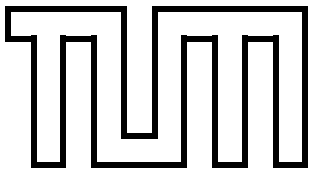


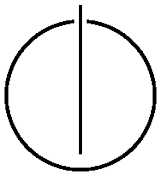
FAKULTÄT FÜR INFORMATIK

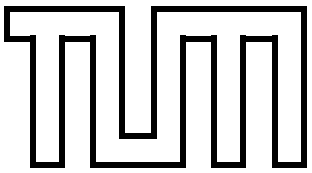
DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Coreference resolution in biomedical full text

Kujtim Rahmani





FAKULTÄT FÜR INFORMATIK

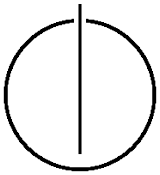
DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Master's Thesis in Informatics

Coreference resolution in biomedical full text

Koreferenz Auflösung in der biomedizinischen Volltext

Author:	Kujtim Rahmani
Supervisor:	Prof. Dr. Bukhard Rost
Advisor:	Mr. Juan Miguel Cejuela Pérez
Date:	December 12, 2014



I assure the single handed composition of this master's thesis only supported by declared resources.

Ich versichere, dass ich diese Masterarbeit selbständig verfasst und nur die angegebenen
Quellen und Hilfsmittel verwendet habe.

München, den 12. Dezember 2014 Kujtim Rahmani

Abstract

This thesis introduces a new coreference resolution system for biomedical texts. This system identifies and links expressions (definite nouns and pronouns), which refer to proteins. I focus on improving the current state-of-the-art in coreference resolution for abstracts and, a novel aspect, apply the system on the full-text articles. The system uses syntactic rules to resolve pronouns, and domain knowledge together with an ordered set of syntactic and semantic rules to resolve definite noun anaphoras.
During this thesis I collected statistics from the existing BioNLP corpus of abstracts and reviewed current coreference resolution methods in other non-biomedical corpora.

The system, in the BioNLP development set, achieves 75% precision, 68.3% recall and 71.5% F-score, which is an improvement of +4.1% of the current state of the art. In the test set the system achieves 60.92% precision, 65.53% recall and 63.14% F-score, which is an improvement of +2.24% of the current state of the art.

On full-text articles, the system achieves 82.3% precision, 61.35% recall and 70.2% F-score. These results are the first results in protein coreference resolution in full text articles.

Acknowledgments

I would first like to thank my supervisor Pr. Burkhard Rost for his support and opportunity to do my master's thesis in his group.

A great thanks goes to my thesis advisor Juan Miguel Cejuela. His guidance in research, organization, communication, and bioinformatics, was evident and encouraging in every moment of my thesis.

I want to extend my gratitude to the whole Rostlab group for creating a great environment to work at. Special thanks goes to Tim Karl, who kindly responded to all my requests.

Contents

Part I.

Introduction and Theory

1. Introduction

1.1. Coreference resolution

Coreference resolution is the process of determining whether two or more expressions in natural language refer to same entity in the real world [4]. Coreference resolution is an important and challenging task of natural language processing (NLP) [1][5]. It is believed to be useful in question answering, information extraction, relationship extraction and document summarization among others.

Example (1):

> *M-CSF treatment was also associated with a rapid induction of* ***the jun-B gene****, although expression of* ***this gene*** *was prolonged compared to that of c-jun.*

In example (1), ***"the jun-B gene"*** and ***"this gene"*** refer to same entity in the real world. Therefore, we can say that these two expressions corefer. Another characteristic of these two expressions is that the second expression is semantically dependent on the first.

This dependency of corefering expressions leads to a new linguistic phenomenon called anaphora. Linguists have given various different definitions for these two linguistic terms, which leads to a conflusion between coreference and anaphora. Sometimes researchers do not distinguish these two terms and say that they are synonyms [2]. For clarification, in the following paragraphs I will give the definitions of anaphora and coreference that we will use throughout the thesis.

Anaphora (anaphoric expression) is the linguistic phenomenon of pointing back to a previously mentioned expression in the text. The expression mentioned in the text which an anaphora points to is called the antecedent. Anaphora resolution is the process of finding the antecedent of an anaphoric expression [73]. In other words, it means finding pairs of anaphoras and antecedents. As illustration, in example (1) ***"the jun-B gene"*** is an anaphoric expression (anaphora) because it refers back to its antecedent, ***"this gene"***.

If the antecedent and anaphora refer to a same entity in the real world they are coreferential [73].

Gasperin defines anaphora as a directed relation between two linguistic expressions in the text, where the reader, in order to interpret the second expression, is referred back to the first one. In Figure 1.1 we can see the intersection and difference between coreference and anaphora. The intersection of these two terms is coreferent anaphora. Thus, a coreferent anaphora is when the anaphora is dependent from the antecedent and they refer to same entity.[49]

Associative anaphora is a relation between two expressions where the anaphora's interpretation depends on the antecedent, but nevertheless, they do not semantically refer to a same entity. In example 2, *"the door"* is an associative anaphora and its antecedent is *"the room"*, because to interpret the expression*"the door"* we need to refer back to the expression *"the room"*.

Example (2):

*The student entered in **the room**. He did not close **the door**.*

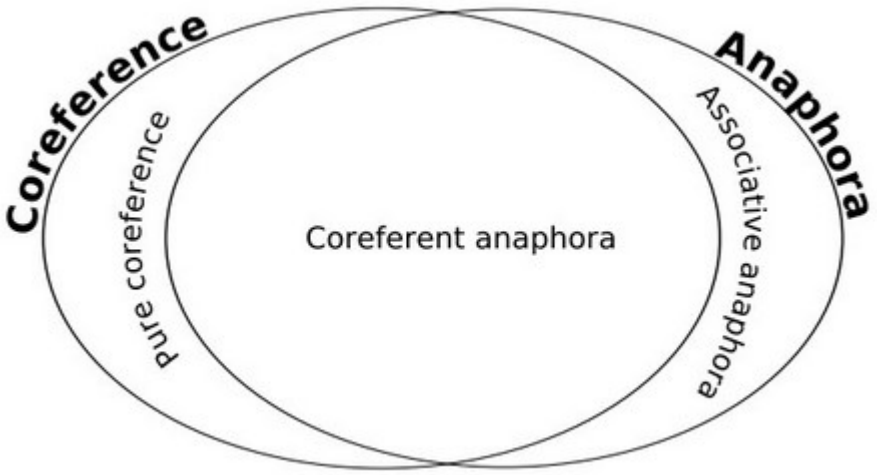


Figure 1.1.: Difference between anaphora and coreference [49]

Kees van Deemter and Kibble [24] define coreference relation of two expressions as the logical equivalence relation where both expressions *"refer to the same entity"*.

Mathematically, defining Referent(α_1) as the*"the entity referred to by α_1"*, a coreference relation for expressions α_1 and α_2 holds if and only if

$$\text{Referent } (\alpha_1) = \text{Referent } (\alpha_2)$$

Hence, coreference is an equivalence relation (reflexive, symmetric and transitive relations). From symmetric relation, the conclusion is that α_1 and α_2 are

semantically independent. Furthermore, the authors [24] define anaphoric relation of two expressions α_1 (antecedent) and α_2 (anaphora), if and only if α_2 depends on α_1 for its interpretation.

Summarizing, anaphora resolution is the task of finding mentions (anaphoras) that are dependent in their interpretation on a second expression (their antecedent). On the other hand, coreference resolution can be seen as a clustering task, where we must find all mentions in the text that refer to same entity.

1.2. Coreference resolution in Biomedical literature

In biomedical literature, there exists just one corpus that deals with coreference resolution. This corpus was published by the BioNLP group. In it, the community defined the task named Protein Coreference Resolution. I will describe in the following sections the detailed specification of this task.

1.2.1. Data description

The BioNLP corpus consists of 1210 annotated abstracts. The corpus is divided into 3 subsets: 800 abstracts in the training set, 150 in the development set, and 250 in the test set. Each document is annotated with proteins, anaphoric expressions and their references. For each abstract, three files exist:

- A text file, which contains the content of the abstract.
- A protein annotation file, which contains the annotated proteins.
- A relation file, which contains the coreference relation.

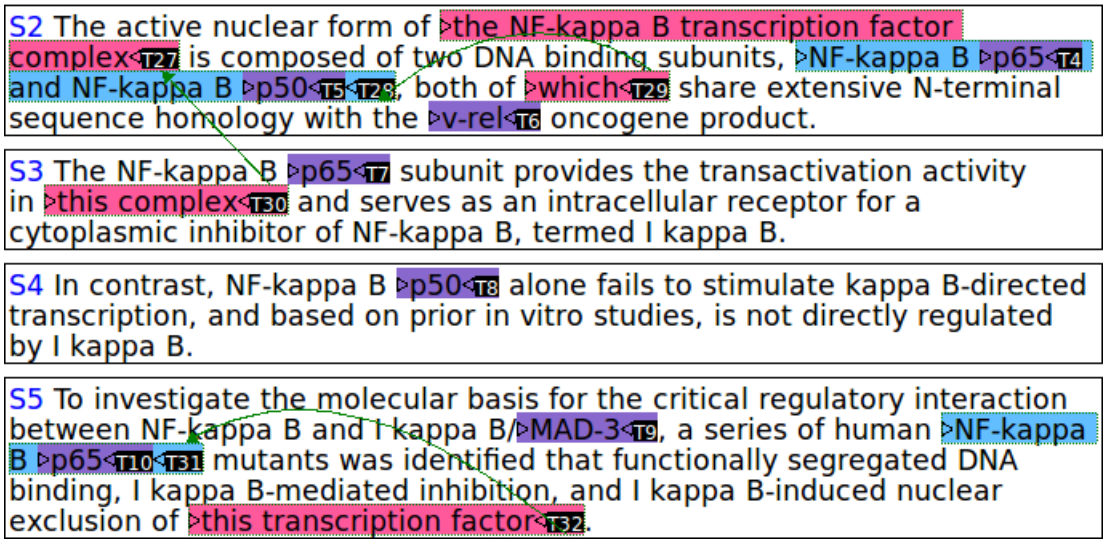


Figure 1.2.: Example of file annotation[1]

I will clarify the content of the protein annotation and coreference relation files by examples. In Figure 1 annotations T4-10, marked with purple color, are protein annotations. Every protein name mentioned in the text is annotated. Each annotation consist of its id, offsets (begin span and end span) and its string content, and this information is inserted in the protein annotation file as in Example (3).

Example (3):

```
T4    Protein 275 278    p65
T5    Protein 294 297    p50
T6    Protein 367 372    v-rel
T7    Protein 406 409    p65
T8    Protein 597 600    p50
T9    Protein 843 848    MAD-3
T10   Protein 879 882    p65
```

In Example (3), the first line indicates that in the text there is a protein reference ***"p65"*** *that begins at 275th character and its content ends before 278th character. This annotation is indexed by the id, "T4" [1].*

The coreference relation file contains three types of annotations: anaphoric annotations, their antecedent, and coreference relations. In Example (4), annotations T27-32 are of the first and second type and the R1-R3 describe corefer-

[1]Figure taken from [63]

ence relation of expressions. Most of the anaphoric expressions are: pronouns (relative, personal, possessive, reflexive) and definite nouns (e.g. this protein).

Example (4)[2]*:*

```
T27 Exp 179 222  the NF-kappa B TF compl 215 222 compl
T28 Exp 264 297  NF-kappa B p65 and NF-kappa B p50
T29 Exp 307 312  which
T30 Exp 459 471  this complex   464 471 complex
T31 Exp 868 882  NF-kappa B p65
T32 Exp 1022 1047this TF 1027 1047 trans.  factor
R1    Coref Anaphora:T29 Antecedent:T28    [T5, T4]
R2    Coref Anaphora:T30 Antecedent:T27
R3    Coref Anaphora:T32 Antecedent:T31    [T10]
```

In Example (4), the first line indicates that in the offset 179-222 of the abstract there is a protein mention "the NF-kappa B transcription factor complex". Its minimal , which still carries its meaning, is "complex" (215-222), and this annotation has the index "T27". In this way, every annotation (a protein mention or an expression that refers to a protein) is described in the coreference relation file.

The last line describes the coreference relation between mentions T32 and T31. Also, we can see that mention T32 is an anaphora and refers to T31, which in its string contains a protein name T10 (p65). So, here we see that we have two types of coreference relations. The first type of relation is between two corefering mentions, like T32 with T31, and the second type of relation is between a mention and a protein name (T32 with T10).

1.2.2. Task description and evaluation

For both types of coreference relations the BioNLP community (mentioned in the previous section) defined a sub-task. First relation type is named surface coreference resolution and the latter is named protein coreference resolution.

For both tasks, the resolver receives a text and a protein annotation file, and as output should return all protein mentions with their offsets and coreference relations between mentions. The result should be a file similar to Example (3), with all detected protein mentions and relations.

1.2.3. Evaluation Measures

In this section, I will define the Protein Coreference Resolution task(in abstracts and full text) mathematically and its evaluation measures. The task(f)

[2]Transcription factor is shorten in TF and complex in compl because of limited line space

of protein coreference resolution, which is defined by the BioNLP community, I define formally in the following way:" given a text (abstract) T with annotated protein names P, the system should return ordered couples (x, y) where y is the antecedent of x".

$$f(X, P) = \{(x_i, y_i)|i = 1..n \ \ \& \ \ Antecedent(y_i) = x_i \ \ \& \ \ y_i = protein\ name\} \\ X - string$$

P is a set of ordered set defined in the following way:

$$P = \{(start_j, end_j, protein_j)|j = 1..n \ \ \& \ \ start_j < end_j \ \ \& \ \ |protein_j| = \\ end_j - start_j\}$$

The system that I built produce three types of anaphora-antecedent links(pairs):

- correct response links
- missing links
- spurious response links.

From these three types of links we can derive:

- $true\ positive\ (tp)$ - number of correct anaphora-antecedent links - links that the system correctly predicted
- $true\ negative\ (tn)$ - number of missing anaphora-antecedent links - links that the system did not find
- $false\ positive\ (fp)$ - number of spurious anaphora-antecedent links - links that the system predicted wrong

To measure the accuracy of the system we use the traditional precision, recall and F-score measurement which are defined in the following way
Precision:

$$P = \frac{tp}{tp+fp}$$

Recall:

$$R = \frac{tp}{tp+tn}$$

F-score is taken as equally weighted harmonic mean of precision and recall:

$$F(\beta) = (1 + \beta^2) \cdot \frac{P \cdot R}{\beta^2 P + R}$$

and for $\beta = 1$

$$F(1) = 2 \cdot \frac{P \cdot R}{P + R}$$

1.2.4. Evaluation measures in BioNLP Protein coreference resolution

In the surface coreference mode, the resolvers should find anaphoric protein expressions and their antecedents, regardless of whether antecedent actually embed protein names or not. The protein coreference mode evaluates the performance of finding anaphoric protein references, which their antecedent contains a protein name. My main focus in this thesis was to build a system which achieves high accuracy in protein coreference mode. Thus, I will describe the scoring principle of protein coreference mode using an example.

From example (4) we take relation *"R1 Coref Anaphora:T29 Antecedent:T28 [T5, T4]"*. Antecedent T28 contains two protein names *p50* and *p65*. So, in the protein coreference mode, we check whether (T29, T5) and (T29,T4) are predicted or not. For each of these pairs, if the system predicted, it gets 1 point. The whole R1 has two points. If the system finds just (T29,T5), it gets 1 point. Wrong predicted pairs decrease the system's precision.

1.2.5. Importance of coreference resolution

Why is coreference task important in biomedical domain?

Coreference resolution is believed to be useful for improved event and relationship extraction [2]. Until now we do not have concrete results to prove that the coreference resolution improves the Information extraction in biomedical domain, because the protein coreference task did not have enough good results and the motivation of this thesis was to build a high accurate coreference resolution system.

In other domains, anaphora resolution improves results in different information extraction and machine translation tasks (see section 2), but still the improvements are not as expected because of the accuracy of current anaphora resolution systems[9,10,11,12].

Anaphora resolution is not a high-accuracy task for information extraction (IE) systems but properly resolving some kinds of coreference is usually difficult even for humans annotators, who achieved about 80 percent [8].

2. Natural Language Processing

2.1. Natural Language Processing: Concepts

Natural language processing (NLP) is a research area for the analysis, understanding, and generation of written and spoken natural language [13,14]. Coreference resolution is a task of NLP.

I had to use different NLP tools to analyse the text. These tools helped the coreference resolution system to understand the text and to find coreference links in it. The quality and robustness of these tools directly influence the performance of my system. To build a high performance coreference resolution system for biomedical domain, I used NLP tools that are built and trained in biomedical data sets.

In the following sections I will explain some NLP tasks that are used to build the used tools and the coreference resolution system. Additionally, I will explain terms that I will use in the next chapters. In the beginning, I will describe the tasks, which help coreference resolution to increase its accuracy. In the latter sections, I will describe the NLP tasks that coreference resolution (is believed) to help.

2.1.1. NLP tasks that help coreference resolution

Sentence segmentation (sentence splitting) is the process of dividing a text into sentences. This process usually receives a text (paragraph or article) as input and returns the text divided into sentences as output. In the English language, this process is rather simple because typically sentences end with a full stop. Nevertheless, the full stop can be ambiguous when we use it in numbers as decimal separator (9.85), in websites (www.in.tum.de), and in abbreviations (M.Sc.). In the biomedical domain, these ambiguous phenomenon occurrences are evident. Despite these occurrences, the sentence splitter GeniaSS [55] is an accurate sentence splitter system, which is trained in biomedical data. GeniaSS achieves an F-score of 99.7 on 200 unseen GENIA abstracts [15].

Word segmentation is the process of dividing a sentence into words. In languages that have a specified word delimiter the word segmentation task is trivial. The English language uses space as word delimiter. There are cases where between two words does not exist a separator because they are connected by an apostrophe. This construction of two words (without separa-

tor) is handled accurately by word segmentation systems. In the English language, the accuracy of the best systems is 99.4-100%. Languages that do not use a word separator word segmentation is not an easy task. Despite this, languages that do not have word delimiters (e.g., Chinese and Japanese) the accuracy of word separators is in the range of 92-98%.

Tokenization is the process where a text is divided into small individual units. Each of these units is meaningful and it is called token. Most of the tokens coincide with words of the text in which the tokenization is being performed [8]. In tokenization we define a scheme how to divide the text.

Part of speech (POS) tagging is the process of assigning the part of speech class to a token/word. The English language has eight part of speech classes, namely: nouns, verbs, adjectives, adverbs, pronouns, prepositions, interjections and conjunctions.
Each language has its part of speech classes. For example in the Albanian language there are 11 part of speech tags.

Parsing(deep parsing) is the process of representing the sentence's syntactic structure in a tree, where the leaves of the tree are tokens. This process provides a detailed structure of the sentence (the tree). Deep parsing is one of the most important tasks in text and speech processing, because it gives the detailed structure of the sentence, relationship between words/tokens and their role in the sentence.

Shallow parsing (or light parsing or chunking) is a similar task to deep parsing, but it does not give a detailed structure of the sentence. This process just identifies the tokens and assigns part of speech class to each chunk.

Parsing direction refers to the direction in which a sentence is parsed. Parsing can be performed in two directions: from left to right (forward direction), and from right to left (backward direction). The direction of parsing a sentence has an important role in the precision and the speed of a parser. Note that this is also important considering that different languages are written in different directions.

Lemmatization is the process of removing the prefixes and/or suffixes of a word/token. The lemmatization process transforms a word to its base (pure) form, which is called lemma. This transformation of the word does not change its meaning. The lemmatization process uses dictionaries to find the lemma of the word and deals with irregular forms of words. For example the words transform, transformed, transforming, transformation are all formed from the lemma *transform*. A similar process to lemmatization is steaming, which finds

the root of a word. Root of the word is the part, which carries the semantic content of it. The stemming process can not handle the irregular forms of the words and do not use dictionaries.

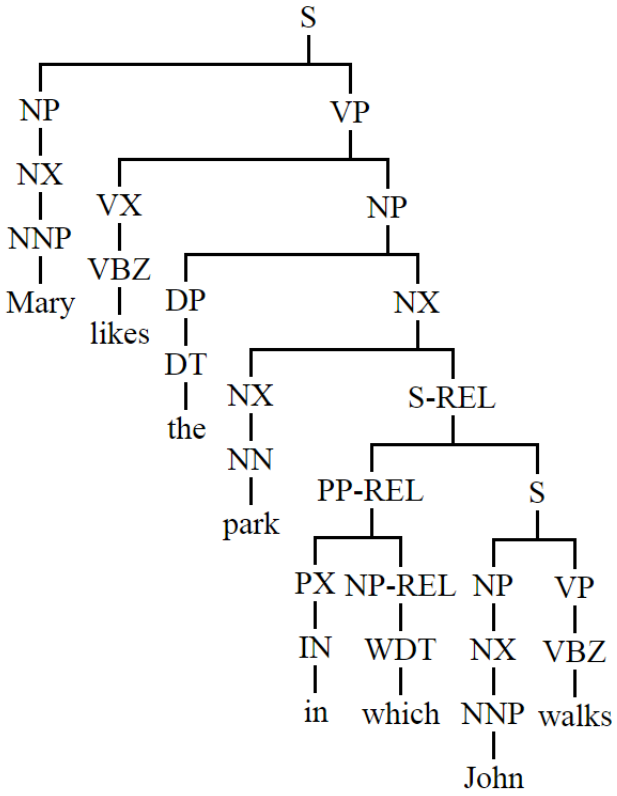


Figure 2.1.: Tree structure of a sentence

Spelling correction is the process in which misspelled words are corrected. Spelling correction systems are called spell checkers. Modern spell checkers (like Google's) not only correct individual misspelled words of a sentence but also have a holistic approach and try to correct the words such that the sentence as a whole becomes semantically correct. To achieve better results, spelling correctors use dictionaries and *n-grams*.

Word-sense disambiguation is the process of determining the meaning of a word. In the real world, it can happen that a word can have two or more meanings. For example, the word desert has two different meanings,a territory with sand and a dish. These "two" words, which have the same spelling, but different meanings are called homonyms.

Another linguistic phenomenon, polysemy, is when two or more words have the same spelling, and distinct and related meanings. The word "mouth"

can be considered as polysemes because it can mean orifice on a face or opening of a cave or river.

Another case of word sense disambiguation are capitonyms. This is when a word has one meaning when it is capitalized, and different meaning when is not capitalized. For example the words ”polish” and ”Polish”, which mean to make shiny and a person from Poland, respectively.

Abbreviation expansionis the process of resolving the full expansion of an abbreviation. A system that deals with abbreviation expansions saves every pair of abbreviation and its full expansion; and then in every new occurrence of an abbreviation it refers back to the saved abbreviation and its expansion. Abbreviations in biomedical text appear often.

2.1.2. NLP tasks which coreference resolution help

Name entity recognition (NER) is the process in information retrieval that identifies and classifies mentions in a text that corresponds to a target group (entities) in which we are interested in. For example, in the biomedical domain we are interested in identifying proteins, genes, mutation and viruses.

Relationship extraction is the process of identifying relations (in a text) between predefined entities. In biomedical information extraction community, the motivation of building an efficient coreference resolution system is that the community believe that coreference resolution improves accuracy of relationship extraction of biomedical entities.

Machine translation is the process of translating a text or a speech from one language to another. These software products that translate a text from one language to another use different statistics and syntactic rules to translate one text to another. Modern machine translation systems use probabilistic models and rules to produce better translation.

Question answering and natural language interfaces are tasks that allow humans to communicate with a computer bot. This communication is done in question and answer format. In this situation, humans ask a computer bot and the computer bot answers the humans by voice or text. In conversion (text or voice), people use pronouns and anaphoric expressions and the computer bot should resolve these anaphoric expressions. For this purpose, coreference and anaphora resolution are important components of question answering and natural language interfaces.

Document summarization is the process of identifying the essence of the

text. This process takes a large text, as input, and return a grammatically and semantically correct smaller text, which contains the essence and the main content of the inputted text.

2.2. Syntax

Syntax(etymology: a Greek word syntaxis=syn [together] + taxis [arrangement]) is a discipline of linguistic, which studies the rules and patterns of ordering and connecting words in the sentence, such that the created structure of the sentence is grammatically correct.

2.2.1. Words

Words and grammatical rules are the fundamental elements to creating a grammatically correct sentence. We should be aware, that if a sentence is grammatically correct, it does not mean that it has a meaning.

These smallest units, words, are classified in different classes based on a characteristics and their function in the sentence. The main characteristic of every word is its part of speech class. Part of speech classes are divided in two subclasses: opened and closed. In the open class are: nouns (N), adjectives (ADJ), adverbs (ADV) and verbs (V). Open part of speech classes have a meaning and every open part of speech class can be extended, thus in these classes can enter new words. In the closed set are: determiners (DET), prepositions (PREP), conjunctions (CONJ) and pronouns (PR). These part of speech classes are closed and their set of words is a fixed set.

2.2.2. Phrases

Phrase is a sequence of words, that inherits the characteristics of one of the words, which is called the head word. The phrase has a part of speech class and it inherits its part of speech class from its head word. From this characteristic, logic follows that a phrase can contain phrases and it has a recursive structure.

Other words of the phrase modify, distinguish the head object from other similar objects and give more information about the head object. These words are called modifiers.

Based on the head word, we have different type of phrases; for example noun phrases (NP) and verb phrases (VP).

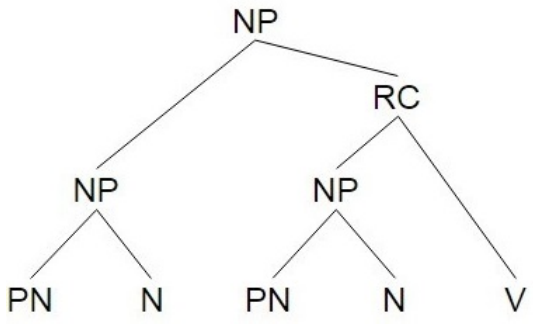


Figure 2.2.: Recursive structure of a phrase

Phrases that their head word is a noun are called noun phrases. Usually singular noun phrases start with a singular determiner (a, an, the, this, any) or start with a proper or a common noun. Plural noun phrases start with a plural determiner (these, those, some..) or start with a plural noun. Pronouns (it, he, they, itself...) are also considered noun phrases.

Example (5):

Some rules for creating noun phrases:

- $NP \rightarrow N$ *student*
- $NP \rightarrow DET\ N$ *the student*
- $NP \rightarrow DET\ ADJ\ N$ *the intelligent student*
- $NP \rightarrow ADJ\ N$ *intelligent student*
- $NP \rightarrow NP\ CONJ\ NP$ *a student and a professor*
- $NP \rightarrow N\ PP$ *student at library*

Phrases that their head word is a verb are called verb phrases. Verb phrases consist of a verb and are followed by a noun phrase and/or prepositional phrase.

Example (6):

Some rules for creating verb phrases:

- $VP \rightarrow V$ *learn*
- $VP \rightarrow V\ NP$ *learn history*
- $VP \rightarrow V\ PP\ NP$ *learn at home*

Prepositional phrases, another type of phrases in English language, consist of a preposition and a noun phrase. Prepositional phrases are used to define more specifically a noun phrase. Thus, we can say that their role is similar to adverbs and adjectives.

2.2.3. Clauses

Clauses are linguistic units that contain a noun phrase and a verb phrase. The clauses are divided into two types: dependent and independent.

2.2.4. Independent clause

An independent clause is a grammatically correct group of words that contains a noun phrase that acts as the subject of clause and a verb phrase.

The independent clause has a meaning, expresses a thought and can form alone a sentence. If a sentence has more than one clause then at least one of the clauses is independent. One of the independent clauses is main and it carries the subject and meaning of the sentence, and is called main clause.

If the sentence has two or more independent clauses they are called coordinated clauses and they are connected with a coordinating conjunction. The coordinated conjunctions are: and, but, for, nor, or, so, and yet.

Alternatively instead of a conjunction, we use a comma to connect two coordinated clauses.

2.2.5. Dependent clauses

A dependent clause is a grammatically correct group of words that contains a noun phrase (NP) and a verb phrase (VP), and it has not semantic meaning. Dependent clauses are also called subordinate clauses, because their role in the sentence is secondary. This means that if a sentence has a dependent clause, then there exist another clause that is main and independent, and the subordinate clause is dependent on the main clause. The role of subordinate clauses in the sentence is similar to adverb or adjective; they are used to describe a property, cause or a consequence of an object or a noun phrase of the main clause. Usually, the subordinate clauses are connected with the main clause or another subordinate clause with an subordinate conjunction or relative conjunction. Examples of relative conjunctions include: who, whom, that, which, what and whose; and examples of subordinate conjunctions include: although, because, if, unless, as, until, even though, even, before and when.

2.2.6. Sentences

We use sentences to express our needs, emotions, ideas, thoughts and curiosity among others. Depending what we express, we use 4 types of sentences:

- declarative sentences, to make statements (.)
- interrogative sentences, to ask questions (?)
- imperative sentences, to give commands (.or !)
- exclamatory sentences, to express strong feelings (!)

In the biomedical articles sentences are in declarative form. Thus, I will describe the structure of the declarative sentences. In the text below, I will use the terms declarative sentence and sentence interchangeably.

A declarative sentence has at least one independent clause, and express a thought. It contains a subject (what or whom the sentence is about) and a verb (the action that the subject takes). There are cases that a sentence expresses a thought and do not contain either verb or subject. This type of sentence is called ellipsis. Ellipsis are common in dialogue. In ellipses, the speaker (writer) assumes a prior knowledge and common sense knowledge from the co-speaker (co-writer).

English language has a defined structure of the sentence and it is an ordered Subject-Verb-Object structure.This structure helps to know the position of the main parts of the sentence.

Another classification of sentences is based on the number and type of clauses that compose the sentences. These types are: simple, compound, complex and compound-complex sentences.

Simple sentence consist just from one independent clause.

Example (7):

Proteins are assembled from amino acids.

Compound sentences consists of two or more coordinated clauses connected with coordinated conjunction.

Example (8):

Bioinformatics is a science and bioinformaticians research bioinformatics.

Complex sentences consists of one independent clause and one or more subordinate clause.

Example (9):

> *Natural language processing is important because it helps computers to understand human language.*

Compound-complex sentences consists of three or more clauses, which two or more are coordinated clauses and one or more subordinate clauses.

Example (10):

> *Bioinformatics is a science and bioinformaticians research bioinformatics, because they want to develop mathematical models and software tools to get knowledge from biological data.*

As we can see, all these types of sentences have a specific construction. We can see that the subjects of clauses in a sentence are correlated and they tend to have same lemma, are synonyms or coreferential expressions. This construction of the sentences and this correlation between subjects of clauses helps to resolve coreference resolution.

Part II.

Coreference Resolution system

3. Coreference resolution

3.1. History

First algorithms for coreference and anaphora resolution date in the ′70 of the last century. The first algorithm was published by Hobbs[46,47] in 1976.This algorithm, a pronominal anaphora resolution algorithm, resolved pronouns ″he″, ″she″, ″it″. Basically, this algorithm was based on syntax rules and heuristics and these rules are:

- anaphora and antecedent should corresponds in their gender and number
- antecedent of a pronoun is the first NP that can be reached with fewer number of steps
- the algorithm uses common sense knowledge (candidate selection)

Hobbs tested his algorithm manually in three different gender texts (historical, novels and news). He claimed that the naive approach algorithm, without using common sense knowledge, achieves accuracy 88%, and with common sense knowledge and with ″careful″ selection of antecedents achieves accuracy more than 91%. These accuracies, the algorithm achieved because Hobbs assumed that the parser that build the syntax structure of sentences is a perfect parser.

In practice, a pronominal anaphora resolution algorithm is hard to achieve accuracy of around 90%, because does not exist a perfect parser and does not exist a tool that has a common sense knowledge.

Algorithm 1 Hobbs's algorithm

1: Find the NP node immediately dominating the pronoun.
2: Go up the tree to the first NP or S node encountered. Call this node X, and call the path used to reach it p.
3: Traverse all children of node X to the left of path p in a left-to-right (L-R), breadth first (BF) fashion. Candidate as antecedent any NP node that appear.
4: **if** node X is the highest node in the sentence **then**
5: traverse the previous sentences in the text in order of recency, the most recent first; each tree is traversed in a L-R, BF manner, and candidate as antecedent any NP.
6: **else**
7: (X is not the highest node in the sentence) continue to step 9.
8: From X, go up to the first NP or S node. Call this new node X, and the path p.
9: **if** X is an NP node and if the path p to X did not pass through the NP node that X immediately dominates **then**
10: propose X as the antecedent
11: Traverse children of X to the left of p in a L-R, BF manner. Propose any NP node encountered as the antecedent
12: **if** X is an S node **then**
13: traverse children of X to the right of path p in a L-R, BF manner, but do not go below any NP or S node encountered.
14: Propose any NP node encountered as the antecedent.
15: Go to step 4

The trend of building coreference and anaphora resolution algorithms based on common sense knowledge and rules continued in the years 90s(1990-1995). Another characteristic of this period is that did not exist an metric to measure performance of anaphora and coreference resolution systems. Because of this, scientist defined specific evaluation metrics for their algorithms in anaphora and coreference resolution.

In this period scientists Barbara J. Grosz,Aravind K. Joshi, Scott Weinstein [48] publish a theory in coreference resolution. This theory, the centring theory, go further in analysis of coreference and anaphora. To resolve anaphora in their theoretical model they add "relationships among focus of attention, choice of referring expression and perceived coherence of utterances within a discourse segment" [14]. This is the first theoretical model that uses the content of the text in anaphora and coreference resolution.

In mid 90s scientists started to apply coreference and anaphora resolution in information extraction.

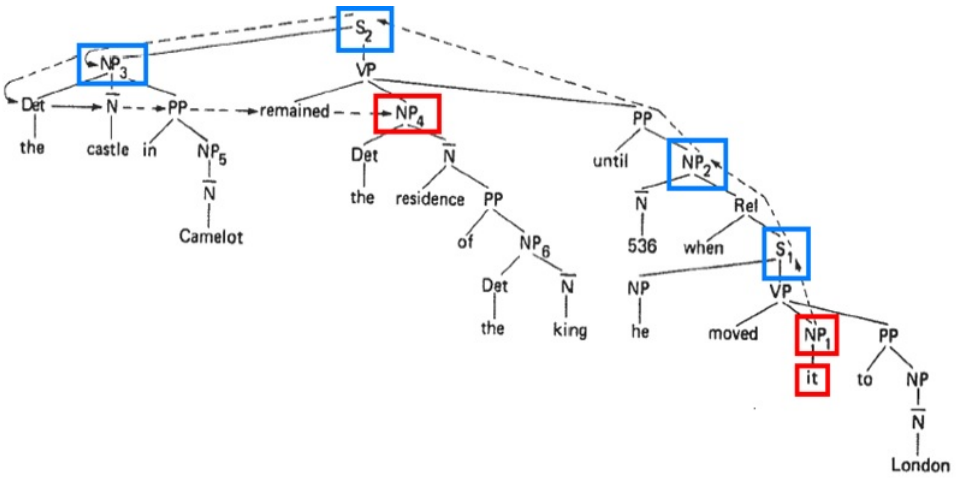


Figure 3.1.: Simulation of Hobbs's Algorithm [30,31]

When scientists started to develop coreference resolution systems they needed a standard of coreference resolution algorithms. To have a standard measurement metrics for evaluation of coreference and anaphora resolution algorithms, the computational linguistic community built and annotated a corpus, and defined evaluation measurements, which are the official standards of coreference and anaphora resolution. The corpus and standards were published in the MUC-6 (Message Understanding Conference) conference. The corpus consisted of 60 English language newswire articles. Two years latter the community published a new corpus MUC-7, which was an update of the previous corpus. These two corpuses are most used and famous corpuses. After publishing these two corpuses, different NLP communities published corpuses and evaluation metrics of coreference resolution. Another known corpus, which is used in the computational linguistic is the Automatic Content Extraction (ACE1) corpus. During this period the community was concentrated in building a base theory of coreference resolution. As a result, they defined standards in annotation schemas and the terminology (e.g. markable detection, mention) that is required in coreference and anaphora resolution.

In the beginning of the year 2000, scientists started using machine learning algorithms in coreference and anaphora resolution. In these methods, they used morphological and structural features: number, gender and part of speech classes of noun phrases to achieve higher accuracy in the resolution. The researcher Soon et al. (2001) to resolve anaphora generated a set of 12 features. In this set of features Ng and Cardie added 53 other features, including positional, morphological, lexical, syntactic, semantic and even pragmatic features. These methods achieved accuracy around 60-67%.

Meanwhile, the computational linguistic community started to get more interested in coreference and anaphora resolution. As a result, new corpuses in different domains and languages were published. Now, there are annotated and published corpuses in domains: scientific articles, dialogues, news and clinical domain; and in languages:Arabic, Chinese, Dutch, English, French,Italian, Japanese,Spanish and Tibetan language.

The community realised that a coreference resolution algorithm perform well in one corpus, but achieves poor results in another corpus. This property of algorithms leads to new perception of coreference resolution. Scientists started to develop corpus based coreference and anaphora resolution algorithms. In the last ten years, scientists are using domain specific features and syntactic rules. Depending on the corpus, the results of the best methods achieve accuracy around 70% in anaphora resolution. [43,65]

3.2. Coreference resolution in biomedical domain

Coreference resolution is believed that helps relationship extraction. This inspired, in 2008, the BioNLP community to build a corpus, with research paper's abstracts, to support other information extraction in biomedical domain. Until now a public corpus with full text articles with protein coreference annotation does not exist.

3.3. The state of the art in protein coreference resolution

In 2011 was organized a competition in Protein Coreference Resolution. The purpose of the task was to evaluate the effect of coreference resolution in biomedical relationship extraction. In the competition the results in protein coreference resolution were not good enough, the best performing algorithm in protein coreference resolution had an accuracy of 34.1%. These results were too low to improve relationship extraction tasks in biomedical domain.[1] Organizers of the competition also provided detailed results of best performing algorithms, which are shown in Table 3.1.

	Precision	Recall	F-score
1st place	73.26	22.18	34.05
2nd place	54.45	21.48	30.96
3rd place	63.22	19.37	29.65

Table 3.1.: Results of BioNLP Protein coreference resolution competition [1]

After the competition some research groups were interested to improve the results of protein coreference resolution. One year later, in 2012, were published two papers in protein coreference resolution. In one of the papers authors use semantic classification and rules to resolve protein coreference and achieved an accuracy of 51.3%. In the second paper authors use an hybrid approach of rules and SVMs and it achieves an accuracy of 60.9%. This algorithm is the current state of the art in protein coreference resolution.

In the table 3.2, I present results of the best performing algorithms in protein coreference resolution.

	Precision	Recall	F-score
Hybrid approach	55.6	67.2	60.9
Event miner	50.4	62.7	55.9
Semantic classification	52.2	50.2	51.3
Reconcile	73.3	22.2	34.1

Table 3.2.: Results of the best performing algorithms in Protein coreference resolution

A full text article corpus with protein coreference annotation does not exist. To evaluate the results of my algorithm in the protein coreference resolution in full text articles I have annotated 8 full text articles, and results of my algorithm in this corpus will be the first results in the protein coreference resolution in full text.

3.4. Available tools for biomedical information extraction

To complete the coreference task, I needed supporting NLP tools for preprocessing and structuring the data set.

For coreference resolution, I used tools that are trained in biomedical domain, because they have better performance than other tools. In the following paragraphs I will describe the tools that i used in my system.

I used the deep parser "ENJU parser", which is trained in biomedical data [50]. Empirically, it is shown that this parser achieves better results in biomedical domain than other parsers [63]. This parser is a Head-driven phrase structure grammar (HPSG)parser. HSPG parsers use semantic and syntactic rules, as well as dictionary to build the tree structure of the sentence[51-57]. The average time of parsing a sentence is 500ms. This parser achieves precision 87.85% and recall 86.85% on the Penn Treebank and the average parsing time is 360 ms per sentence [66].

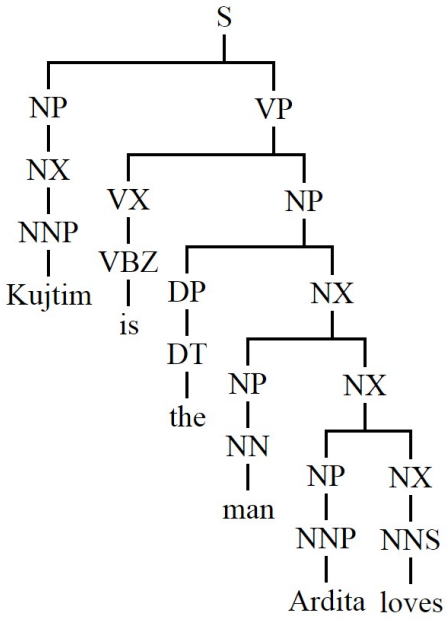


Figure 3.2.: Example output of Enju parser [59]

An advantage of HPSG parsers is that they give rich linguistic informations. The Enju parser returns the syntactic structure of the sentence in tree. This structure is returned as XML object. In each internal node (not leaf) of the XML object are included these informations: syntactic category (verb, noun phrase preposition,...), extra syntax category (coordinate clause, subordinate clause), and a unique identifier for the node.

For each token, leaf node of the tree, the parser provides information about the following linguistic characteristics: unique identifier of the token, syntactic category, Penn Treebank-style part of speech tag, base form, and for each of part of speech gives its characteristics (words, punctuations). For example for verb, gives following information: the person, the number, the tense, aspect

and voice(passive/active).

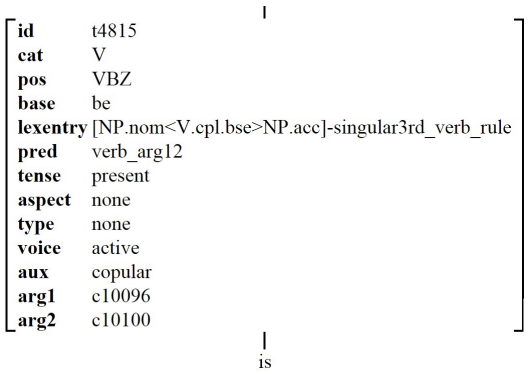


Figure 3.3.: Information that *Enju parser* gives for the token *is*[59]

Another tool that I used in this thesis is the sentence splitter, GENIA sentence splitter (GeniaSS) [55]. This sentence splitter receives a text (a text file) as input and returns array of sentences (a file where each sentence is written in a specific line). This sentence splitter is trained on supervised learning method using maximum entropy modelling; and it is implemented in C++ [61].

GeniaSS is high accurate parser sentence splitter in the Biomedical domain, it achieves an F-score of 99.7 on 200 unseen GENIA corpus [15].

4. Coreference resolution using rules and domain-dependent features

4.1. Overview

In this chapter, I will describe the implementation of a new method in protein coreference resolution. I used corpus based rules to deal with problem of pronominal coreference resolution.

The corpus consists of research papers' abstracts, thus, I can assume that the text is written in formal way and authors followed syntax rules. From this hypothesis, I got an idea to generate syntax rules that model relationship between clauses' subjects in sentence.

For definite noun coreference resolution, I use domain dependent features, rules and semantic classification. Also, I realised that the order of the rules influences the accuracy of the system, thus, I defined an order of rules, which I applied on definite noun coreference resolution. This new method effects the result by improving both the recall and precision of the system.

To build the system, I used the paradigm "divide and conquer". Each type of anaphoric expression is resolved in different way based on its syntax role in the sentence structure. I divided the anaphoric expressions in two main types: pronominal anaphoric expressions and definite noun (phrase) anaphoric expressions.

Because pronouns can have different role in the sentence, I divided pronouns into three classes:

- relative pronouns: that, which, whom and whose
- personal pronouns: it and they
- possessive pronouns:its and their

To extract these rules, I used the training data set of the BioNLP corpus. First, I generated different statistics of the corpus and calculated the distribution of anaphoric expressions, antecedents and coreference links.

I used the statistics to get knowledge about the corpus and they led me to research following characteristics of the corpus:

- the structure of the training data set via mining
- patterns
- structure of the language used in the articles
- deeper statistics on anaphors, antecedents and their relationship
- clauses relationship in the sentence
- anaphora's role in the sentence
- syntactic rules
- window size of the search space

4.2. Data set

As I have described in the first chapter, I have used the training set of BioNLP corpus to derive the rules. I have tested the accuracy of the algorithm in the test set and in the corpus of 8 full text documents that I have annotated in a format similar to the BioNLP corpus. I have annotated this corpus of 8 documents on the www.tagtog.net [67,68], which is a web application tool for annotation. I have compared my algorithm with current state of the art algorithms in the development and the test set of BioNLP corpus.

4.3. Coreference resolution system architecture

Coreference resolution systems have a pipeline architecture. My protein coreference resolution system I developed in Java and published as an open source tool.[1]

[1] https://github.com/kujta1/CoreferenceResolution

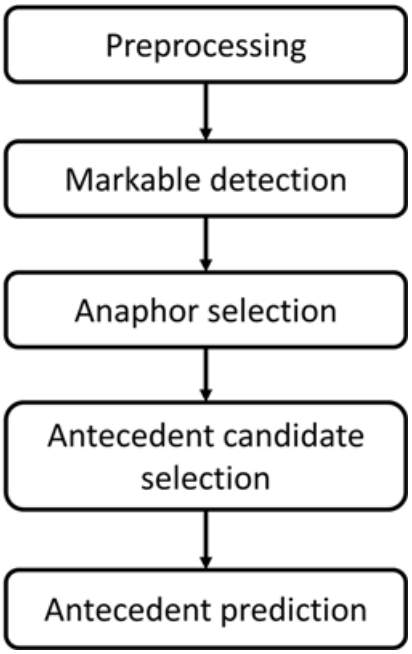


Figure 4.1.: Architecture of the system. As input the system receives two files, the protein annotation file and the text file. The protein annotation file should follow the BioNLP annotation schema. The system after the fifth step, antecedent prediction, will create a file in which the system will write the coreference links of the inputted text file [3]

The first step is the preprocessing step where the text is dived into sentences. Then, I select all nun phrases the are potential antecedents of an anaphoric expression. In the next step, I select potential antecedent candidates. When all anaphoric expressions are selected then for each anaphoric expression I select its potential antecedents. In the final step, I select the most likely antecedent candidate for each anaphoric expression.

In the following sections, I will describe the 5 components of my coreference resolution system.

Input (PMID - 7964516)	T cell hybridomas respond to activation signals by undergoing apoptotic cell death, and this is likely to represent comparable events related to tolerance induction in immature and mature T cells in vivo. Previous studies using antisense oligonucleotides implicated the c-Myc protein in the phenomenon of activation-induced apoptosis. This role for c-Myc in apoptosis is now confirmed in studies using a dominant negative form of its heterodimeric binding partner, Max, which we show here inhibits activation-induced apoptosis.
Preprocessing of sentence S3	S1: T cell hybridomas respond to activation signals by undergoing apoptotic cell death, and this is likely to represent comparable events related to tolerance induction in immature and mature T cells in vivo. S2: Previous studies using antisense oligonucleotides implicated the c-Myc protein in the phenomenon of activation-induced apoptosis. S3: This role] for c-Myc in apoptosis is now confirmed in studies using a dominant negative form of its heterodimeric binding partner, Max, which we show here inhibits activation-induced apoptosis.
Markable detection	S1: *[T cell hybridomas]* respond to *[activation signals]* by undergoing *[apoptotic cell death]*, and this is likely to represent *[comparable events related to tolerance induction]* in *[immature and mature T cells [in vivo]]*. S2: [Previous studies using *[antisense oligonucleotides]*] implicated *[the c-Myc protein]* in *[the phenomenon of [activation-induced apoptosis]]*. S3: *[This role for [c-Myc] in [apoptosis]]* is now confirmed in *[studies]* using a *[dominant negative form of [[its] heterodimeric binding partner,[Max]]]*, *[which] [we]* show here inhibits *[activation-induced apoptosis]*.
Anaphora selection	S1: T cell hybridomas respond to activation signals by undergoing apoptotic cell death, and **[this]** is likely to represent comparable events related to tolerance induction in immature and mature T cells in vivo. S2: Previous studies using antisense oligonucleotides implicated the c-Myc protein in the phenomenon of activation-induced apoptosis. S3: This role for c-Myc in apoptosis is now confirmed in studies using a dominant negative form of **[its]** heterodimeric binding partner, Max,[this] we show here inhibits activation-induced apoptosis.
Antecedent candidate selection and antecedent	S1: [T cell hybridomas] respond to [activation signals] by undergoing [apoptotic cell death], and this is likely to represent [comparable events related to tolerance induction] in [immature and mature T cells [in vivo]]. S2: [Previous studies using [antisense oligonucleotides]] implicated [the c-Myc protein] in [the phenomenon of [activation-induced apoptosis]]. S3: [This role for [c-Myc] in [apoptosis]] is now confirmed in [studies] using [a dominant negative form of ***[its]*** heterodimeric binding partner, Max], which we show here inhibits activation-induced apoptosis.
Predict antecedent	S1: T cell hybridomas respond to activation signals by undergoing apoptotic cell death, and this is likely to represent comparable events related to tolerance induction in immature and mature T cells in vivo. S2: Previous studies using antisense oligonucleotides implicated the c-Myc protein in the phenomenon of activation-induced apoptosis. S3: This role for ***[c-Myc]*** in apoptosis is now confirmed in studies using a dominant negative form of ***[its]*** heterodimeric binding partner, Max, which we show here inhibits activation-induced apoptosis.

Table 4.1.: A simulation of the protein coreference resolution system workflow (the text adopted from [3])

4.4. Preprocessing

In this step, I have used sentence segmentation and syntax parsers to structure the inputted text. For sentence segmentation I used Genia sentence splitter and for sentence (deep) parsing I used Enju parser.[50,54]

4.5. Markable detection

After the preprocessing step I knew the structure of each sentence, I selected text chunks that are candidate coreferential expressions. In the MUC-7 jargon, these text chunks are called markables. Noun phrases and pronouns were considered as markables. To reduce the redundant information in markables, I filtered noun phrases that shared the same head word and I selected as potential candidate just the longest noun phrase (Figure 4.2).

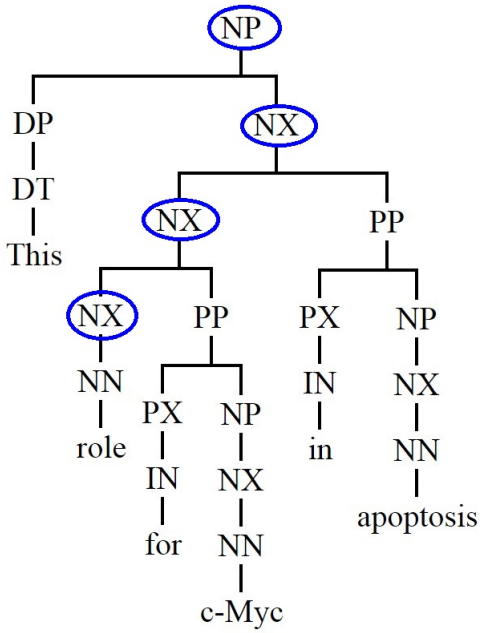


Figure 4.2.: Filtering out noun phrases that share the same head word. In this example, all four noun phrases have the same headword *"role"*. I considered as potential candidate just the longest noun phrases *"This role for c-Myc in apoptosis"*.(In the Enju parser noun phrases are marked with abbreviations NP and NX)

Additionally, I did not consider as markables noun phrases that have a subordinate clause in their content (Figure 10).

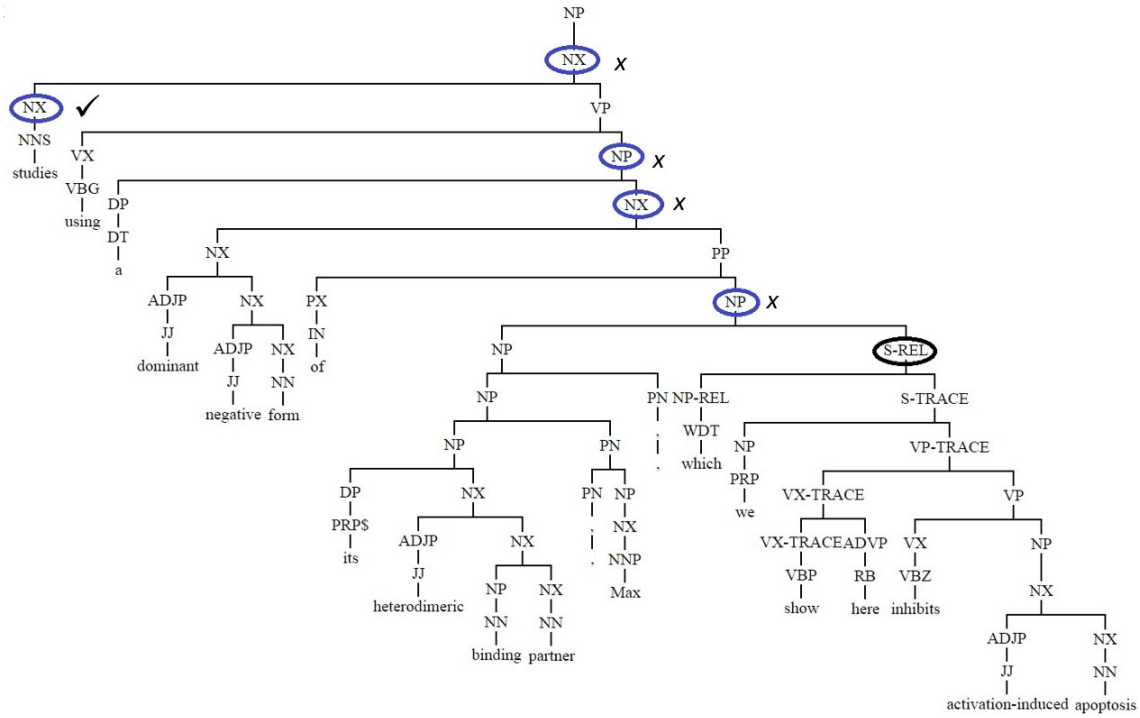


Figure 4.3.: Filtering out noun phrases that have a subordinate clause in their content. In this example, I removed noun phrases (marked with ***x***) that in their content have the subordinate clause *"which we show here inhibits activation-induced apoptosis"*

4.6. Anaphora selection

Anaphora selection is the step in which I selected text parts that are anaphoric expressions (pronouns and definite noun phrases).

I selected the following types of pronouns:

- relative pronouns (who, whom, whose, which and that)
- personal pronouns (it, they and them)
- possessive pronouns (its and their)
- definite noun phrases that refer to protein

From the personal pronouns, I just selected the pronouns *"it"* and *"they"* as anaphoric expressions. Other personal pronouns (*I,you,he,she, we and you*) cannot refer to proteins, thus, I filtered and did not select them as anaphoric expressions. Another case of filtering is when the pronoun *"it"* did not refer to any object, for example *"It is important for people to live healthily"*. When

the pronoun *"it"* does not refer to an object, it is called *pleonastic*[2] *it*. I used patterns to filter pleonastic mentions of the pronoun *it* in similar way like in [3,69].

I used these patterns to identify pleonastic it:

- *It be* $[Adj|Adv|verb]$* *that*
- *It be Adj* $[forNP]$ *to VP*
- *It* $[seems|appears|means|follows]$ *that*
- *It has been shown*

From possessive pronouns, I only selected the pronouns *its* and *their* as anaphoric expressions. Other possessive pronouns *(my, mine, our, ours, his, her, hers, your, yours)* cannot refer to proteins, thus, I filtered and did not select them as anaphoric expressions.

Identifying anaphoric definite nouns, is a difficult task, because we do not know if a noun phrase is anaphoric or not. In the biomedical domain, many definite nouns do not have an antecedent as the referenced concepts refer to a biomedical entity which is not introduced before or to a general concept. These non-anaphoric expressions appear in scientific texts, because authors assume that the reader has some knowledge in the domain and the authors use the definite noun phrases when they refer to a biomedical entity [3,69].

To select the potential anaphoric definite noun phrases, I used semantic knowledge [3,70] and rules which were generated from the statistics of the corpus. I selected as anaphoric expressions only noun phrases that satisfied the following conditions :

- their head noun was protein(s) or gene(s) and start with the words: the, this and these. For example the protein, these two genes and this protein.
- their head noun was factor(s), molecule(s), element(s), family, inhibitor(s), receptor(s), complex, construct(s).

Using these filters I covered more than 85% of the coreference links in the training and the development set.

Indefinite pronouns can act as anaphoric expressions. These cases are not so evident in the biomedical domain. Although these types of expressions are very difficult to identify and my system does not process or try to find these types of anaphoric expressions. In Example 11 the expression "a protein" is an anaphoric expression and its antecedent is "P53".

[2] pleonastic (etymology: from Greek word pleonasm which means much, too much) is the phenomena when we use more words than necessary to express a statement to express an idea

Example (11):

P53, a protein that in humans is encoded by the TP53 gene.

4.6.1. Distribution of anaphoric expressions

	BioNLP training data set	BioNLP development data set	BioNLP test data set	Full text data set
Relative pronouns	32%	34%	26%	30%
Personal pronouns	10%	12%	8%	12%
Possessive pronouns	26%	29%	31%	19%
Definite noun phrases	25%	19%	24%	31%
Other	7%	6%	11%	8%

Table 4.2.: Distribution of anaphoric expressions that refer to protein by syntactic category

From Table 4.2, we can see that in *the full text data set* the anaphoric definite noun phrases occurs more often than in abstracts. On the other hand, possessive pronouns occur more often in abstracts than in full text.

4.7. Antecedent candidate selection

In my system, I used rules to resolve pronominal anaphora, rules and domain dependent features to select the antecedent candidate of a definite noun anaphoric expression. These rules I derived based on the assumption that the subjects of clauses in a complex or a compound sentence refer to the same object, and two coordinated noun phrases refer to the same object. To select antecedent candidates for definite noun phrases, I used syntactic structure of the sentences and domain dependant features. Most scientists use the nearest candidate to resolve definite noun anaphoric expressions and also claim that the nearest candidate approach is better than best candidate. In my system I implemented a hybrid approach of nearest first and best first [2,4].

4.7.1. Statistics of the data sets

The first step in antecedent candidates selection of anaphoric expressions was to find a relationship between antecedents and anaphoras. I measured the distance between an anaphoric expressions and their antecedent. This distance,

I measured in sentences as I wanted to know where to look the antecedent of an anaphoric expression.

	Test set			
	D=0	D=1	D=2	D>2
Relative pronouns	100%	0%	0%	0%
Personal pronouns	80%	20%	0%	0%
Possessive pronouns	83%	11%	5%	1%
Definite noun phrases	31%	46%	8%	15%
Other	47%	29%	15%	9%
Total	72%	19%	4%	5%

Table 4.3.: Distribution of anaphoric expressions that refer to protein by category in the **test set**

	Test set			
	D=0	D=1	D=2	D>2
Relative pronouns	100%	0%	0%	0%
Personal pronouns	84%	16%	0%	0%
Possessive pronouns	90%	10%	0%	0%
Definite noun phrases	42%	34%	0%	8%
Other	72%	14%	14%	14%
Total	82.6%	12.4%	1%	4%

Table 4.4.: Distribution of anaphoric expressions that refer to protein by category in the **development set**

From these statistics, I decided to use ”divide and conquer” techniques to generate the rules and the window size of the search space for each of 4 types (relative pronouns, personal pronouns, possessive pronouns and definite pronouns) of anaphoric expressions. And the search space for each type is following:

- relative pronouns - window size of one sentence
- personal and possessive pronouns - window size of two sentences
- definite nouns window size of three sentences

4.7.2. Syntax structure of data set

From the definition of syntax in section 2.2, we saw that syntax is a set of rules and words. This was a good signal that I could use syntax rules to build a coreference resolution system. An important rule from syntax is that anaphora and antecedent should correspond to their number (singular or plural)[3].

The Subject-Verb-Object (SVO) structure of declarative sentences in the English language is the most important rule in coreference resolution. This structure can be defined as: "when people express a thought (clause), first they say the subject of the clause (what or whom the sentence is about) then the verb, and in the end the object of the clause".

From this property of the English language I derived rules and hypothesis for resolving personal and possessive pronouns.

I will describe in the following sections all rules and heuristics that I used in the system.

4.7.3. Domain dependent heuristics

In the section 3.1 (History) we saw that the accuracy of algorithms is dependent from the corpus. Because of this property of coreference resolution systems, I decided to use corpus features in antecedent and anaphora coreference resolution.

I created regular expressions to identify biomedical entities. The regular expression classify every word as biomedical entity, if it contains a capital letter in its content or it contains a number or a special character.

For each definite noun anaphoric expression I select different candidates based on their number (plural or singular) and based on the head noun. I used the following rules to select potential antecedents of a definite noun anaphoric expression:

- If the anaphoric expression is a ***plural*** definite noun, and its head word is ***genes/proteins*** then as potential candidate consider:
  - Noun phrases that their head word is genes/proteins
  - Noun phrases that their head word is *"family"*
  - If the previous two rules do not find a potential antecedent candidate then as potential antecedent candidates consider noun phrases that contain 2 or more proteins
- If the anaphoric expression is a ***singular*** definite noun, and its head word is ***gene/protein*** then as potential candidate consider:

[3] In other domains is important that the antecedent and anaphora to not have different genders.

  - Noun phrases that their head word is gene/protein
  - If the previous rule does not find a potential antecedent candidate then as potential antecedent candidates consider noun phrases that is a protein name

- If the anaphoric expression is a ***plural*** definite noun, and its head word is ***inhibitors, elements, complexes, molecules or receptors*** then as potential candidate consider:
  - Noun phrases that contain two or more biomedical entities
  - Noun phrases that have same head word with the anaphoric expression
  - If the previous two rules do not find a potential antecedent candidate then as potential antecedent candidates consider noun phrases that contain 2 or more proteins

- If the anaphoric expression is a ***singular*** definite noun, and its head word is ***inhibitor, element, complex, molecule or receptor*** then as potential candidate consider:
  - Noun phrases that have same head word with the anaphoric expression
  - Noun phrases that contain a biomedical entity

4.8. Antecedent prediction

The last step of the coreference resolution process is antecedent selection. For pronoun coreference resolution I have used rules and these rules generate just one candidate. In the definite noun coreference resolution, it is shown in other domains that the nearest potential candidate is the most likely candidate to be the antecedent of the definite noun anaphoric expression. Other methods in selecting the antecedent of an definite noun anaphoric expression were based in choosing the best candidate. The best candidate was selected based in some criteria. All ”best” candidate methods were less accurate than the nearest first.

I created a hybrid method based on nearest candidate first and the ”best rules” first. In this hybrid approach, I just select candidates that fulfill the requirements of one the best rules and if these rules find candidates I select the nearest candidate. If these ”best rules” do not find a candidate then I run other less accurate rules to find the nearest antecedent that fulfill one of these less accurate rules.

4.8.1. Relative pronouns

I realised that relative pronouns and their antecedent are always in the same sentence. Because of this relationship, I measured the distance between the end offset of antecedent and the begin offset of the anaphora. The results were that this distance was less than 5 characters in more than 90% of relative pronoun anaphora links. These numbers tell us that if we have a relative pronoun then its antecedent is the nearest noun phrase.

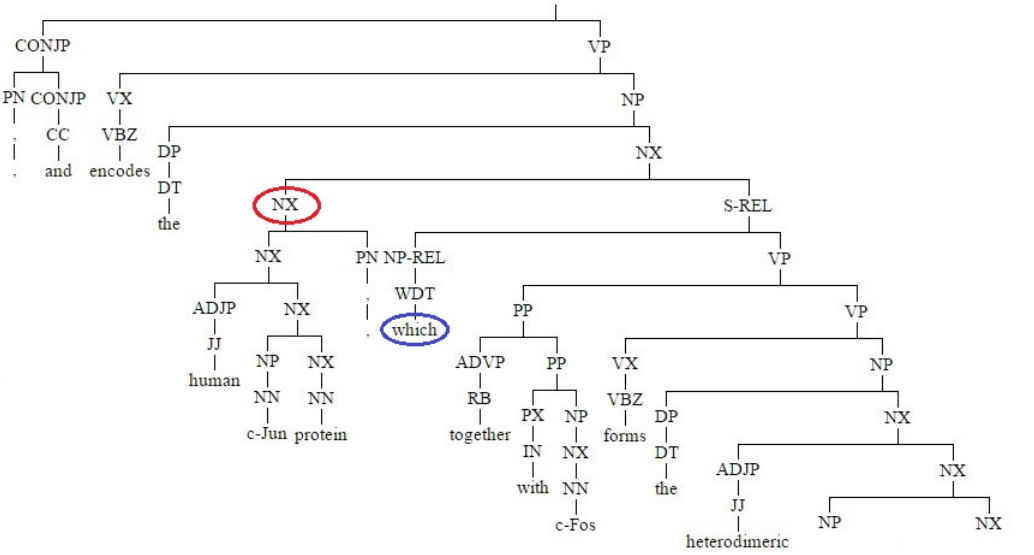


Figure 4.4.: DESCRIPTION

4.8.2. Personal pronouns

From the statistics and mining the data set I realised that when the personal pronoun appears in the second or third clause of the sentence then its antecedent is the subject of the previous (nearest) clause, in the first or in the second clause, respectively. The same rule is when in the sentence we have two coordinated noun phrases.

If the personal pronoun appears in the first clause then the personal pronoun is the subject of the sentence. In this case, I take as antecedent the subject of the previous sentence. I base this rules based on the assumption:*"In scientific writing people try to follow the rules of cohesion and coherence. This mean that in one paragraph they do not introduce so many subjects and describe just one subject and its property."*

[3]Sentence "Thus, although CD3, CD28, and CD2 activate many of the same signaling molecules, they differed in their capacity to induce the tyrosine phosphorylation of HSI." is taken from PMID-9794375

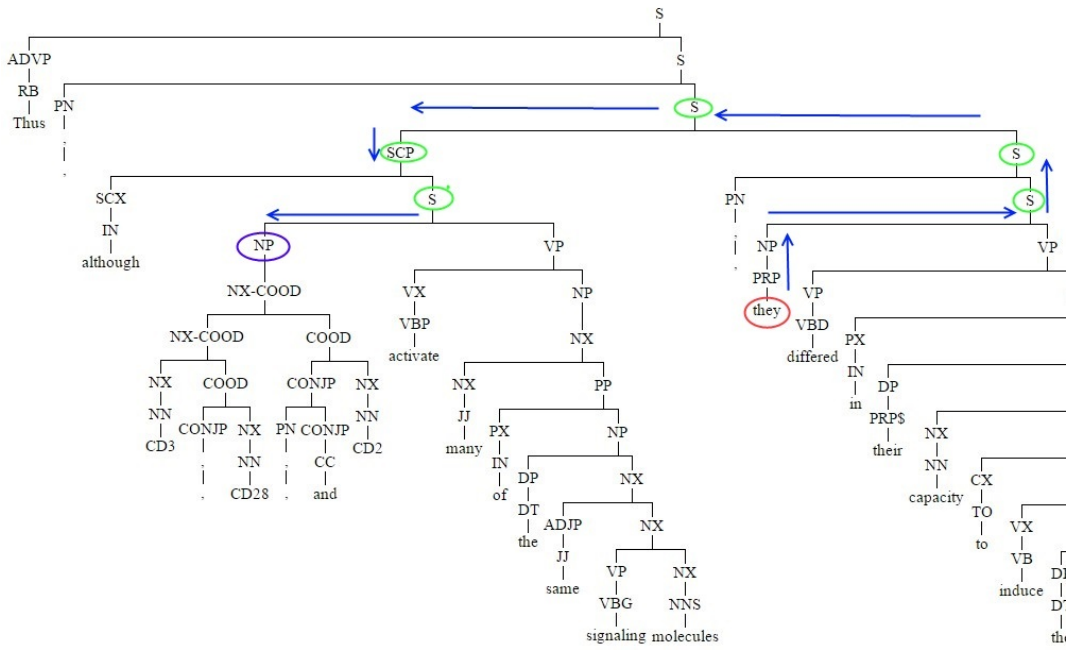


Figure 4.5.: Simulation of the rule when the anaphora is in the second clause and the antecedent is in the first clause of a complex sentence.

An exception is when we have two coordinated verb phrases. If the anapha (it or they) appears in a coordinated verb phrase, and it is in the second verb phrase, then I do not search for antecedent in the first verb phrases because "the verb phrase describe an activity that is taken from the subject". Thus, both verb phrases of coordinated phrases describe an activity, and the activity is taken by the subject of the sentence and I select as antecedent the subject of previous clause (sentence) in which the anaphora appears.

[4]Sentence "Therefore, our data not only reveal that Nmi can potentiate STAT-dependent transcription, but also suggest that it can augment co-activator protein recruitment to at least some members of a group of sequence-specific transcription factors." is taken from PMID-8108414

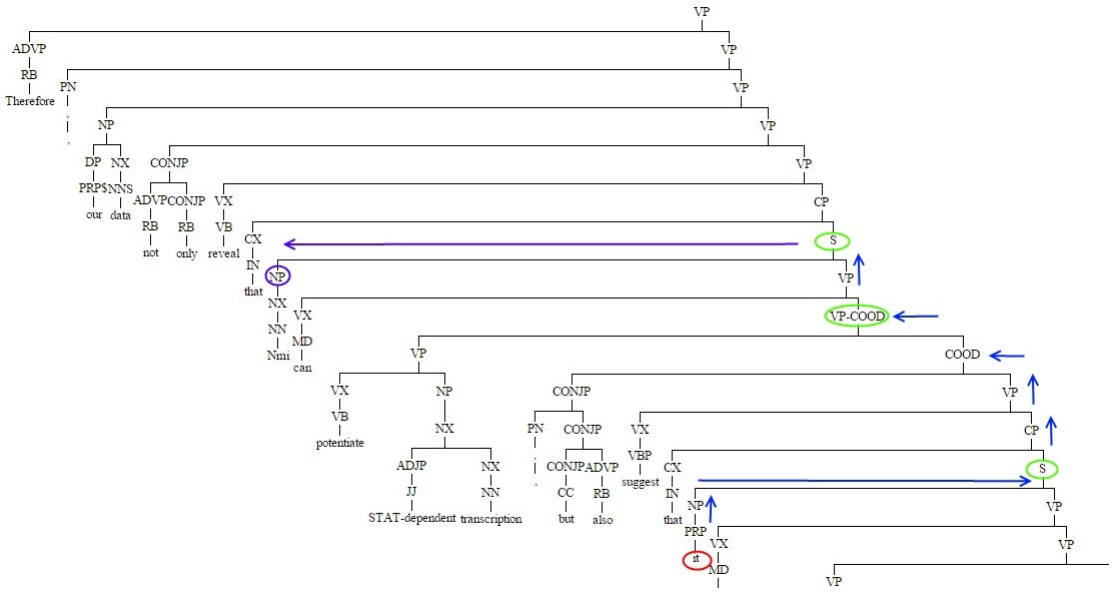


Figure 4.6.: Simulation of the rule when the anaphora is in the second verb phrase of a coordinated verb phrase and I chose as antecedent the subject of the nearest clause (sentence).[4]

4.8.3. Possessive pronouns

The same idea and the same rules I applied in possessive pronouns. From the mining of the corpus,I realised that possessive pronouns appear more in coordinated noun phrases.

[5]Sentence "Therefore, our data not only reveal that Nmi can potentiate STAT-dependent transcription, but also suggest that it can augment coactivator protein recruitment to at least some members of a group of sequence-specific transcription factors." is taken from PMID-8108414

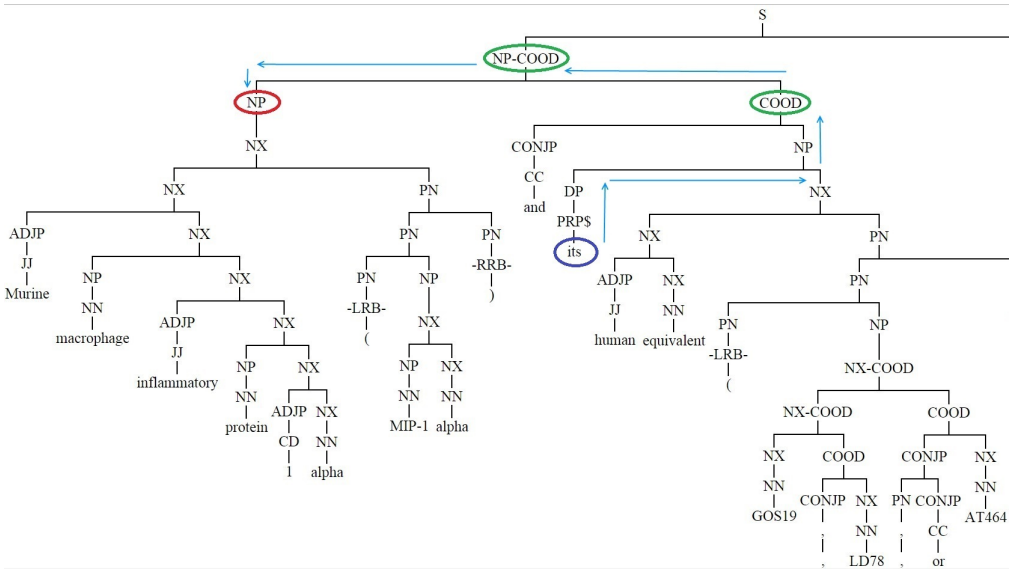


Figure 4.7.: Simulation of the rule when the anaphora is in the second coordinated noun phrase of a coordinated noun phrase and I chose the antecedent from the previous noun phrase[5]

4.8.4. Definite noun phrases

I select the nearest candidate of the candidates that fulfill the requirements of the ”best” rules, then if there is no candidate selected then I run the other rules to select the nearest noun phrase that fulfill the requirements of the less accurate rules.

[6]Sentence ”Murine macrophage inflammatory protein 1 alpha (MIP-1 alpha) and its human equivalent (GOS19, LD78, or AT464) are members of the -C-C family of low-molecular-weight chemokines.” is taken from PMID-7760807

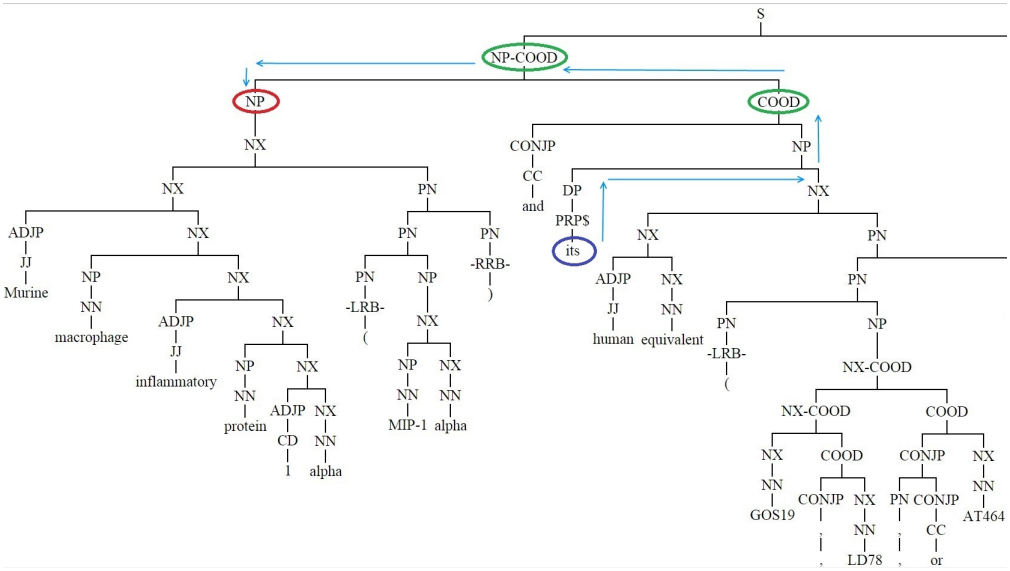


Figure 4.8.: Simulation of the rule when the anaphora is in the second coordinated noun phrase of a coordinated noun phrase and I chose the antecedent from the previous noun phrase[6]

5. Results

5.1. Results

In this chapter I will show the performance results for the protein coreference resolution system.The evaluation is done in the way that is described in the first chapter. To evaluate the system in the development and the test set I used the BioNLP online testing system [71][72]. In the development set there are 202 coreferent links the system should predict and in the test set there are 284 coreferent links that the system should predict.

I will show the result of the whole system and I will compare with the current state of the art method and other best methods.

For the development data set I will compare results with current state of the art methods in each of 4 types (personal,possessive, relative and definite noun phrases).

Additionally, I will present the first result in 243 coreferent link in coreference resolution in full text articles.

Methods	Recall	Precision	F
CR resolver	52.5	50.2	51.3
Current state of the art	55.6	67.2	60.9
My system	**60.92**	**65.53**	**63.14**

Table 5.1.: Results of the current the best two results and the results of the system that I build

Table 5.1 shows the result of the system and the current state of the art algorithms in Protein coreference resolution in the test data set. The evaluation measures precision (P), recall (R), and F-measure (F) are in presented in percentage. The system performance is tested in the online system of the Protein coreference task website. This test is performed in the test data set which consists of 250 documents and in this data set appear 284 protein coreference links. The new methods that I implemented in the system improved for 5.3% the recall and doped the precision for 1.7%. My system outperforms the current state of the art method in the F-measure by 2.24%.

The table 5.2 shows the result of the system in the full text. The system performance is tested in 8 full text documents, in which appear 243 protein

coreference links. The result that I present in this thesis are first results in full text articles. This results are first because the protein coreference resolution is a new research filed and does not exist an annotated corpus with full text articles.

Methods	Recall	Precision	F1
My system	**61.3**	**82.3**	**70.2**

Table 5.2.:

	Current state of the art			My system			
	R	P	F	R	P	F	Diff.
Relative pronouns	83.8	83.8	**83.8**	85.3	78.4	**81.7**	**-2.1**
Per. & poss. pronouns	63.8	77.9	**70.2**	75.6	74.7	**75.1**	**+4.9**
Definite noun phrases	36.8	58.3	**45.2**	45.2	66.7	**55.4**	**+10.2**
Whole system	59.9	77.1	**67.4**	68.3	75	**71.5**	**+4.1**

Table 5.3.: Distribution of anaphoric expressions that refer to protein by syntactic category

List of Tables

List of Figures

6. Bibliography

[1] Ngan Nguyen, Jin-Dong Kim and Jun'ichi Tsujii. Overview of the Protein Coreference task in BioNLP Shared Task 2011.*In Proceedings of BioNLP Shared Task 2011 Workshop*, pages 74-82 Portland, Oregon, USA, 24 June, 2011. 2011 Association for Computational Linguistics

[2] Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2004a Improving noun phrase coreference resolution by matching strings. *In Proceedings of the First International Joint Conference on Natural Language Processing*, pages 22-31.

[3] Ngan Nguyen, Jin-Dong Kim, Makoto Miwa, Takuya Matsuzaki and Jun'ichi Tsujii. Improving protein coreference resolution by simple semantic classification. November 2012 BMC Bioinformatics, 13:304

[4] Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to coreference resolution of noun phrases 2001. *Computational linguistics*, 27(4):521-544

[5] Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008b. BART: A modular toolkit for coreference resolution. *In Proceedings of the ACL-08: HLT Demo Session*, pages 9-12.

[6] Laura Hasler, Constantin Orasan, and Karin Naumann.2006. NPs for events: Experiments in coreference annotation. *In Proceedings of the 5th International Conference on Language Resources and Evaluation,* pages 1167-1172.

[7] Jin-Dong Kim, Sampo Pyysalo,Tomoko Ohta,Robert Bossy, Ngan Nguyen,Jun'ichi Tsujii Overview of BioNLP Shared Task 2011, *In Proceedings of BioNLP Shared Task 2011 Workshop*, pages 1-6 Portland, Oregon, USA, 24 June, 2011. 2011 Association for Computational

[8] Ronen Feldman and James Sanger.The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, Cam-bridge, MA, USA, December 2006.

[9] Fabio Celli and Massimo Poesio. Improving Relation Extraction with Anaphora

In Italian DAARC 2011

[10] M.Novak. Utilization of Anaphora in Machine Translation in *WDS 2011 Proceedings of Contributed Papers: Part I, ser*. WDS 11, 2011, pages. 155 - 160.WDS'11 Proceedings of Contributed Papers, Part I, 155-160, 2011

[11] Christian Hardmeier and Marcello Federico. 2010. Modeling Pronominal Anaphora in Statistical Machine Translation. *In International Workshop on Spoken Language Translation (IWSLT)*, Paris, December 2nd and 3rd, 2010, pages 283 -289

[12] Ruslan Mitkov ,Sung-Kwon Choi and Randall Sharp. ANAPHORA RESOLUTION IN MACHINE TRANSLATION in *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 87-98

[13] Liddy, E.D. Natural Language Processing. *In Encyclopedia of Library and Information Science, 2nd* Ed. NY. Marcel Decker, Inc. 2001

[14] Gobinda G Chowdhury. Natural language processing,2003. *In Annual review of information science and technology* vol 37, pages 51-89

[15] SÃ¦tre, Rune, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Yuichiro Matsubayashi and Tomoko Ohta. AKANE System: Protein-Protein Interaction Pairs in BioCreAtIvE2 Challenge, PPI-IPS subtask. *In Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 209-212, April 2007. CNIO.

[16] Joseph Olive, Caitlin Christianson, and John McCary. 2011.*Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer Publishing Company, Inc., 1st edition.(page 97)

[17] Huihsin Tseng, Pichuan Chang, Galen Andrew, Daniel Jurafsky and Christopher Manning. A conditional random field word segmenter for sighan bakeoff 2005, *In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 168-171, Jeju Island, Korea.

[18] Pi-Chuan Chang, Michel Galley, and Christopher D. Manning. Optimizing Chinese Word Segmentation for Machine Translation Performance 2008 *in Proceedings of the Third Workshop on Statistical Machine Translation*, pages 224-232

[19] Hai Zhao, Chang-Ning Huang and Mu Li An. Improved Chinese Word Segmentation System with Conditional Random Field in *Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing* 2006, pages 162-165, Sydney.

[20] Manabu Sassano. An empirical study of active learning with support vector machines for japanese word segmentation. *In Proceedings of the 40th Annual Meeting of the Association for ComputationalbLinguistics*, pages 505-512, 2002

[21] Josef Steinberger , Massimo Poesio Mijail A. Kabadjov and Karel Jezek. Two uses of anaphora resolution in summarization in *J. Steinberger et al.Information Processing and Management* 43 (2007), pages 1663-1680

[22] Ruslan Mitkov, Richard Evans, Constantin Orasan, Le An Ha, and Viktor Pekar. Anaphora Resolution: To What Extent Does It Help NLP Applications? DAARC 2007, LNAI 4410, pages. 179-190, 2007

[23] Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, pages 1-54, January 2013. [24] Kees van Deemter and Rodger Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics,* 26(4):629-637. Squib

[25] Marco Maggini, Natural Language Processing Part 3: Syntax grammar chunking constituents

[26] Jirka Hana. Intro to Linguistics - Syntax 1, November 7, 2011

[27] Jean Eggenschwiler and Emily Dotson. Cliffs Quick Review Writing: Grammar, Usage, and Style Paperback - May 29, 2001 by Jean Eggenschwiler (Author), Emily Dotson Biggs (Author)

[28] Andrew Radford. An Introduction to English Sentence Structure, Cambridge, 2004, ISBN-13 978-0-511-50666-6

[29] Andrew Carnie. Syntax,Oxford (2001)

[30] Adrian Brasoveanu Anaphora Resolution Spring 2010, UCSC

[31] Massimo Poesio, Simone Ponzetto, and Yannick Versley. 2011. Computational models of anaphora resolution:A survey. Unpublished.

[32] Xiaofeng Yang, Jian Su, Jun Lang, Chew Lim Tan, and Sheng Li. 2008a. An entity-mention model for coreference resolution with inductive logic programming. *In Proceedings of ACL-08: HLT*, pages 843-851.

[33] Yimeng Zhang and Yangbo Zhu. Machine Learning for Coreference Resolution: Recent Developments

[34] Shumin Wu and Nicolas Nicolov. Coreference Resolution, A Machine Learning Approach

[35] Pradheep Elango. Coreference Resolution: A Survey,Technical Report, University of Wisconsin Madison.

[36] Hannaneh Hajishirzi Leila Zilles Daniel S. Weld Luke Zettlemoyer. Joint Coreference Resolution and Named-Entity Linking with Multi-pass Sieves. *In proceeding Conference on Empirical Methods on Natural Language Processing*, pages 289-299. ACL.

[37] Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov logic. *In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 649-658, Honolulu, HI. ACL.

[38] Shasha Liao and Ralph Grishman. Large Corpus-based Semantic Feature Extraction for Pronoun Coreference. *In Proceedings 23rd International Conference on Computational* pages 60-68

[39] Ruslan Mitkov. 1999. Anaphora resolution: The state of the art. Technical Report (Based on the COLING/ACL-98 tutorial on anaphora resolution), University of Wolverhampton, Wolverhampton.

[40] Fredrik Olsson. A Survey of Machine Learning for Reference Resolution in Textual Discourse. Technical report, Swedish Institute of Computer Science, Kista, 2004.

[41] Aria Haghighi and Dan Klein. 2009. Simple Coreference Resolution with Rich Syntactic and Semantic Features. In Proceedings of the 2009 Conference on Empirical Conference in Natural Language Processing.

[42] Delip Rao and Paul McNamee and Mark Dredze. Streaming cross document entity coreference resolution. *In: Conference on Computational Linguistics (COLING) (2010)*

[43] Rahman Altaf and Vincent Ng. 2009.Supervised models for coreference resolution. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968-977, Suntec. Vincent Ng

[44] David Bean and Ellen Riloff. 2004. Unsupervised learning of contextual role knowledge for coreference resolution. *In Proceedings of HLT/NAACL*, pages 297-304.

[45] Unsupervised Models for Coreference Resolution, Vincent Ng

[46] Hobbs, Jerry R., 1976. Pronoun Resolution. Research Report 76-1, Department of Computer Sciences, City College, City University of New York. August 1976.

[47] Hobbs, Jerry R., 1978, "Resolving Pronoun References", Lingua, Vol. 44, pp. 311-338. Also in Readings in Natural Language Processing, B. Grosz, K. Sparck-Jones, and B. Webber, editors, pp. 339-352, Morgan Kaufmann Publishers, Los Altos, California. (a shorter version of the original)

[48] Grosz, Barbara. J., Aravind. K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):202-225. (The paper originally appeared as an unpublished manuscript in 1986.).

[49] Gasperin Statistical anaphora resolution Caroline Gasperin and Ted Briscoe *in Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)* pages 257-264, Manchester , August 2008

[50] http://www.nactem.ac.uk/enju/ Enju parser

[51] Yusuke Miyao and Jun'ichi Tsujii. 2002. Maximum Entropy Estimation for Feature Forests. *In Proceedings of HLT 2002*.

[52] Yusuke Miyao and Jun'ichi Tsujii. 2003. Probabilistic modeling of argument structures including non-local dependencies. *In Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP)* 2003, pages, 285-291

[53] Yusuke Miyao, Takashi Ninomiya, and Jun'ichi Tsujii. 2004. Corpus-oriented Grammar Development for Acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank. In Proceedings of IJCNLP-04.

[54] Yusuke Miyao and Jun'ichi Tsujii. 2005. Probabilistic Disambiguation

Models for Wide-Coverage HPSG Parsing. *In Proceedings of ACL-2005*, pages. 83-90.

[55] http://www.nactem.ac.uk/y-matsu/geniass/ Genia sentence splitter

[56] Takashi Ninomiya, Takuya Matsuzaki, Yoshimasa Tsuruoka, Yusuke Miyao and Jun'ichi Tsujii. 2006. Extremely Lexicalized Models for Accurate and Fast HPSG Parsing. *In Proceedings of EMNLP 2006.*

[57] Takashi Ninomiya, Takuya Matsuzaki, Yusuke Miyao, and Jun'ichi Tsujii. 2007. A log-linear model with an n-gram reference distribution for accurate HPSG parsing. *In Proceedings of IWPT 2007*.

[58] Yusuke Miyao and Jun'ichi Tsujii. 2008. Feature Forest Models for Probabilistic HPSG Parsing. *Computational Linguistics.* 34(1), pages 35-80, MIT Press.

[59] http://www.nactem.ac.uk/enju/demo.html

[60] SÃ ¦ tre, Rune, Kazuhiro Yoshida, Akane Yakushiji, Yusuke Miyao, Yuichiro Matsubayashi and Tomoko Ohta., AKANE System: Protein-Protein Interaction Pairs in BioCreAtIvE2 Challenge, PPI-IPS subtask.*In Proceedings of the Second BioCreative Challenge Evaluation Workshop*. pages. 209-212, April 2007. CNIO.

[61] Yoshimasa Tsuruoka., A simple C++ library for maximum entropy classification, http://www-tsujii.is.s.u-tokyo.ac.jp/ tsuruoka/maxent/ , .

[62] Jin-Dong Kim, Tomoko Ohta, Yuka Tateishi, and Jun'ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics, 19(Suppl.1)*:180-182.

[63] http://2011.bionlp-st.org/ bionlp shared task 2011

[64] Miji Choi, Karin Verspoor and Justin Zobel.2014. Evaluation of coreference resolution for biomedical text.*In proceeding of MedIR 2014, page 2.*

[65] Yannick Versley, Alessandro Moschitti, Massimo Poesio and Xiaofeng Yang.2008. Coreference systems based on kernel methods. *In Proceedings of COLING*, pages 961-968.

[66] Ninomiya, Takashi, Yoshimasa Tsuruoka, Yusuke Miyao, and Jun'ichi Tsujii. 2005. Efficacy of beam thresholding, unification filtering and hybrid parsing in probabilistic HPSG parsing. *In Proceedings of the 9th International*

Workshop on Parsing Technologies, pages 103-114, Vancouver

[67] www.tagtog.net

[68] Cejuela,J.M., McQuilton,P., Ponting,L. et al. tagtog: interactive and text-mining-assisted annotation of gene mentions in PLOS full-text articles. Database (2014) Vol. 2014: article ID bau033; doi:10.1093/database/bau033.

[69] Yu hsiang Lin and Tyne Liang. 2004. Pronominal and sortal anaphora resolution for biomedical literature. *In In Proceedings of ROCLING XVI: Conference on Computational Linguistics and Speech Processing*.

[70] Jennifer D'Souza and Vincent Ng. Anaphora Resolution in Biomedical Literature: A Hybrid Approach. *In Proceeding of BCB '12 Proceedings of the ACM Conference on Bioinformatics*, Computational Biology and Biomedicine Pages 113-122

[71] http://bionlp-st.dbcls.jp/CO/eval-test/

[72] http://bionlp-st.dbcls.jp/CO/eval-development/

[73] Mitkov, Ruslan. 2002. Anaphora Resolution. London: Pearson Longman.