-kappa B p50 are subtypes of NF-kappa B.

This Complex

The anaphor *this complex* is a noun phrase with a demonstrative determiner. Entity *nuclear transcription factor complex* contains the string complex. The concept of the entity *nuclear transcription factor complex* is *nuclear transcription factor complex*. According to the UMLS database, *complex* has a child *protein complex* which is the parent of *transcription factor complex*. Moreover, *transcription factor complex* is the parent of *nuclear transcription factor complex*. From the above, there are three paths between *complex* and *nuclear transcription factor complex*. All of them indicate a parent-child (narrower/broader) relationship. Therefore, the entity *nuclear transcription factor complex* is the antecedent of *this complex*. This is half true according to figure 5. In the original text, the true antecedent of this complex should be *the NF-kappa B transcription factor complex.* After analyzing the phrase syntactically, I got the structure in A.I.(2).

A.I.(2)
[[**determiner** the] [[**modifier** NF-kappa B] [**NP** transcription factor complex]]]

NF-kappa B is a modifier of *the NP transcription factor complex* and thus syntactically congregated with the NP. However, in the UMLS, the phrase has been mapped to two separate concept *NF-kappa B* and *transcription factor complex*, which makes it impossible to automatically recognize the phrase as a single entity co-referencing with the anaphor. It is reasonable for the next step to combine the method of semantic relatedness with syntax-based algorithms. For the scope limit, we may consider these possibilities in the further studies.

This Transcription factor

Same as *this complex*, *this transcription factor* is a noun phrase in which a demonstrative determiner *this*. However, the entity *nuclear transcription factor complex* which partially matches the string "transcription factor" does not have any eligible path linking to *transcription factor* in the UMLS. The result makes sense from the syntactic point of view. The head of the phrase *this transcription factor* is *transcription factor*. The semantic meaning of a phrase is determined by its head. Thus, *this transcription factor* is a transcription factor. The only potential antecedent containing the sequence "transcription factor" is *nuclear transcription factor complex*. The head of this potential antecedent is *complex* instead of either *nuclear transcription factor* or *transcription factor*, which means this potential antecedent refers to a *complex* rather than a *transcription factor*. Hence, we will speculate the antecedent by calculating the semantic relatedness. The entities in the same main clause are extracted. The entities include *segregations, dna binding, IKappaB, Inhibition, I-, Nuclear Factor kappa B, Nuclear Factor I/B and Transcription*. The semantic relatedness is calculated and shown in table 5 and table 6. Due to the space limitation, the column of average is shown in table 7.

	e3 segregations	**e4 dna binding**	**e5 IKappaB**	**e6 Inhibition**
nuclear transcription factor complex	0.0336	0.4948	0.8679	0.4055
NF-kappa B p65	0.0816	0.5595	0.8172	0.4194
NF-kappa B p50	0.039	0.4157	0.7705	0.394
v-rel Oncogene	0.0018	0.2099	0.8385	0.326

Protein Subunit	0.1379	0.371	0.2704	0.7195
NF-kappa B p65 Subunit	0.0816	0.5595	0.8172	0.4194
Transactivation	0.029	0.2324	0.5052	0.2305
intracellular	0.1628	0.0919	0.6231	0.3226
Receptor	0.0683	0.3541	0.7075	0.5234
Cytoplasmic	0.044	0.1766	0.6522	0.2573
Term	0.122	0.1413	0.1089	0.257
IKappaB/MAD-3	0.0123	0.2844	0.9259	0.2208
Human	0.1081	0.1598	0.5103	0.2906
mutants	0.2337	0.13	0.2665	0.206

Table 8 The score of semantic relatedness for mention >with< in the fourth sentence (part1)

	e7 I-	**e8 Nuclear Factor kappa B**	**e9 Nuclear Factor I/B**	**e10 Transcription**
nuclear transcription factor complex	0.1866	0.9106	0.6676	0.6373
NF-kappa B p65	0.1986	0.8988	0.8155	0.7992
NF-kappa B p50	0.1812	0.8753	0.7049	0.7103
v-rel Oncogene	0.1812	0.8108	0.5634	0.4887
Protein Subunit	0.1992	0.4295	0.4973	0.4374
NF-kappa B p65 Subunit	0.1986	0.8988	0.8155	0.7992
Transactivation	0.1271	0.5738	0.5646	0.6835
intracellular	0.3863	0.5677	0.0965	0.0914
Receptor	0.3071	0.7318	0.4076	0.3574
Cytoplasmic	0.2609	0.6254	0.1748	0.1642
Term	0.1469	0.1849	0.136	0.1503
IKappaB/MAD-3	0.1854	0.8291	0.5729	0.5198
Human	0.3743	0.414	0.3248	0.1895
mutants	0.4446	0.2998	0.2387	0.3548

Table 9 The score of semantic relatedness for mention >with< in the fourth sentence (part2)

Entities before the mention this not including e1 and e2 (Potential co-referenced entities)	**Average**
nuclear transcription factor complex	0.5255
NF-kappa B p65	0.5737
NF-kappa B p50	0.5114
v-rel Oncogene	0.4275
Protein Subunit	0.3827
NF-kappa B p65 Subunit	0.5737
Transactivation	0.3683
intracellular	0.2928
Receptor	0.4322
Cytoplasmic	0.2944
Term	0.1559
IKappaB/MAD-3	0.4438
Human	0.2964
mutants	0.2718

Table 10 The score of semantic relatedness for mention >with< in the fourth sentence (part3)

In this case, *NF-kappa B p65* and *NF-kappa B p65 Subunit* are in tie and achieve the highest score. The reason for this scenario is that *NF-kappa B p65* and *NF-kappa B p65 Subunit* are mapped to the same concept in the UMLS. According to figure 5, *NF-kappa B p65* (or *NF-kappa B p65 Subunit*) and *this transcription factor* have a co-reference relation.

REFERENCES

Charniak, E. (1972). Toward a model of children's story comprehension. AI TR-266, Massachusetts Institute of Technology Artificial Intelligence Laboratory.

Jurafsky, D & Martin, H. J. (2016). "Chapter 15 Vector Semantics". *Speech and Language Processing*.

Haghighi, A. & Klein, D. (2009). Simple Coreference Resolution with Rich Syntactic and Semantic Features.

Kilicoglu, H., Rosemblat, G., Fiszman, M., & Rindflesch, T. C. (2016). Sortal anaphora resolution to enhance relation extraction from biomedical literature. BMC Bioinformatics, 17(1):1–16

Hobbs, J. (1978). Resolving pronoun references. Lingua, Volume 44, Issue 4, 311-338

Huddleston, R. (1984). Introduction to the Grammar of English. Cambridge University Press.

Lamy, J., Venot, A. & Duclos, C. (2015). "PyMedTermino: an open-source generic API for advanced terminology services." In MIE, pp. 924-928

Liu Y., Bill R., Fiszman M., Rindflesch T., Pedersen T., Melton G.B., et al. (2012). Using SemRep to Label Semantic Relations Extracted from Clinical Text. AMIA Annu Symp Proc. 587–95

Liu Y., McInnes B.T., Pedersen T., Melton-Meaux G. & Pakhomov S. (2012). Semantic Relatedness study using second order co-occurrence vectors computed from biomedical corpora, UMLS and WordNet. In Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. Miami, Florida, USA: ACM; 363–372.

MUC-6. (1995). Proceedings of the Sixth Message Understanding Conference. Morgan Kaufmann, San Francisco, CA.

MUC-7. (1998). Proceedings of the Seventh Message Understanding Conference. Morgan Kaufmann, San Francisco, CA.

Mccarthy, J. & Lehnert, W. (2000). Using Decision Trees for Coreference Resolution.

McInnes BT, Pedersen T, Pakhomov SV. (2009). UMLS-Interface and UMLS-Similarity: Open Source Software for Measuring Paths and Semantic Similarity. AMIA Annu Symp Proc. 431–435

Mitkov, R. (2014). Anaphora resolution. Routledge, 2014

Rindflesch, T.C. & Fiszman, M. (2003) 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

Rios, A. ()

Winograd, T. (1972). Understanding natural language. New York: Academic Press.