FAKULTÄT FÜR !NFORMATIK

Coreference Resolution in Clinical Practice Guidelines Focusing on Hypernym/Hyponym Relations

DIPLOMARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

im Rahmen des Studiums

Wirtschaftsinformatik

eingereicht von

Marco Romauch
Matrikelnummer 0226387

an der
Fakultät für Informatik der Technischen Universität Wien

Betreuung:
Betreuer/Betreuerin: Ao.Univ.Prof. Mag. Dr. Silvia Miksch
Mitwirkung: Mag. Dr. Katharina Kaiser

Wien, 24.04.2009

(Unterschrift Verfasser/in) (Unterschrift Betreuer/in)

Abstract

Medical knowledge is often only available in natural language text documents, which makes the automated processing of the information they contain a highly expensive, labour-intensive, and time-consuming task. Therefore, research efforts have been made to find ways of making selected medical documents processable for automated systems.

That applies also for clinical practice guidelines (CPGs) since these documents represent the state-of-the-art knowledge in a certain medical field. The use of computerised CPGs can be beneficial in several ways, especially in patient-specific decision support, since they provide the possibility to automatically generate recommendations about what medical procedures to perform tailored to an individual patient.

The proper automated processing of information provided by CPGs relies heavily on the correct interpretation of a certain semantic proposition in natural language text, namely coreference relations. Coreference detection and resolution is an important task in natural language processing (NLP). Two or more terms in a text are coreferent if they refer to the same real-world entity. Authors often use this semantic structure in order to prevent word repetition. Its correct interpretation helps to understand what is going on in a discourse of text. There exist several different types of coreference in natural language text such as name-alias coreference, pronoun coreference, and definite description coreference.

In this thesis work we will especially deal with the latter one. After the presentation of the theoretical background of coreference resolution, including an outline of existing algorithms and systems, we introduce our coreference resolution approach for CPGs. The focus lies on the detection and resolution of hypernym/hyponym coreference relations, a special kind of definite description coreference, since they represent the most frequent type found in CPGs. A hypernym/hyponym coreference exists if a coreferent relation holds between a more general expression (hypernym) and a more specific expression (hyponym). In order to accomplish this task the resolution algorithm firstly determines all possible phrases and selects the relevant ones for further processing. Secondly, we apply several tests that use information provided by external tools, namely MetaMap Transfer (MMTx) and the Unified Medical Language System (UMLS) in order to identify the candidates that can possibly be part of a coreference relation. Finally, a set of resolution rules is used to determine coreference relations that hold between the candidates.

We developed an initial algorithm and implemented it prototypically in order to test and improve it. The resulting algorithm was then evaluated with the help of set of test documents. During this evaluation our coreference resolution algorithm achieved 84,96% in recall and 68,49% in precision.

Kurzfassung

Medizinisches Wissen steht oft nur in Form von natürlichsprachigen Textdokumenten zur Verfügung. Dieser Umstand macht eine automatisierte Verarbeitung dieser Informationen zu einer extrem kostspieligen, arbeitsintensiven und zeitaufwändigen Tätigkeit. Aus diesem Grund wurden vielfach Anstrengungen mit dem Ziel unternommen, ausgewählte medizinische Dokumente automatisch verarbeitbar zu machen.

Diese Anstrengungen gelten besonders für medizinische Leitlinien (engl.: clinical practice guidelines (CPGs)), da diese Dokumente das aktuell gültige Wissen in einem bestimmten medizinischen Bereich repräsentieren. Die Verwendung rechnergestützter CPGs bietet verschiedenste Vorteile, besonders im Bereich der patientenspezifischen Entscheidungsunterstützung. Mit ihrer Hilfe ist es möglich, individuelle, auf Patienten speziell zugeschnittene Behandlungsvorschläge automatisch zu erstellen.

Die korrekte automatisierte Verarbeitung der Informationen in den CPGs beruht unter anderem auf der richtigen Interpretation eines speziellen semantischen Theorems, der so genannten Koreferenzbeziehung. Die Erkennung und Auflösung dieser Struktur ist eine wichtige Teilaufgabe im Bereich des Natural Language Processing (NLP). Zwei oder mehrere Ausdrücke in einem Text sind koreferent, wenn sie auf dasselbe reale Objekt referenzieren. Diese semantische Struktur wird oft zur Verhinderung von Wortwiederholungen eingesetzt. Eine korrekte Interpretation hilft dabei, den Inhalt eines Textes zu verstehen. Es existieren verschiedene Arten von Koreferenz in natürlichsprachigen Texten, wie zB Name-alias Koreferenz, Pronomen-Koreferenz und Definite-Description Koreferenz.

Diese Arbeit fokussiert auf die Identifizierung des letzteren Typus. Nach der Vorstellung des theoretischen Hintergrundes zum Thema Auflösung von Koreferenzbeziehungen, die auch einen Überblick über existierende Ansätze und Systeme beinhaltet, präsentiert diese Arbeit unseren Korefernzidentifizierungsalgorithmus für CPGs. Ein Hauptaugenmerk liegt auf der Erkennung und Auflösung von Definite Description Koreferenz, und dabei speziell auf hypernymen/hyponymen Korefernzbeziehungen. Diese stellen den in CPGs am häufigsten auftretenden Typ dar. Eine hypernyme/hyponyme Korefernz liegt dann vor, wenn eine Koreferenzbeziehung zwischen einem generelleren Ausdruck (Hypernym) und einem spezielleren Ausdruck (Hyponym) besteht. Um diese Aufgabe erfüllen zu können, identifiziert unser Algorithmus zuerst alle möglichen Phrasen und selektiert die relevanten für die weitere Verarbeitung. Im zweiten Schritt werden alle Kandidaten identifizieren, die möglicherweise Teile einer Korefernzbeziehung sind. Dazu verwenden wir verschiedene Tests die Informationen von externen Informationen, nämlich MetaMap Transfer (MMTx) und dem Unified Medical Language System (UMLS) beziehen. Schließlich werden „Resolution Rules“ eingesetzt um Koreferenzbeziehungen, die zwischen den Kandidaten existieren zu ermitteln.

Wir entwickelten einen Algorithmus, den wir prototypisch implementierten um ihn in weiterer Folge anhand von Trainingsdokumenten zu verbessern. Der endgültige Algorithmus wurde danach anhand von Testdokumenten evaluiert. Unser Algorithmus zur Identifizierung von Koreferenzen erreichte bei dieser Evaluierung Werte von 85,96% Vollständigkeit (Recall) und 68,49% Genauigkeit (Precision).

Table of Contents

List of Tables

Table of Figures

1 INTRODUCTION

"The doctor is often more to be feared than the disease."
Latin Proverb

1.1 Motivation

Nowadays, critical decision-making in several crucial fields such as economy and engineering relies heavily on information provided by specialised computer system. Those applications support the decision-maker by automatically processing the available input data in consideration of valid domain depended information and knowledge. Additionally they give suggestions and recommendations about the possible effects and outcomes of a certain decision.

This scenario does not seem to exist in medical science. In this major field of human well-being, quick and correct decision-making is utterly essential. Nevertheless, a medical doctor, who in this case is the decision-maker, mainly relies on his personal knowledge and experience gained during his studies or professional life. Although this is adequate in most cases, there are various situations in which the incorporation of a sophisticated supporting system can be beneficial. A precondition for the development of such a system is the availability of all essential valid medical domain knowledge. Unfortunately, the majority of medical information is only available in natural language text such as in the form of clinical practice guidelines (CPG).

Per definition, CPGs are "systematically developed statements to assist practitioner and patient decisions about appropriate health care for specific clinical circumstances" [Field and Lohr, 1990]. CPGs play an important role in various fields of medical healthcare. These documents briefly identify, summarize, and evaluate the best evidence and most current data about prevention, diagnosis, prognosis, management, and therapy. They also serve as a set of recommendations concerning these clinical actions by pointing out potential treatment options and corresponding outcomes. CPGs furthermore aim to minimize errors and provide consistent quality in care by reducing variation in practice and setting up certain standardized procedures. In a nutshell, a CPG represents the state-of-the-art knowledge in a certain medical field.

Since most CPGs exist only in natural language text it is extremely difficult to integrate them and the information they hold in electronic clinical supporting or patient data management systems although this would be a desirable step in order to improve clinical decision-making. The use of computerised CPGs can be beneficial in several ways, especially in patient specific decision support. For example, they provide the possibility to automatically generate recommendations about what medical procedures to perform tailored for an individual patient.

Considering this fact, medical science shows great interest in finding ways of making CPGs computer-interpretable. Existing computerised formats require the manual adaptation of regular CPG documents which is a highly expensive, labour-intensive, and time-consuming task. Latest research results present natural language processing (NLP) as a promising approach to find a (semi-)automatic method for the creation of computer-interpretable CPG documents.

1.2 Background

Correct interpretation of certain semantic propositions in natural language texts is an important task to provide high quality results in many fields of natural language processing (NLP), such as information extraction, question answering, and text summarization.

A NLP system usually consists of several subtasks that form a NLP pipeline and therefore process one step after another. An important task is the detection and resolution of so-called coreference relations both within and across sentences. Coreference is a certain linguistic structure that holds between two textual expressions whereas both are related to the same referent in the real world. Such a proposition can be frequently observed in natural language text corpora since a human author tries to avoid word repetition by using a variety of noun phrases that describe the same object. “While humans have little trouble mapping a collection of noun phrases onto the same entity, this task of noun phrase (NP) coreference resolution can present a formidable challenge to a NLP system.” [Cardie and Wagstaff, 1999] The ability to discover and resolve such linguistic structure gives a NLP system the potential to interpret natural language texts correctly or in other words it helps the system to understand what is going on in a discourse of text.

Due to the complexity of natural language several types of coreference exist in natural language texts. They can be basically divided into three main groups:

(1) **Name-alias coreference:** A coreferent relation holds between two expressions that stand for the same name (“Marco Romauch” – “Mr. Romauch”)

(2) **Pronoun coreference:** A coreferent relation holds between a pronoun (“he”, “it”, “they”…) and its antecedent substantive.

(3) **Definite description coreference:** A coreferent relation holds between two definite terms that can only be resolved with the help of domain specific background knowledge ("amoxicillin" – "the antibiotic")

The computational resolution of the different types of coreference requires the application of different types of background knowledge. These information sources range from domain independent syntactic knowledge to process pronoun coreference to highly sophisticated domain specific semantic knowledge to resolve definite description coreference. Consequently NLP applications, like the one presented in this work, that aim to resolute definite description coreference have to be designed to operate in a specific domain dictated by the broad subject matter of the processed texts. In this particular case the analysed documents are CPGs and therefore it is necessary to develop an effective coreference processing approach considering that specific domain of discourse.

1.3 Overview of the Thesis

After a short introduction, including the motivation and background of this work, Chapter 2 analyzes the requirements and gives the basic theoretical knowledge necessary to design a coreference detection and resolution algorithm. Next to a classification of the two similar concepts of anaphora and coreference, this section describes the several different types of coreference that are presented in computational linguistic literature as well as a method that helps to measure the performance of coreference resolution algorithms.

Chapter 3 focuses on the theoretical background of coreference resolution algorithms. At first, a general resolution approach including its sub-steps is presented. Secondly, this section embraces an analysis of existing coreference and anaphora resolution algorithms including examples in the (bio)medical domain. The approaches investigated range from the very beginning of research interest in this area to state of the art algorithms and include both knowledge-based as well as machine learning systems.

Chapter 4 presents resources required for definite description coreference in the (bio)medical domain, namely the UMLS (Unified Medical Language System) and its three knowledge sources, the Metathesaurus, the Semantic Network, and the Syntactic Lexicon. Furthermore, the functionality of the MetaMap algorithm which maps (bio)medical text to UMLS concepts is explained with the help of a detailed example.

The main part of this thesis is presented in Chapter 5. It contains a detailed description of our developed coreference resolution algorithm. Therefore, its three main steps phrase detection, relevant markable determination, and the actual coreference resolution is subject to a deep inside investigation. This section also contains a description of the three coreference types our approach is able to resolve.

The performance evaluation of our coreference resolution algorithm is the main topic of Chapter 6. It describes how this process is performed and what information is required in order to calculate the main performance measures. Additionally, with respect to this theoretical background the actual results of our algorithm are also presented and interpreted in this section.

In the final chapter, we summarize the theoretical background as well as the developed coreference resolution algorithm and the results we were able to achieve. Finally, we take a look in the future and present ideas how the approach can be improved.

2 PROBLEM ANALYSIS

This section gives the basic theoretical knowledge necessary to design a coreference detection and resolution system. First of all linguistic background knowledge about the theoretical concept of the coreference phenomenon is given. Additionally, the most important types of coreference concerning CPGs and the corresponding resolution strategies are presented. Finally a scoring system is introduced, that makes it possible to measure and consequently improve the performance of different coreference resolution approaches.

2.1 Linguistic Definitions

Several significant linguistic terms are used throughout this thesis. This section serves as an aggregation in order to provide a clear definition and to avoid ambiguous understanding [Trask, 1993].

phrase: "A phrase is a group of words that functions as a single unit in the syntax of a sentence."

noun phrase: "A noun phrase (abbreviated NP) is a phrase whose head is a noun or a pronoun, optionally accompanied by a set of modifiers."

headword: "The head of a noun phrase. A word that is qualified by modifiers."

2.2 Coreference vs. Anaphora

Since this work is primarily about the detection and resolution of coreferent relations it is firstly necessary to define the theoretical concepts behind this semantic phenomenon and distinguish it from other linguistic structures that hold between parts of a text.

[Kibble and van Deemter, 1999] point out that "the terms coreference and anaphora tend to be used inconsistently and interchangeably in much empirically-orientated work in NLP, and this threatens to lead to incoherent analyses of text and arbitrary loss of information."

Therefore, it is elementary to clarify the notions of these two semantic concepts in order to avoid ambiguous understanding.

The following example shows the usage of coreference and anaphora:

"Prevention of the disease, or failing that, minimising its consequences by early detection, are key goals."

In this sentence, a dependent relation exists between the noun phrase "the disease" and the pronoun "its". In linguistics, such a relation is called an anaphoric relation between one expression, the anaphor, that points back to another earlier mentioned expression, the antecedent. The antecedent ("the disease") provides the information the reader requires in

order to interpret the anaphor (“its”). This knowledge is necessary to fully and correctly understand the sentence.

Additionally another semantic relation exists between the two expressions “the disease” and “its”. They both refer to the same real world entity, or in other words they are coreferent. The ability to detect and resolve such coreferential relations is critical to discourse analysis and language understanding in general [Soon et al., 2001].

[Trask, 1993] gives some textbook definitions for both relation types:

Anaphora: “An item with little or no intrinsic meaning or reference which takes its interpretation from other item in the same sentence or discourse, its antecedent.”

Coreference: “The relation which obtains between two noun phrases (usually two NPs in a single sentence) both of which are interpreted as referring to the same extralinguistic entity.”

Considering these definitions and according to [van Deemter and Kibble, 2000] and [Kibble and van Deemter, 1999], coreference and anaphora are two different things. The following facts help to explain why:

- Coreference is an equivalence relation, which means it is reflexive, transitive and symmetrical. This does not generally hold for anaphoric relations.
- Anaphora in contrast to coreference is context-sensitive of interpretation. This implies that a coreferential relation can hold between two non-anaphoric noun phrases if both refer independently to the same real world entity without mutual dependence.
- In principle anaphora can be without coreference, because anaphora, unlike coreference, does not require a referent in the real or conceptual world.

This leads to the conclusion that “anaphoric and coreferential relations can coincide, of course, but not all coreferential relations are anaphoric, nor are all anaphoric relations coreferential” [van Deemter and Kibble, 2000].

Actually, it is almost always true that coreference and anaphora coexist, or in other words, that two expressions refer to the same real world entity and one of these expressions, the anaphor, depends on the other expression, the antecedent, for its interpretation. Due to this fact, the terms antecedent and anaphor are frequently used in literature concerning coreference detection and resolution and therefore this work adapts to this terminology.

2.3 Types of Coreference

Many types of coreference exist in natural language texts. Unfortunately, literature does not present a distinct classification schema. According to [Bagga, 1998] there are eleven coreference classes whereas on the other hand [Denber, 1998] identifies only six types.

Due to this ambiguousness and the fact that different types of coreference are unequally important in different domains of discourse this work focuses on the detection and resolution of coreference types that occur frequently in clinical guidelines. It is highly

important to be aware of these types of coreference in order to guarantee best possible background knowledge for the detection and resolution process.

A semantic analysis of three Scottish Intercollegiate Guidelines Network (SIGN)[1] guideline documents [SIGN, 2003a], [SIGN, 2003b] and [SIGN, 2003c] pointed out that the most frequent coreference types used in this kind of texts is definite description and demonstrative description coreference. On the other hand only a small percentage of coreference found in the guidelines was pronominal. This assumption is also confirmed by other sources. [Torii and Vijay-Shanker, 2007], for example, investigated 50 Medline abstracts. They found slightly over hundred sortal (definite and demonstrative) anaphoric expressions and by contrast only four occurrences of "they", none of "he" or "she" and just seven anaphoric uses of "it". The following sections aim to give a short description of the most relevant types of coreference.

2.3.1 Definite/Demonstrative Description Coreference

In this case the coreference term is either a noun phrase preceded by a definite article ("the") or a demonstrative determiner ("this", "that", "these", "those") [Poesio and Vieira, 1998]. According to [Lin and Liang, 2004] also noun phrases with the modifiers "either", "both" or "each" have to be considered as member of this type of coreference.

[Vieira et al., 2003] distinguish between two different classes of definite description coreference in English natural language text:

1) **Same head coreference**
   The coreferent expressions share the same headword. This class is also called *direct coreference*. ("the cancer" – "the brain cancer")

2) **Bridging coreference**
   In this class the coreferent expressions have a different headword. This type is also called *indirect coreference*. ("amoxicillin" – "the antibiotic")

Since bridging descriptions cannot be identified through headword equality, a more complex form of lexical or common sense knowledge is necessary to detect and resolve this type of coreference. [Vieira and Teufel, 1997] identified several classes of bridging descriptions that can be resolved using a lexical knowledge base. Among these classes there are two, namely synonym ("the tumour" – "the cancer") and hypernym/hyponym ("the drug" – "the antibiotic"), which can be often found in guideline documents.

2.3.2 Hypernym/Hyponym Coreference

Even though hypernym/hyponym coreference is a subtype of bridging coreference it deserves a separate consideration because of its frequent occurrence in CPG texts.

[Rindflesch and Fiszman, 2003] define the hypernymic proposition as "a semantic structure in which two concepts, a hyponym and a hypernym, are in a taxonomic relation." A hypernym/hyponym coreference exists if a coreferent relation holds between a more general expression (hypernym) and a more specific expression (hyponym).

[1] http://www.sign.ac.uk/ (last assessed: March 12, 2009)

There are three major syntactic strategies that encode a hypernymic proposition in English natural language texts [Rindflesch and Fiszman, 2003]:

1. The specific NP is subject of the verb "be" and the general NP is represented by its complement. Other verbs such as "remains" are also possible.

   "Nimodipine is an isopropyl calcium channel blocker which readily crosses the blood–brain barrier."

2. Two NPs occur next to each other and they are separated by commas or parentheses. The NPs can also be linked by lexical items like "such as", "including", "especially" and "particularly" (see [Hearst, 1992]).

   "Non-steroidal anti-inflammatory drugs such as indomethacin attenuate inflammatory reactions."

3. The hypernym and the hyponym term occur in the same NP. One concept represents the head noun while the other one serves as a modifier.

   "An increase in blood pressure was also seen in patients who were taking adjunctive antihypertensive medications."

Nevertheless, hypernym/hypnonym coreference can also be frequently found without encoded in such syntactic patterns as shown in the following example:

"To reduce the incidence and mortality rate of cervix cancer, effective screening and preventive strategy must be actively pursued, in addition to early detection of the disease."

A main goal of this work is the development of a proper resolution strategy for such type of coreference.

2.3.3 Pronominal Coreference

Although pronominal coreference is very common in regular natural language text (see [Bagga, 1998]) it represents only a small percentage of coreference found in CPGs. Nevertheless it cannot be ignored during a coreference processing task.

Pronouns that occur in CPG texts are mostly neuter third person and reflexive pronouns. Following [Lin and Liang, 2004] noun phrases with "it", "its", itself", "they", "them", "themselves" and "their" have to be considered as coreference terms in this specific domain of discourse.

A special type of pronouns that can be found in natural language text are pleonastic pronouns. Such pronouns (usually "it") do not refer to any particular antecedent and therefore cannot be considered anaphoric or coreferent. Nevertheless, a prior identification of such occurrences is important so that the coreference resolution system does not attempt to determine a correct antecedent [Dimitrov, 2002].

2.4 Coreference Chain

Coreferent relations do not exist exclusively between two expressions. The following example shows a so-called coreference chain that consists of three coreferring expressions:

> "Melanoma, especially when diagnosed at an advanced stage, can cause serious morbidity and may be fatal despite treatment. Prevention of the disease, or failing that, minimising its consequences by early detection, are key goals."

The noun phrase "the disease" refers to the earlier mentioned expression "Melanoma" and additionally the pronoun "its" refers to "the disease" as in the above example. This leads to a coreference chain "Melanoma" → "the disease" → "its".

The ability to detect coreference chains is very important in many natural language processing tasks in order to enhance systems discourse analysis ability.

2.5 Scoring

Like any other NLP system, a coreference detection and resolution system requires scores, or metrics, in order to measure and improve its performance and eventually compare it with results provided by similar systems or humans.

According to [Lehnert et al., 1994] there are two main evaluation scores in NLP:

- **Recall**
  "The recall score measures the ratio of correct information extracted from the texts against all the available information present in the texts."

  recall = correct / available

- **Precision**
  "The precision score measures the ratio of correct information that was extracted against all the information that was extracted."

  precision = correct / extracted

In coreference processing a high recall score indicates that almost all existing coreference relations were found while a high precision score indicates that almost all found coreference relations are relevant. It is difficult to optimize both scores, because an increase in one score most likely results in a decrease of the other.

Next to these two single scores another combined measure of recall and precision exists. It is called F-measure and serves as an indicator for overall performance of the system. The F-measure was introduced to compare the performance of two or more systems that process the same text corpora, since it is almost impossible to do that on the basis of two separate scores. Consequently, the F-measure of a sole system is not significant.

F-measure = (P * R) / (b * P + (1-b) * R)

The formula uses three variables precision (P), recall (R) and b that is the relative importance of recall over precision. Values of b can vary from 0 to 1. The most common form is 0.5, which gives precision and recall an equal influence on the overall score.

3 RELATED WORK

After a short introduction in a general approach to coreference resolution this chapter presents different computational coreference resolution systems introduced in computational linguistic literature from the very beginning of research interest in this area to state of the art algorithms. Both knowledge-based and corpus-based approaches will be considered in this outline that embraces domain independent solutions as well as approaches designed for the (bio)medical domain. Selected algorithms will be analyzed in order to develop effective strategies for coreference resolution concerning medical domain texts, especially clinical practice guidelines.

3.1 A General Approach to Coreference Resolution

Accurate coreference resolution is an essential task for any kind of text understanding system. The process can be typically divided into two main stages:

1. Textual elements that could be part of potential coreferential relations, so-called markables, have to be determined in the given text corpora. During this step linguistic information concerning the markables is also collected.

2. A resolution algorithm attempts to resolve coreferential relationships between each markable and its correct antecedent. This is the most challenging task in the whole process of coreference resolution, because in most cases the possible anaphor holds not only one but a set of potential antecedent candidates. Typically all markables preceding the current markable have to be considered as candidates. At first a filter mechanism eliminates candidates that are incompatible with the markable under investigation. In a second step the most likely antecedent is selected among the remaining candidates. Both, filtering and selection the correct candidate require the application of different types of knowledge source.

3.1.1 Knowledge Sources for Coreference Resolution

The resolution of coreference relations in a given text requires considerable background knowledge. According to [Hoste, 2005] this information includes morphological and lexical knowledge like number agreement and the knowledge about the type of the markables, syntactic knowledge such as the grammatical functions of anaphor and potential antecedent within the sentence and semantic knowledge which allows to identify semantic propositions such as synonyms or hypernyms/hyponyms. [Mitkov, 2003] points out that discourse knowledge is also valuable in case of antecedent selection because the most salient element among the potential candidates is most likely the right antecedent. In some cases, however, not even the most extensive morphological, lexical, syntactic, semantic and discourse knowledge can provide the information necessary to select the correct antecedent when real-world knowledge is required in order to find the candidate that makes common sense. Systems that are designed to handle such phenomenon rely mostly on hand-crafted resourses of lexico-semantic knowledge provided through ontologies or digital lexical

resources like WordNet [Feldbaum, 1998] or UMLS [Humphreys et al., 1998] for the medical domain.

3.1.2 Markable Determination

A crucial prerequisite for accurate coreference resolution is the determination of all discourse entities that could be involved in a coreferential relationship. These so-called markables are the union of pronouns, definite and demonstrative description noun phrases found in the text corpora [Soon et al., 2001]. In order to obtain all of the markables the raw input text runs through a NLP pipeline that consists of several text-processing modules. The goal of this preprocessing step is to locate all markables including their relevant linguistic and textual boundary information required in the following coreference resolution steps. Figure 1 schematically illustrates this preprocessing procedure.

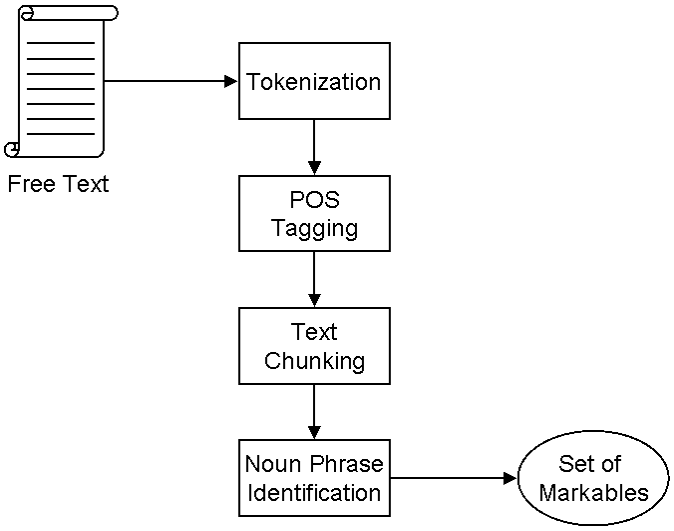

Figure 1: NLP pipeline for markable determination

The NLP pipeline consists of the following modules [Soon et al., 2001][Hoste, 2005]:

- **Tokenization**
  The text is split into sentences and sentences are split into words, which represent the smallest linguistic units with semantic meaning.

- **Part-of-speech tagging**
  This step identifies the linguistic categories (noun, verb, adjective…) of words found during the tokenization stage.

- **Text chunking**
  Syntactically related words that were identified in the input text are combined to non-overlapping phrases.

- **Noun phrase identification**
  For the determination of markables only noun phrases including a headword and eventual existing premodifieres such as adjectives and determiners are selected.

3.1.3 Correct Antecedent Candidate Selection

The selection of the correct antecedent among a set of possible candidate markables is a complex step since the decision-making process involves several sources of background knowledge. Different directions can be taken in using this information in order to find the correct antecedent for a certain markable. Literature concerning coreference and anaphora resolution presents two main strategies for solving this problem [Mitkov, 1999][Mitkov, 2003][Eiken, 2005]:

1. **Knowledge-based approaches** filter correct candidates using hand-coded rules that represent linguistic, domain and/or real-world knowledge.

2. **Corpus-based approaches** find the most likely candidate based on statistical or machine-learning techniques.

The first coreference resolution systems presented in computational linguistic literature were knowledge-based. These approaches can be basically divided in “approaches that generally depend upon linguistic knowledge and approaches in which discourse structure is taken into account” [Hoste, 2005].

The use of constraints and preferences in knowledge-based approaches

Approaches that belong to the first group use linguistic knowledge such as lexical, morphological, syntactic and semantic information in order to define resolution factors. According to [Mitkov, 2003] there are two types of such factors, namely constraints and preferences. At first constraints eliminate unlikely markables from the set of possible candidates. In a second step preferences are used to favor certain candidates over other with the help of preference rules.

Centering theory in knowledge-based approaches

Discourse structure can also be applied in order to resolve corefernce relations. Focusing and centering theory assume that “certain entities mentioned in an utterance are more central/in focus than others and this imposes certain constraints on the referential relations in a text” [Hoste, 2005]. Literature presents forward as well as backward looking centers, which are in both cases sets of markables ranked by their number of appearances in the text corpora. Finally constraints are used to select the correct antecedent among the set of candidates.

The advantages of corpus-based approaches

The main problem that comes with knowledge-based approaches for coreference resolution is the high amount of human input like labor and knowledge that is necessary to build and maintain such systems in order to make them work properly in a wider range of domains. Computational linguistic literature points out that as a consequence the focus in research over the past years has changed to corpus-based systems. These approaches are based on the theory that the correct antecedent among a set of candidates can be found with information presented in annotated text corpora. A prerequisite for the rise of corpus-based

techniques was the fact that coreferentially annotated corpora have become better, larger, and more available.

Some approaches obtain collocation patterns from coreferentially annotated text while other techniques mind linguistic rules which are used in combination with certain heuristics in order to filter out unlikely antecedent candidates.

Machine learning techniques in corpus-based approaches

Most of corpus-based approaches for coreference resolution, however, use machine-learning techniques. The information, if two markables are coreferent or not, is represented in a feature vector. It contains "distance, semantic, syntactic, morphological and lexical information on the candidate anaphor, its candidate antecedent and the relation between both" [Hoste, 2005]. The goal of these techniques is to train a machine learner to automatically decide if a certain pair of markables is coreferential or not. Literature presents two types of machine learning techniques. In supervised approaches the machine learner receives feedback about the coreferentiallity of a given pair of markables whereas in unsupervised methods there is no such information. The baseline-supervised machine learning approach can be recast as a classification task. A coreference classifier is trained on training texts for determining if two markables represented by a feature vector are coreferent or not. The instances are labelled positive if a coreferent relation holds between the two markables or negative if not. For accurate coreference resolution the classifier uses the information gained during the training stage to select the correct antecedent among a set of potential candidates.

3.2 Computational Approaches to Coreference Resolution

This section aims to present an outline of coreference resolution approaches introduced in computational linguistic literature over the last decades. The goal of this study is to find strategies that help to solve the problem of coreference resolution in clinical practice guidelines. Therefore the main emphasis is on approaches designed for the (bio)medical domain even though the problem of coreference resolution in contrast to anaphora resolution is not frequently tackled in this specific application area. As a result of this observation domain independent algorithms are also kept in mind and they are analyzed for strategies that can be applied in order to resolve coreference relations in CPG documents. Since different types of coreference require different resolution strategies this outline also focuses on approaches that are able to handle definite/demonstrative description coreference especially when encountered in a synonym or hypernym/hyponym relation (bridging coreference), because this type is, unlike pronominal coreference, very common in guideline texts.

At the beginning it should also be noted that some of the approaches presented below were originally designed to compute anaphoric relations. This has two reasons:

1. The main research focus during the time they have been introduced lay on anaphora and not on coreference resolution.

2. Research in anaphora resolution is still popular in the computational linguistic community and remains as important as ever for NLP systems.

Nevertheless the presented anaphora resolution algorithms could basically also be used to detect and resolve coreference relations since, as stated above, in this work it is almost always true that coreference and anaphora coexist.

3.2.1 General Knowledge-Based Approaches

Knowledge-based approaches for coreference/anaphora resolution usually use linguistic (lexical, morphological, syntactic and semantic), domain and/or real-world knowledge in order to define two types of hand-crafted rules:

1. **Constraints** are used to filter out unlikely antecedent candidates in the first place.

2. **Preferences** are subsequently applied to select the most likely antecedent among the remaining candidates.

Hobbs' tree search algorithm

One of the first knowledge-based approaches presented was the tree search algorithm by [Hobbs, 1978]. It is designed to find the correct antecedents for anaphoric pronouns. The algorithm relies mostly on syntactic knowledge and implies constraints concerning this kind of information. Therefore it uses a surface parse tree, which generates a search space that includes all previous sentences in the text corpora, as well as the sentence with the anaphor to be resolved. This tree is basically a syntactic representation of the discourse text under investigation. In order to find the correct antecedent the tree is searched in a left-to-right, breath-first manner. During this process the syntactic constraints are used to filter out unlikely NPs. The search terminates when a noun phrase is found that matches in gender, number and person with the anaphoric pronoun in question.

A full syntactic parsing algorithm by Lappin and Leass

Another approach for the resolution of pronominal anaphora is the one by [Lappin and Leass, 1994]. Their full syntactic parsing algorithm relies on measure of salience derived from syntactical structure of the text corpora in order to select the correct antecedent among a list of potential candidates. No semantic conditions or real-world knowledge is employed during the resolution process. At first syntactical and morphological constraints are applied to filter out unlikely NPs. Then the remaining NPs are weighted with several salience parameters (such as grammatical role, parallelism of grammatical roles, frequency of mention, proximity and sentence recency) in order to prefer certain candidates over other. Finally, the candidate with the highest salience score gets selected as the correct antecedent.

The CogNIAC system by Baldwin

Due to the high error rate of full syntactic parsing algorithms, like the one mentioned above, several alternative approaches have been proposed whereas most of them use a part-of-speech tagger and other text processing units like NP recognition in order to derive more sophisticated linguistic information about the lexical items found in the text corpus.

One of these approaches is the CogNIAC system by [Baldwin, 1997] for the resolution of anaphoric pronouns. As a preprocessing step the input text runs through a part-of-speech tagging and noun phrase recognition module. The resolution process itself is performed with the use of a limited set of predefined anaphora resolution rules. For each pronoun found in the text starting from left to right the rules are applied in a predefined order. If a correct antecedent is found with the help of one rule, no further rules are tried for that specific anaphoric pronoun. If no antecedent can be found after the application of all existing rules, the pronoun remains unresolved.

Mitkov's pronoun resolution approach

Another example is the pronoun resolution approach by [Mitkov, 1998]. At first the input text runs through a part-of-speech tagger in order to identify the textual entities and gather their linguistic information. For each pronoun all preceding NPs within a two-sentence distance are identified and considered as possible antecedents. Noun phrases that do not agree in gender and number to the anaphor are filtered out. In the next step the remaining antecedent candidates are ranked by the application of so-called antecedent indicators, such as definiteness, lexical reiteration, distance, etc. Candidates are assigned a positive or negative score (2, 1, 0, -1) for each indicator and the candidate with the highest aggregate score is selected as the correct antecedent.

A semantic driven algorithm by Munoz and Palomar

All the presented approaches so far can only handle pronominal coreference/anaphora. [Munoz and Palomar, 2001] present a semantic-driven algorithm for definite description (DD) resolution. They state that this type of coreference is more difficult to treat, because of two reasons:

1. The distance between a DD and its corresponding antecedent can be much larger than between a pronoun and its antecedent. Additionally DD hold more semantic information than pronouns.

2. DD unlike pronouns are not always anaphoric. They may introduce a new entity in the discourse instead of refer to a preceding NP.

In order to handle this more complex type of coreference their algorithm consists of three main components, namely a semantic network generation module, a set of semantic constraints and a set of preferences obtained from an empirical study.

At first a semantic network is generated in order to establish a mechanism that previously identifies anaphoric and non-anaphoric DD. The authors state that "a DD will not be anaphoric if it is not semantically compatible to a previous NP". But if one DD exist that belongs to the same semantic category like any other previous found NP, only the application of a resolution algorithm can tell if the NP is anaphoric or not. Therefore, the ontological (semantic) concept of every NP's headword found in the input text is extracted from the lexical resource WordNet [Feldbaum, 1998]. Furthermore, the system distinguishes between the different possible types of the NP. If the NP under consideration is not a DD then the NP is added to the list of the corresponding ontological concept. If the list does not exist yet, it is created firstly and then the NP is added as its first member. If the NP under consideration is a DD then the system checks if the corresponding ontological concept

already exists in the semantic network. If it does not exist the DD and its ontological concept are added to the semantic network and the DD under consideration is classified non-anaphoric. However, if the ontological concept already exists all of its member head nouns are semantically compared to the head noun of the DD under consideration. If no semantically related member can be found the DD is classified non-anaphoric. But, if there exists any semantic relation between head nouns, an algorithm to solve references has to be applied in order to classify the DD under consideration as anaphoric, which furthermore includes the selection of the correct antecedent, or non-anaphoric. The resolution algorithm applies constraints and preferences to the members of the corresponding ontological concept in order to select the correct antecedent if any exists.

At first a set of semantic constraints is used in order to filter out unlikely anaphoric NP-DD dependencies due to non-compatible semantic relations. The authors present two constraints that rule out possible candidates:

1. Two members of the same ontological concept can only be coreferent if they share an equivalent head noun (same head coreference) or if their head nouns are in a synonym or hypernym/hyponym relationship (bridging coreference).

2. If a semantic comparison of the modifiers of the DD under consideration and a NP results in an antonym relationship they cannot be coreferent.

In a second step a set of six preferences obtained from an empirical study is used to select the correct antecedent for a given anaphoric DD among the remaining members in the same ontological concept. The system defines the set of preferences as follows:

1. Previous appearances of the same DD (same head noun and modifiers).

2. Semantic relation (synonym or hypernym/hyponym) of the candidate NP and the DD modifiers plus same head noun.

3. Bridging relation (synonym or hypernym/hyponym) of candidate NP and DD head noun.

4. Antecedent without modifiers. Candidates with an equivalent head noun are selected over ones with a semantic relation between head nouns.

5. Gender and number agreement of candidate NP and DD.

6. Closest antecedent candidate. This preference is only applied if there is still more than one candidate left.

The approach presents two ways of managing the sets of preferences:

1. An ordered application that dismisses candidates that do not fulfill a preference if there is at least one that does.

2. A weight management where experimentally chosen salience values are applied for each preference in order to select the candidate with the highest aggregate score as the correct antecedent.

The authors also state in their conclusion that the weight management approach outperforms the ordered application approach.

3.2.2 General Machine Learning Approaches

Machine learning approaches for coreference resolution have become very popular over the last years since coreferentially annotated corpora have become better, larger, and more available. As mentioned above there are supervised and unsupervised approaches, whereas the vast majority presented in literature is supervised.

Soon et al.'s machine learning approach baseline system

The basic concept for supervised learning systems is presented in [Soon et al., 2001]. The center part of the approach is a trained classifier that is derived from a C5 decision tree learning algorithm, which is an updated version of C4.5 [Quinlan, 1993]. For the classifier-training phase a set of training documents is manually annotated with coreference chains of NPs. Therefore all markables in the documents have to be obtained firstly. The markable determination for the training documents as well as during the actual coreference resolution step is processed through a pipeline of language-processing modules (e.g., part-of-speech tagging, noun phrase identification) that also collect linguistic information about the determined markables. This knowledge is subsequently used to form feature vectors, which are sets of features that describe pairs of markables and hold their coreferentiallity information. Each markable starting with the most right one is paired with every one of its antecedents to form an instance and then associated with a feature vector. During training an instance is only labeled as a positive example if its markables are immediately adjacent in the same coreference chain. Instances can also be labeled negative if markables that belong to no or any other coreference chain exist between the two markables of the investigated antecedent-anaphor pair. After training the classifier is able to process a new document. Each test instance derived from the text corpora is associated with the corresponding feature vector. The trained classifier returns a number between 0 and 1, which represents the likelihood that a coreferential relation holds between the two analyzed NPs. This confidence value is used to rank the markables among the set of potential candidates. Unlikely markable pairs with a score under 0.5 are filtered out. Finally, a "Closest First" algorithm, which selects the candidate closest to the anaphor, is applied to choose the correct antecedent if any exists.

The selection of the features used in the feature vector is the one of the most essential task in designing a machine learning system since the information they provide is used to select the correct antecedent. Literature presents different approaches that vary mostly in number, type (morphological, syntactic, semantic, etc.) and information granularity of the feature vector. [Soon et al., 2001], for example, use twelve features in their baseline system. Five features (two pronouns and a proper name, a definite and a demonstrative noun phrase feature respectively) indicate the type of the markables whereof others are related to their gender and number. A semantic class feature uses information derived from WordNet [Feldbaum, 1998] to test the semantic compatibility of the two markables. Additionally, there is a distance feature that measures the number of sentences between the two markables, a string match feature that tests if the two NPs under consideration are the same string after removing determiners, an alias feature that checks if one of the two markables is

an alias of the other, and an appositive feature that tests if one of the NPs is in apposition to the other.

Since the results of the baseline system were not satisfying enough numerous attempts to improve this supervised machine learning approach have been introduced in NLP literature over the last years.

Two improvements to the baseline system by Ng and Cardie

[Ng and Cardie, 2002a] for example propose three extra-linguistic changes to the machine learning framework and expand the set of features used in the baseline system from 12 to 53. The modifications affect mainly the correct candidate selection algorithm. Instead of choosing the closest NP with a confidence value above 0.5, they suggest to select the candidate with the highest likelihood score. This, however, requires also a different method for generating examples during the classifiers training stage since now the markable pairs with the most likely and not the one with the closest antecedent candidate have to be labeled positive. A third change concerns the string match feature of the baseline system, which is split into several less complex features one for each type of coreference. This first improvement leads to a significant gain in precision. The expansion of the feature set includes a small number of additionally lexical, semantic, and knowledge-based features and a huge increase of grammatical features. This leads to "more complex string matching operations, finer-grained semantic compatibility tests and more sophisticated syntactic coreference resolution rules" [Ng and Cardie, 2002a]. The drawback of this approach, however, lies in the significant performance drop when using the full set of the introduced features. The use of the full 53 feature vectors leads to a significant increase in recall but also to a large decrease in precision. As a result, features, which are responsible for the low precision score are manually eliminated. This action leads to a significantly gain in precision with only a small drop in recall.

In a second attempt [Ng and Cardie, 2002b] tried to increase the precision score of their approach by incorporating an anaphoricity determination component as a preprocessing filter for the actual coreference determination algorithm. Instead of comparing every found NP with every preceding NP like in the baseline system, this trained classifier checks a prior if a given NP is anaphoric or not. If the test result is positive the NP under consideration is considered anaphoric and hence can be compared to the preceding NPs.

The application of string information by Strube et al.

[Strube et al., 2002] introduce an additional feature based on the minimum edit distance (MED) of two strings. "The MED computes the similarity of strings by taking into account the minimum number of editing operations (substitutions, insertions, deletions) needed to transform one string into the other" [Strube et al., 2002]. The feature calculates the minimum edit distance from antecedent to anaphor and vice versa, which leads to a significant performance improvement especially for definite noun phrase coreference.

Yang et al.'s improvements to the baseline system

The approach by [Yang et al., 2005] presents two innovations to the baseline system. Firstly it introduces a set of several string match features that in contrast to a simple headword and full-string comparison also include an accurate investigation of the modifiers (adjective,

preposition, number, possessive, proper noun nonfinite and quantifier) of a markables since these words usually hold essential information in order to perform correct coreference resolution. This set of features is used to capture matching patterns in the modifiers of two NPs. Subsequently the matching degree of the markables is computed including three string distance metrics and two additionally weighting schemes. The performance of the system including the modifier match feature set shows a gain in recall compared with one that only uses the full-string match feature (tightest matching check) and a gain in precision compared with one exclusively using the headword match feature (loosest matching check). A second improvement presented in this approach concerns the training instance selection strategy of the machine-learned classifier. It points out that non-anaphoric NPs also provide important information for coreference resolution especially when they represent a discourse-new entity with no preceding referent NP even if there would be a full string-matching candidate. Since the baseline system provides no adequate training example in the training text the classifier might fail in such cases. This would subsequently lead to a decrease in the precision score. Therefore, negative labeled training instances consisting of a non-anaphoric NP and an antecedent NP containing the same headword have to be generated and presented to the classifier.

3.2.3 Clustering Approaches

Since coreference resolution denotes the process of resolving markables that refer to the same real world entity it is obvious that this task can also be seen as a clustering or partitioning of the set of markables found in a given text corpus. Literature presents both, unsupervised and supervised approaches that tackle coreference resolution as a clustering task instead of a binary classification problem.

The clustering approach by Cardie and Wagstaff

The idea of gathering coreferent markables in the same cluster was firstly introduced by [Cardie and Wagstaff, 1999]. They state that “intuitively, all of the noun phrases used to describe a specific concept will be ‘near’ or related in some way, i.e. their conceptual ‘distance’ will be small”. Consequently a clustering algorithm requires two things, namely a description for each NP and some kind of method that evaluates the distance between two given NPs. Their unsupervised corpus-based approach consists of two main stages:

1. All existing NPs in the input text are determined and considered as markables. In this step a feature vector containing eleven features is auto-generated for each markable and used to describe the NP during further processing. The vector contains information about the markables head noun, number, gender, sentence position, semantic class, etc. whereas the semantic class information is provided by the lexical database WordNet [Feldbaum, 1998].
2. In the beginning each markable forms its own cluster. The clustering algorithm starts at the end of the text and compares every markable to all proceeding markables. It uses a distance metric in order to compute a distance for a given NP pair by comparing each feature in the feature vector of one markable to the corresponding feature of the other. The results of this process are firstly weighted and then summarized. If the calculation for a certain markable pair results in a distance less than a predefined clustering radius, then their clusters are considered for merging

unless there exist markables in the clusters, which are incompatible. Markable pairs with a distance greater than the clustering radius by contrast, cannot be interpreted coreferent and hence not be merged.

Yang et al.'s clustering approach to coreference resolution

Another solution that attempts coreference resolution as a clustering task is the one presented by [Yang et al., 2004]. Their supervised learning approach uses a NP-Cluster based framework in contrast to the NP-NP based framework presented in the baseline approach by [Soon et al., 2001] to process coreference resolution. They state that a coreferential cluster provides much more information to describe a markable it contains than the single noun phrase itself and consequently this expanded knowledge enhances the resolution capability of the system. Therefore, a classifier is trained to choose the correct cluster instead of the correct antecedent for a given markable. A training instance in this approach consists of three elements, namely the markable under consideration, an existing cluster and a markable that represents the cluster. It is likely that a cluster contains more than one reference markable and thus has numerous associated instances. An instance is represented by a feature vector that contains 24 features. Out of these, 18 features describe the relationship between the markable under consideration and the referent markable and six features describe the relationship between the markable under consideration and the cluster itself. For the classifier learning step an instance is labeled positive if the markable under considerations belongs to the cluster, or negative if not. During training all NPs are processed from the beginning to end. One instances is created for every markable and all of its preceding clusters represented by the last NP they contain. The process does not terminate until the correct cluster is found for every markable. The resolution procedure differs to the one used during training because for each cluster under consideration not only one, but multiple instances, one with every containing markable as the referent markable, are created. For every instance the trained classifier computes the likelihood that the given markable can be linked to the cluster under consideration. The confidence value for one cluster is the maximal confidence value of all of its instances. Clusters judged with a confidence value under 0.5 are filtered out while a certain selection strategy, i.e. "Most Recent First" or "Best First" is applied to select the correct one. The NP-Cluster based framework outperforms a NP-NP-based baseline system which uses the same features, except the one that describe the relation between the markable under consideration and the cluster in both, recall and precision.

3.2.4 Approaches Concerning Bridging Coreference

Bridging coreference is very common in natural text since human authors use this kind of proposition frequently in order to avoid word repetition. The resolution of such coreference relations deserves a special investigation because this task is notable harder than the processing of other coreference types. The lexical relations a resolution system has to face in cases of bridging coreference are [Versley, 2007]:

- ***Synonym***
  The antecedent and the anaphor are synonyms like in "the tumor" – "the cancer"

- ***Hypernym/hyponym***
  The anaphor is a generalization of the antecedent like in “the drug” – “the antibiotic”

In such cases a resolution algorithm has to deal with a larger number of antecedent candidates and can no longer rely strictly on syntactic information like in case of pronominal coreference or surface similarities like in case of same head coreference.

The application of lexical and common sense knowledge

Since bridging coreference cannot be identified through headword equality a more complex form of lexical or common sense knowledge is necessary to detect and resolve this type of coreference. Numerous approaches presented in computational linguistic literature use the lexical database WordNet [Feldbaum, 1998] to look for a synonym or hypernym/hyponym relation.

The clustering algorithm by [Cardie and Wagstaff, 1999] for example applies the WordNet node distance of two markable head nouns as a feature in its distance measure that indicates a possible coreferent relation.

The knowledge-based approach by [Munoz and Palomar, 2001] uses WordNet to derive the semantic classes of all found NPs in order to create a semantic network. This is because a given definite description can only be considered anaphoric if and if only there is a semantically compatible NP prior in the input text. A coreference resolution algorithm is only applied if the headwords of the NP-DD pair under consideration hold a semantic relation.

The semantic class feature in the feature vector of the machine learning system by [Soon et al., 2001] assigns a predefined semantic class to every markable found in the text. For comparison the semantic classes are mapped to WordNet. A possible coreference relation of two markables can be indicated if the two semantic classes are equal or in a parent-child relationship.

The downside and disadvantages of lexical resources

The resolution of bridging coreference, especially hypernym/hyponym relations, is central to text understanding. Unfortunately the application of domain independent lexical resources such as WordNet is limited since they are not available for all languages and they are often very incomplete, especially for more domain specific vocabulary and proper names [Garera and Yarowsky, 2006]. Another downside is the high expense that is necessary to create and maintain such semantic taxonomies.

The use of lexico-syntactic patterns

A solution for this problem that does not require any or just little pre-encoded knowledge is the application of so-called “lexico-syntactic patterns” which are textual constructions that indicate a hypernym/hyponym relation and occur frequently in natural language text across genre boundaries. Several of these linguistic structures were firstly identified by [Hearst, 1992] (NP_0 stands for the more general noun phrase while NP_1 … NP_n represents the more specific expressions):

1. NP_0 such as NP_1 (,NP_2, . . . , and/or NP_n)
2. such NP_0 as NP_1 (,NP_2, . . . , and/or NP_n)

3. NP_1 (,NP_2, . . . , NP_n) or other NP_0
4. NP_1 (,NP_2, . . . , NP_n) and other NP_0
5. NP_0, including NP_1 (,NP_2, . . . , and/or NP_n)
6. NP_0, especially NP_1 (,NP_2, . . . , and/or NP_n)

Next to the six original lexico-syntactic patterns numerous other frequently found textual constructions like the ones by [Snow et al., 2005] have been discovered and published:

1. NP_0 like NP_1
2. NP_0 called NP_1
3. NP_1 is a NP_0
4. NP_1, a NP_0 (appositive)

3.2.5 Approaches Concerning the Medical Domain

In contrast to other genres computational anaphora/coreference resolution systems concerning the medical domain can apply semantic information and structured domain knowledge provided by a huge domain specific lexical resource, the Unified Medical Language System (UMLS) [Humphreys et al., 1998], which is a long-term National Library of Medicine research project that integrates information from multiple biomedical information sources.

All anaphora/coreference resolution algorithms presented below use the UMLS to derive semantic information (UMLS semantic types and Metathesaurus concepts) in order to select the correct antecedent among the set of potential candidates.

Castano et al.'s anaphora resolution approach for biomedical literature

The knowledge-based approach by [Castano et al., 2002] treats coreferential pronominal and sortal (definite description) anaphora in Medline abstracts. Those two types of anaphora are common in biomedical texts, whereas it was found that sortal anaphors are prevalent. The algorithm relies on syntactic features, semantic information, and the textual information by the string itself. During preprocessing a POS tagger is applied in order to identify all NPs, which are subsequently represented by a so-called *Syntactic Chunk Object (SCO)* that contains syntactic features (gender, number…) gained during tagging, semantic type information derived from the UMLS type system and string information. The resolution algorithm itself consists of two main stages, namely anaphor and antecedent recognition. At first all relevant anaphors are selected through a two-step filtering strategy. The first selection is based on syntactic information. Only definite NPs and third person personal, possessive, and reflexive pronouns are considered as potential anaphors. First and second person pronouns are excluded because they are not relevant. The second selection relies on semantic information. Since the approach only handles a certain subset of all possible entities in the corpus, only those candidate anaphors in the predefined biomedical semantic UMLS type are selected. Additionally, the number of antecedent required by each anaphor is identified and stored in the SCO object, because singular anaphors may only refer to one antecedent, while plural anaphors usually point to plural antecedents. The actual coreference resolution is processed during the second stage. Each of the filtered anaphors is resolved by selecting the SCO with the highest salience score as their correct antecedent. An

anaphor is compared with all preceding candidate antecedents starting with the closest from right to left. The initial score for each antecedent candidate is zero. Syntactically, preference is given to antecedent-anaphor pairs with equal number and person. This is especially important for pronominal anaphora. In case of resolution of sortal anaphora, additionally morphological and semantic preferences are applied. Morphological information is used to compute a score of string similarity between the antecedent and anaphor through the application of the *Longest Common Subsequence (LCS)* algorithm [Black, 1999], which denotes the fact that the anaphor and its antecedent are morphological variants of each other. A semantic comparison is made by matching the corresponding semantic types of the antecedent-anaphor pair since each SCO object is likely to hold more than one UMLS type. The more UMLS types the anaphor and antecedent share the higher the salience score. The correct antecedent is finally selected by using the "nearest fit" strategy. If an antecedent candidate reaches a high enough score, no further comparisons are made. In case of a tied maximum salience score preference is given to the closest antecedent candidate. If no SCO pair reaches a predefined minimum score, then the anaphor is most likely global-referring and marked as a global anaphor. In case of multiple antecedents the resolution algorithm determines further antecedents based on a combined salience measure of the anaphor and the first antecedent.

The anaphora resolution approach for biomedical literature by Lin and Liang

[Lin and Liang, 2004] present an improvement for the approach by [Castano et al., 2002]. Instead of selecting the closest antecedent candidate that reaches a certain salience score ("nearest fit"), their algorithm applies a "best fit" selection strategy. The candidate with the highest overall salience value is chosen as the correct antecedent. The features used to select the correct antecedent are very similar to the baseline system. They also apply syntactic information (number agreement checking), semantic knowledge (UMLS type checking) and morphological preferences (Longest Common Subsequence). The approach furthermore presents some rules to filter out pleonastic it instances:

1. It be [Adj/Adv/verb]* that
2. It be Adj [for NP] to verb
3. It [seems/appears/means/follows] [that]*
4. NP [makes/finds/take] it [Adj]* [for NP]* [to verb]

The anaphora resolution approach by Torii and Vijay-Shanker

A similar algorithm was presented by [Torii and Vijay-Shanker, 2007]. Their machine learning approach aims to resolute sortal (definite and demonstrative) anaphora in Medline abstracts since as they state this is the most frequently occurring type of anaphoric expressions in the biology domain. At first a parser is used to obtain all markables NPs from the text. Among the markables only NPs with a definite ("the") or demonstrative ("this" and "these") article are considered as potential anaphor for the purpose of reference resolution. The resolution algorithm selects the correct antecedent for a given anaphor out of all antecedent candidates where candidates are all preceding markables found in the text. For each candidate a likelihood of being the correct antecedent is calculated using a set of weighted features whereas the antecedent candidate with the highest value is selected. The system applies a number agreement feature to check if the antecedent candidate anaphor pair

agrees in number, a distance feature that measures the number of sentences between the anaphor and the antecedent candidate and a semantic type match feature with semantic knowledge information provided by the UMLS. While the number of terms in this knowledge base is extensive, the authors state that there were still nouns in the input texts that could not be found in the UMLS dictionary. In such cases the algorithm uses so-called name-internal features in order to extend the coverage of the dictionary. If a certain term cannot be found in the UMLS the dictionary is searched for terms containing the same headword as the current one starting with terms that share the two rightmost tokens. If there is still no satisfying result the search string is reduced to the single rightmost token. Additionally, the algorithm applies features that indicate if a NP markable is in a subject position of either a clause or a sentence or if the NP is within a prepositional phrase attached to the subject NP. The system also uses several string match features, namely a common head feature to compare headword equality, a common string feature to compare words except of the heads, a common suffix feature to show if the headword of one NP is subsumed by that of the other and a common phrase feature to point out if a headword is rephrased. The last set of features presented in this approach addresses textual patterns that can be frequently found in scientific texts. One of these patterns concerns acronyms where the short form is put in parentheses and succeeds the long form of the NP. For each occurrence of the acronym in the text the corresponding acronym-feature is set. Similar features are applied for NPs found in other patterns that introduce new entities in the discourse such as appositive constructions and phrases like “named (as)” and “called (as).

Rindflesch and Fiszman’s hypernymic proposition interpretation approach

[Rindflesch and Fiszman, 2003] present a hypernymic proposition interpreter for biomedical text. Their system uses syntactic analysis and structured domain knowledge from the UMLS to identify and interpret hypernym/hyponym (“IS_A”) relations in Medline abstracts. Furthermore, the knowledge-based application MetaMap [Aronson, 2001] is used to find the best mapping between the text of a NP and a concept in the UMLS Metathesaurus [USNLM, 2008]. The Metathesaurus is a large medical vocabulary that groups equivalent terms to unique concepts and additionally provides associative and hierarchical relationships between them. Each Metathesaurus concept is also assigned to one or more semantic UMLS types that categorize the numerous concepts. The system relies on syntactic analysis in order to detect potential hypernym/hyponym relations. The authors identified three major patterns that encode a hypernymic proposition in English natural language texts:

1. The specific NP is subject of the verb “be” and the general NP is represented by its complement. Other verbs such as “remains” are also possible.

2. Two NPs occur next to each other and they are separated by commas or parentheses. The NPs can also be linked by lexical items like “such as”, “including”, “especially”, and “particularly” (see [Hearst, 1992]).

3. The hypernym and the hyponym term occur in the same NP. One concept represents the head noun while the other one serves as a modifier.

After tokenization a POS tagger is used to obtain all NPs from the input text. Then the text is searched for syntactic structures that potentially indicate hypernym/hyponym relations and the involved NPs are identified. Each NP is given a partial internal analysis in order to identify

its headword and modifiers and subsequently augmented with Metathesaurus concepts and semantic UMLS types. The concepts are then subjected to semantic validation. The system firstly matches the concepts and checks if they occur in the same semantic group. To finally proof a hypernym/hyponym relation the concepts themselves have to be in a hierarchical relationship in the Metathesaurus.

3.3 Discussion

In the previous sections some existing coreference resolution approaches were presented. The outline included domain independent systems as well as approaches designed for the medical domain. Furthermore, a distinction was made between knowledge-based and machine-learning algorithms.

Although the majority of coreference resolution approaches presented over the last decade use machine-learning algorithms, systems concerning the medical domain still rely mainly on knowledge-based algorithms. This is due to the lack of existing domain depended coreferentially annotated corpora and the specific knowledge that is necessary to operate with a high precision in this domain.

Most of the approaches obtain syntactic information and semantic domain knowledge from a single information source, the UMLS. With the help of the background knowledge provided by the UMLS, the resolution algorithms apply constraints and preferences or a set of features in combination with a salience measure to denote the likeliness of a coreferent relation between an anaphor and an antecedent candidate.

4 TOOLS AND KNOWLEDGE SOURCES

This section mainly presents the supporting tools and knowledge sources that our coreference resolution algorithm uses to fulfill its task.

As stated above, the process of coreference resolution can be basically divided into two main stages:

1. Markable determination
2. Correct antecedent candidate selection

The first step denotes the detection of all possible phrases within a given text corpora that could be part of a coreferential relation. Syntactic information is necessary to identify related words in the input text and to combine them ton non-overlapping phrases. Among all phrases in a given input text only noun phrases and pronominal phrases can be coreferent. As one can see, again, syntactic knowledge is required to select those markables.

The execution of the second step depends on syntactic information, especially for the resolution of pronominal coreference, as well as on sophisticated domain dependent semantic knowledge. This type of information is essential in order to resolve bridging or indirect coreference, since the application of this knowledge is the only way to find a relationship between the anaphor and the corresponding correct antecedent candidate.

Our coreference resolution approach obtains the required syntactic information from MetaMap [Aronson, 2001], or to be more specific from the MetaMap Transfer (MMTx) program [Divita, 2005]. The semantic knowledge is provided by a huge domain specific lexical resource, the Unified Medical Language System (UMLS) [Humphreys et al., 1998]. Both technologies are presented in the following sections.

4.1 Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS) [Humphreys et al., 1998] is a large (bio)medical knowledge base that integrates information derived from various other machine-readable (bio)medical information sources such as databases, dictionaries, specialized vocabularies, and ontologies. The UMLS was initiated in 1986 by the United States National Library of Medicine (NLM)[2] with the final goal to give an almost complete picture of the current existing (bio)medical knowledge. The structured information provided by the UMLS can be used in various fields, for example as described here in research in natural language processing. The UMLS consists of three main knowledge sources that provide different types of information as illustrated in Figure 2. Those resources are [USNLM, 2008], [UMLS, 2006]:

- Metathesaurus
- Semantic Network
- Specialist Lexicon

[2] http://www.nlm.nih.gov/ (last assessed: February 19, 2009)

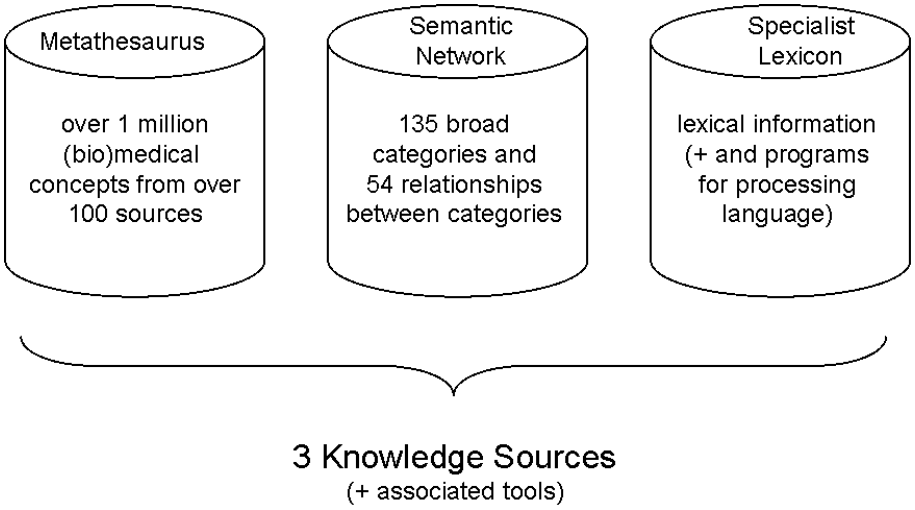

Figure 2: UMLS schematic illustration **[UMLS, 2006]**

4.1.1 The Metathesaurus

The UMLS Metathesaurus is a large repository that consists of more than one million (bio)medical and health related concepts derived from more than 100 (bio)medical information sources like vocabularies, classifications, and coding systems which are used in a variety of purposes and settings such as research, clinical, administrative, or public health reporting. Those various information sources differ in complexity, concept order, notation, terminology, and language. The Metathesaurus integrates the different concept orders in one large common hierarchical data structure and clusters their terms by meaning into unique concepts. These concepts form the organizational core of the Metathesaurus. In other words, all synonyms, views and alternative names of the same concept from the different terminologies are linked together and a concept unique identifier (CUI) is assigned. Furthermore, additional information for each concept such as specific definitions, various attributes, and translations in other languages (up to seventeen) can be found within the Metathesaurus [USNLM, 2008], [UMLS, 2006].

Figure 3 illustrates the composition of the Metathesaurus concept "Addison's disease" from the several information sources.

term	source	term type	source id
Addison's disease	Metathesaurus	PN	
Addison's disease	SNOMED CT	PT	363732003
Addison's Disease	MedlinePlus	PT	T1233
Addison Disease	MeSH	PT	D000224
Bronzed disease	SNOMED Intl	SY	DB-70620
Primary Adrenal Insufficiency	MeSH	EN	D000224

concept	
C0001403	Addison's disease

Figure 3: Metathesaurus Concept **[UMLS, 2006]**

In addition to the concepts itself useful associative and hierarchical relationships between the concepts are also represented. There are several different types of relations that exist within the Metathesaurus (see Table 1), which either come from the source terminologies or are added by editors from the NLM [USNLM, 2008].

Table 1: The relationship types of the UMLS Metathesaurus **[USNLM, 2008]**

Code	Description
AQ	allowed qualifier
CHD	has child relationship in a Metathesaurus source vocabulary
DEL	deleted concept
PAR	has parent relationship in a Metathesaurus source vocabulary
QB	can be qualified by
RB	has a broader relationship
RL	the relationship is similar or "alike"
RN	has a narrower relationship
RO	has relationship other than synonymous, narrower, or broader
RQ	related and possibly synonymous
RU	related, unspecified
SIB	has sibling relationship in a Metathesaurus source vocabulary
SUBX	concept removed from current subset
SY	source asserted synonymy
XR	not related

4.1.2 The Semantic Network

The Semantic Network aims to categorize the concepts in the (bio)medical domain. Therefore, each Metathesaurus concept is assigned to one or more semantic types. The network defines 135 of such broad subject categories like “Disease or Syndrome” or “Pharmacologic Substance”. Additionally, there are several coarse-grained aggregates of semantic types, so-called semantic groups, such as “Chemical and Drugs” or “Anatomy”. The semantic types are linked together with the help of 54 semantic relations such as “prevents”, “location of” or “affects”.

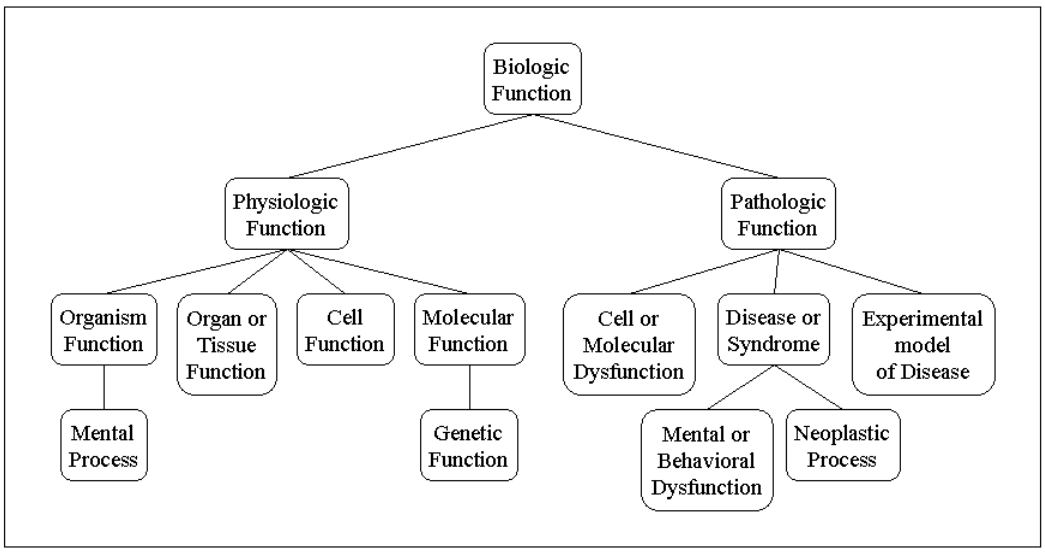

Figure 4: Semantic Type Hierarchy [USNLM, 2008]

Both, the semantic types and the semantic relations have a hierarchical structure. For example Figure 4 shows that the semantic type “Disease or Syndrome” is more specific than “Biological Function”. Furthermore, Figure 5 illustrates that the semantic relation “affects” is more general than “prevents” [USNLM, 2008].

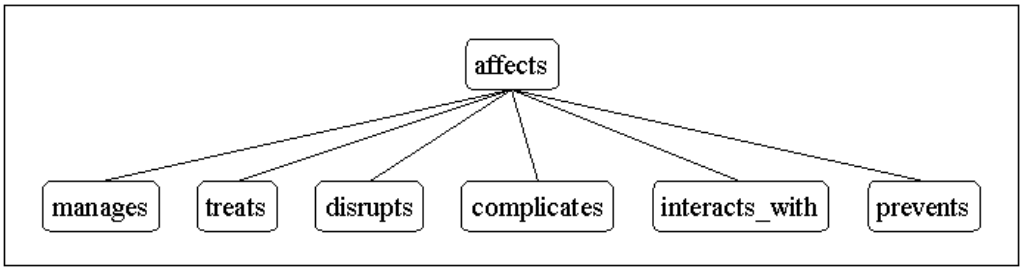

Figure 5: Semantic Relation Hierarchy **[USNLM, 2008]**

The Semantic network itself is organized as a single-inheritance hierarchy, where the semantic types represent the nodes and the semantic relations the links between them. Single-inheritance denotes that every semantic type holds certain hierarchical relations, one to its parent and one to every child. Additionally, the network defines various useful associative relations between the semantic types which represent valid (bio)medical knowledge. Since those relationships only link the semantic types, the do not necessarily hold at the Metathesaurus concept level [USNLM, 2008].

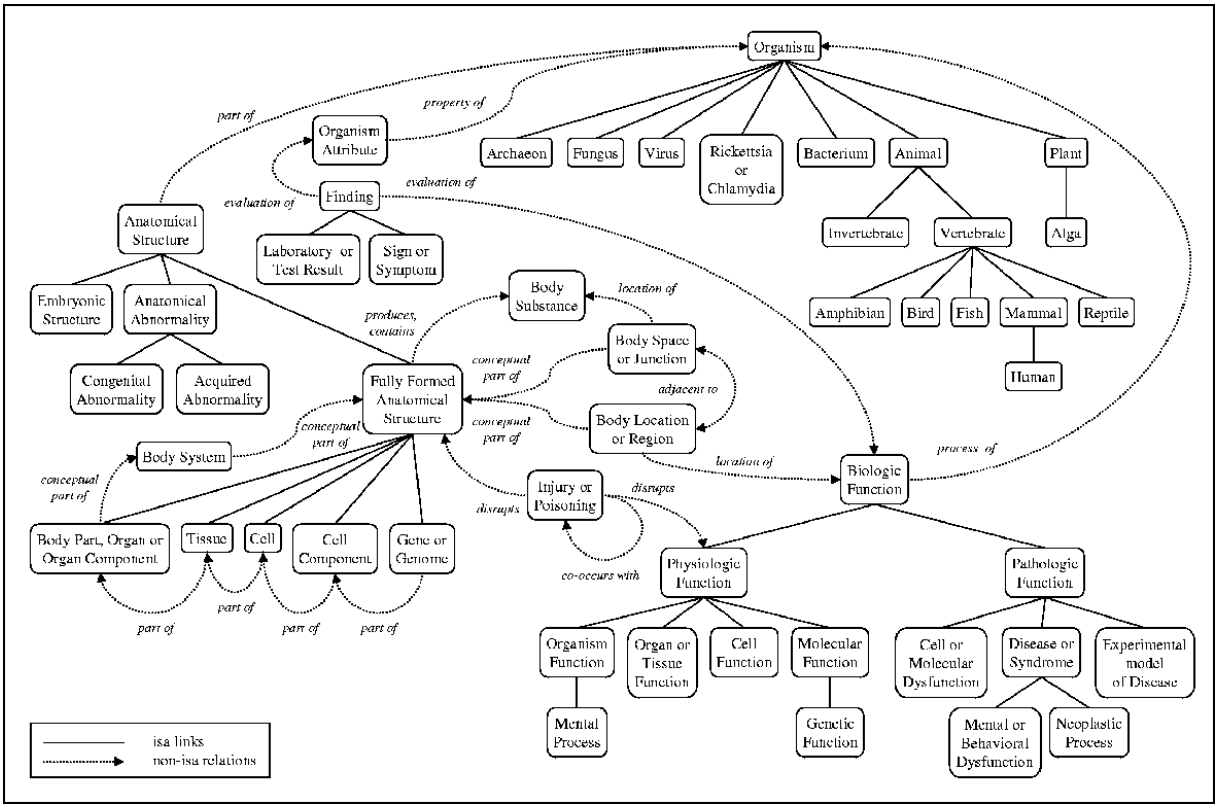

Figure 6: Part of the Semantic Network **[USNLM, 2008]**

4.1.3 The Specialist Lexicon & Specialist NLP Tools

The Specialist Lexicon is a syntactic English lexicon that includes common words as well as (bio)medical terms. It consists of several relational files that contain the following lexical information [UMLS, 2006]:

- Syntax (how words are put together)
- Morphology (inflection, derivation, and compounding)
- Orthography (spelling)

As an example, Figure 7 shows the two unit lexical records the Specialist Lexicon provides for the query string “anesthetic”. The first set includes the information for the noun form of the word and the second for the adjectival form. A unit lexical record consists of slots and filters (<slot> = <filter>). The term in front of the equals sign denotes the slot and the term behind the filter information.

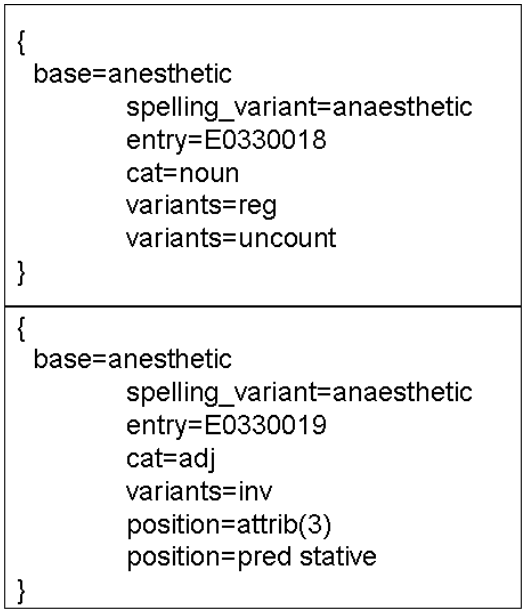

Figure 7: Unit lexical records from the Specialist Lexicon for the entry "anesthetic" **[USNLM, 2008]**.

Every record has a "base" slot that indicates the base form of the entry. Optionally a record can have a set of spelling variants identified by the "spelling_variant" slot. Additionally, there is always an "entry" slot, whose filter denotes the entry unique identifier (EUI), and a "cat" slot that holds the syntactic category or part of speech (POS) information (see Table 2).

Table 2: The syntactic categories in the Specialist Lexicon **[USNLM, 2008]**

Code	Syntactic Category
noun	nouns
adj	adjectives
adv	adverbs
pron	pronouns
verb	verbs
det	determiners
prep	prepositions
conj	conjunctions
aux	auxiliaries
modal	modals
compl	complementizers

The "variants" slot contains the inflectional types of the lexical entry. There are several different inflection types depending on the syntactic category (nouns, verbs, pronouns, adjectives and determiners). This slot is especially important for pronouns since it provides person (singular and plural) and number (first, second, and third) information for the coreference resolution process.

Additionally, there is a "position" slot in the adjectival entry. It denotes if an adjective is attributive, post modifying or predicative. Furthermore, the Specialist Lexicon provides information about the modification types of adverbs, and various features of terms in various categories [USNLM, 2008].

In addition to the Specialist Lexicon, the United States National Library of Medicine provides several so-called Specialist NLP Tools that allow developers to gain more sophisticated

syntactic, morphological and orthographic information within their applications [USNLM, 2008]:

- **Lexical tools** deal with all kinds of variations like lexical variant generation, word normalization and inflections.
- **Text tools** are designed to analyze free input text. They are able to tokenize input strings into sections, sentences, phrases, terms, and words.
- **Spelling tools** offer the possibility to correct misspelled words by finding orthographically related or close terms.

4.2 MetaMap

MetaMap [Aronson, 2001] tackles the task of "mapping (bio)medical text to concepts in the UMLS Metathesaurus, or equivalently, to find UMLS Metathesaurus concepts in (bio)medical text" [UMLS, 2006].

The MetaMap algorithm consists of five steps. At first the input text has to be parsed and tokenized to a phrase level. At next, lexical variants are generated for each phrase. The UMLS Metathesaurus gets consulted in the third step in order to retrieve candidates that match the generated variants. Subsequently the best matching candidates are evaluated via a mapping algorithm. Finally the candidate(s) with the highest mapping score(s) are returned [Aronson, 2001].

Figure 8 shows a sample mapping for an input text that includes four relevant phrases.

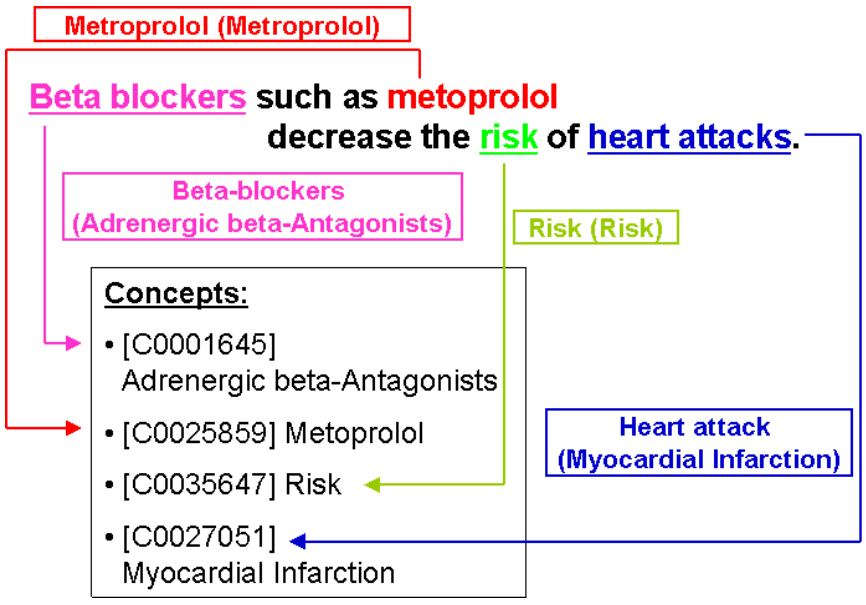

Figure 8: Mapping example

4.2.1 Parsing

In this stage the input text is subject to a shallow syntactic analysis. This task is performed using the Specialist minimal commitment parser [McCray et al., 1994]. The parser uses the Xerox part-of-speech (POS) tagger [Cutting et al., 1992] in order to assign POS labels to

words with ambiguous syntactic category tags in the Specialist Lexicon (like “anesthetic” in the above example). Furthermore, the NPs are given a partial internal analysis with the purpose to identify their headwords, which are the most central parts of the phrase [Aronson, 2001].

Table 3 shows how the output of the parsing stage would look like:

Table 3: NPs, syntactic tags, and headwords for the mapping example.

Phrase	Syntactic Tags	Headword
Beta blockers	noun (Beta blockers)	Beta blockers
such as metoprolol	prep (such as) noun (metoprolol)	metoprolol
the risk	det (the) noun (risk)	risk
of heart attacks	prep (of) noun (heart attacks)	heart attacks

4.2.2 Variant Generation

During this step lexical variants are generated for any meaningful subsequence of words found in the determinate NPs. A meaningful subsequence is either a single word or a combination of words that occur in the Specialist lexicon. Words with a syntactic category such as pronouns, determiners, prepositions, conjunctions, auxiliaries, modals, and complementizers are ignored. Each meaningful subsequence is a source for the generation of variants and therefore called variant generator. A set of variants consists of the variant generator itself and all of its acronyms, abbreviations, synonyms, derivational variants, meaningful combinations of these, and finally inflectional and spelling variants (see Figure 9) [Aronson, 2001].

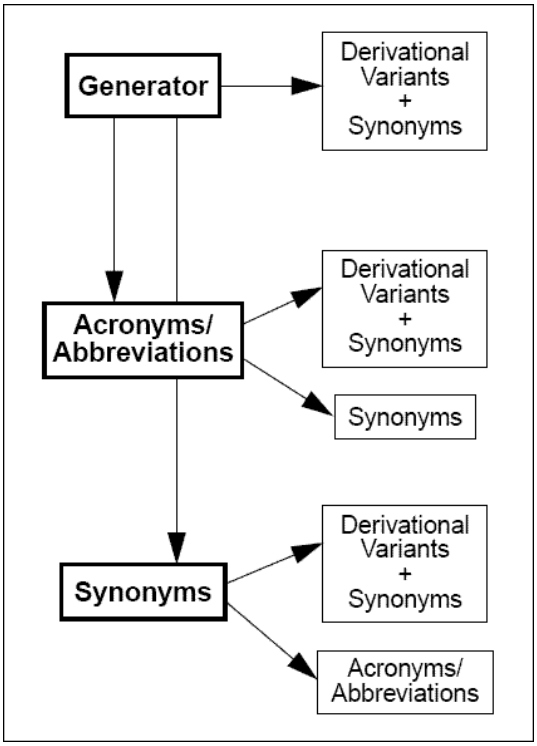

Figure 9: MetaMap variant generation [Aaronson, 2001]

For every variant, the distance from its generator is computed. There are two different views (see Table 4):

1. Numerical variation value
2. Variant creation history

The distance value estimates how much the variant differs from the original string in the phrase. It is the summarization of all distance values for each step taken in the variant generation process. The creation history is a string composed of the different distance labels [Aronson, 2006].

Table 4: Variant distance and labels [Aaronson, 2006]

Variant type	Distance value	Distance label
spelling	0	p
inflectional	1	i
synonym	2	s
acronym/abbreviation	2	a
derivation	3	d

Table 5 shows how the generated variants would look like:

Table 5: Variant generators and variants for the mapping example

Phrase	Variant generators	Variants
Beta blockers	Beta blockers	betablocker [noun, 1="i"] beta blockers [noun, 0=""]
	Beta	beta [noun, 0=""] betas [noun, 1="i"]
	blockers	blockers [noun, 0=""] blocker [noun, 1="i"]
such as metoprolol	metoprolol	metoprolol [noun, 0=""]
the risk	risk	risky [adj, 3="d"] risk [noun, 0=""]
of heart attacks	heart attacks	heart attacks [noun, 0=""]
	Heart	hearts [noun, 1="i"] heart [noun, 0=""]
	Attacks	attacks [noun, 0=""] attack [noun, 1="i"]

4.2.3 Candidate Retrieval

The set of candidates for a given NP consists of all Metathesaurus strings containing at least one of the generated variants in the previous step. When a string itself is not the preferred name for the Metathesaurus concept, the preferred name appears in parenthesis following the string. By default, candidate concepts with an overmatch or a concept gap are filtered out before the evaluation stage. An overmatch occurs when a concept candidate has a non-matching word on either the front or the back end of its string. In case of a concept gap, the non-matching word occurs in the middle of the concept candidate string [Aronson, 2006].

Table 6 shows how the retrieved candidates would look like:

Table 6: Retrieved candidates of the mapping example

Phrase	Candidates
Beta blockers	C0001645:Beta Blockers (Adrenergic beta-Antagonists) C0330390:Beta (Beta plant) C0439096:Beta (Beta greek letter) C1552649:beta (Probability Distribution Type - beta)
such as metoprolol	C0025859:Metoprolol
the risk	C0035647:Risk
of heart attacks	C0027051:Heart Attacks (Myocardial Infarction) C0018787:Heart C1281570:Heart (Entire heart) C0277793:Attack, NOS (Onset of illness)

	C0699795:Attack (Attack device) C1261512:attack (Attack behavior) C1304680:Attack (Observation of attack)

4.2.4 Candidate Evaluation

In this step a linguistically principled evaluation function measures the quality of the match between an input phrase string and a Metathesaurus string. The calculation consists of a weighted average of four metrics: centrality, variation, coverage, and cohesiveness, whereas each component represents a value between zero (weakest match) and one (strongest match). The centrality metric shows the involvement of the headword of the phrase in the Metathesaurus candidate. The variation value estimates the difference of the variants in the Metathesaurus string and the corresponding words in the phrase. Coverage and cohesiveness measure how much of a Metathesaurus candidate matches the input text and in how many pieces. The result of the evaluation process is a final mapping score for a concept and a phrase with a value between 0 (no match) and 1000 (perfect match) [Aronson, 2001][Aronson, 2006].

Mapping Score = (Centrality + Variation + 2 x Coverage + 2 x Cohesiveness) / 6

Table 7 shows how the derived mapping scores would look like:

Table 7: Mapping scores for the mapping example

Phrase	Candidates & Mapping scores
Beta blockers	[1000] Beta Blockers (Adrenergic beta-Antagonists) [861] Beta (Beta plant) [861] Beta (Beta greek letter) [861] beta (Probability Distribution Type - beta)
such as metoprolol	[1000] Metoprolol
the risk	[1000] Risk
of heart attacks	[1000] Heart Attacks (Myocardial Infarction) [861] Heart [861] Heart (Entire heart) [827] Attack, NOS (Onset of illness) [827] Attack (Attack device) [827] attack (Attack behavior) [827] Attack (Observation of attack)

4.2.5 Mapping Construction

The construction of final mappings is the last step in the whole process. Various combinations of Metathesaurus candidates, which participate in matches with disjoint parts of the NP are examined. The strength of the complete mappings is computed with the

evaluation function as in the previous step for candidate mappings. The complete mappings that reach the highest score form the final mappings, which represent MetaMap's best interpretation of the original phrase. Even though, the algorithm tries to find the best mapping, it is possible that more than one concept reaches the highest mapping score. In such a case the final mapping is ambiguous and the final decision has to be made manually [Aronson, 2001][Aronson, 2006].

Table 8 shows how the final mappings would look like:

Table 8: Final mappings for the mapping example

Phrase	Candidate	Semantic Type
Beta blockers	[1000] Beta Blockers (Adrenergic beta-Antagonists)	Pharmacologic Substance
such as metoprolol	[1000] Metoprolol	Organic Chemical, Pharmacologic Substance
the risk	[1000] Risk	Qualitative Concept
of heart attacks	[1000] Heart Attacks (Myocardial Infarction)	Disease or Syndrome

5 The CPG Coreference Resolution Algorithm

This chapter mainly describes the way our coreference resolution algorithm was developed.

The approach aims to detect and resolute specific linguistic propositions, more precisely coreference relations, in clinical practice guideline text using underspecified syntactic analysis and structured domain knowledge.

Our algorithm relies heavily on the Unified Medical Language System (UMLS) that supplies the required domain dependent semantic information via its large (bio)medical repository, the UMLS Metathesaurus, and its Semantic Network. Additionally, syntactic information can be obtained from the UMLS Specialist Lexicon.

Furthermore, we extensively use the functionality provided by the MetaMap Transfer (MMTx) program that allows an analysis of the input text on a syntactic level and the mapping of (bio)medical text to UMLS Metathesaurus concepts.

The following sections give a short introduction into the MMTx program followed by a profound presentation of our CPG coreference resolution algorithm.

5.1 MetaMap Transfer (MMTx)

MMTx is the distributable version of the original MetaMap. It uses the same algorithm as its archetype. In contrast to the original MetaMap, MMtx is written and distributed in the Java programming language with the purpose to address a larger number of developers and end users. MMTx provides an application programming interface (API) in order to embed it into other applications.

“MMTx maps text to UMLS Metathesaurus concepts. As part of this mapping process, MMTx tokenizes text into sections, sentences, phrases, terms, and words. MMTx maps the noun phrases of the text to the best matching UMLS concept or set of concepts that best cover each phrase.” [Divita, 2005]

As described above, the first step of the MetaMap algorithm includes the tokenization and parsing of the input text. The MMTx_API textfeature package provides container classes that allow a horizontal splitting of the input text into several levels with different granularity (see also Figure 10). The following listing presents the most important containers starting with the lowest to the highest granularity [Divita, 2005]:

- **Token**
  A token or word is the smallest element of a text with a semantic meaning. It is delimited by a preceding and succeeding white space.

- **Lexical Element**
  A lexical element represents an entry in the Specialist lexicon. It can consist of one (“heart”) or more (“heart attack”) words.

- **Phrase**
  Syntactically related words within a sentence form a phrase. There are several different types of phrases, whereas the noun phrase is the most important one for

the task of coreference resolution. It consists of a headword preceded and succeeded by modifiers such as adjectives or determiners.

- **Sentence**
  A sentence is an arrangement of semantically connected words that is delimited by punctuations such as period, question mark, exclamation mark, and semicolon.

- **Section**
  A section consists of several related sentences. Sections can be determined by the text structure, because a paragraph usually represents a section.

- **Document**
  The document class has the highest granularity. It is the container for the whole input text and consists of all of its sections.

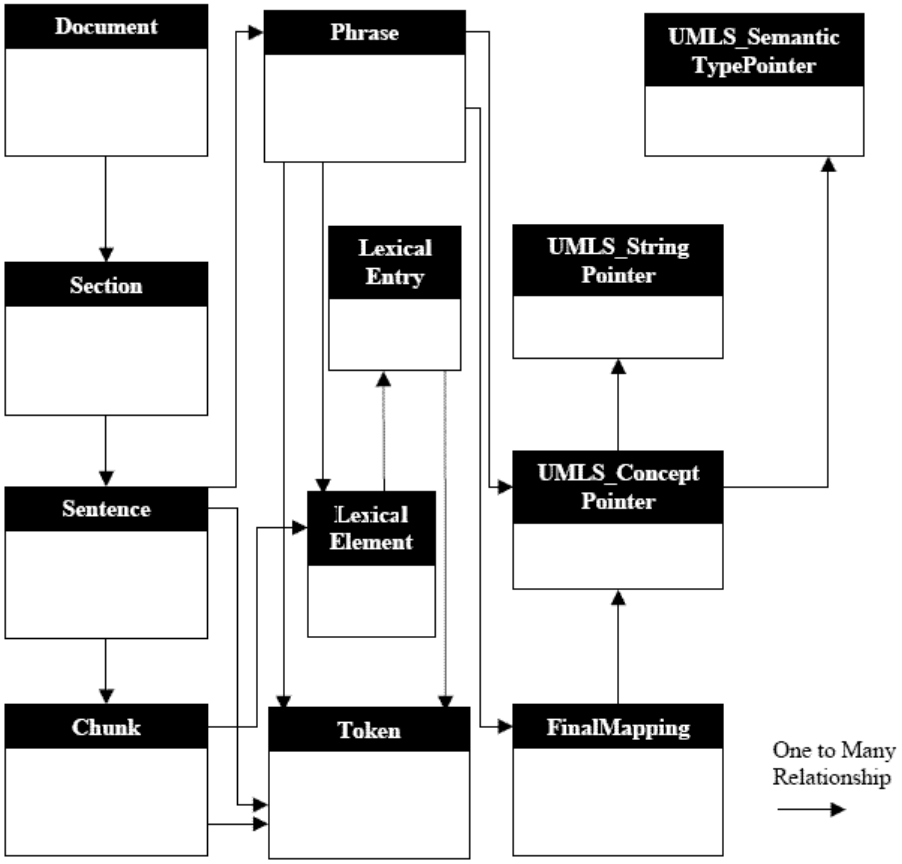

Figure 10: Entity relationship diagram for the textfeature package **[Divita, 2005]**

5.2 The Coreference Resolution Algorithm

Our CPG coreference resolution algorithm aims to determine coreferent relations that hold between textual elements in (bio)medical texts. We focus on the resolution of definite/demonstrative description coreference especially when encountered in an acronym or hypernym/hyponym relation since this type is prevalent in this type of text.

The resolution strategy of our algorithm relies on different types of background knowledge (morphological, syntactic, and semantic information). We use this knowledge in order to gather information that in combination with predefined resolution rules help to determine a possible coreference relation between two markables.

Our knowledge-based approach can be basically divided into three main modules:

1. **Phrase detection**
   The input text gets tokenized and parsed in order to identify all existing phrases. All noun phrases and prepositional phrases that get determined in the input text are mapped to the best matching UMLS concept or set of concepts.

2. **Relevant markable determination**
   All existing phrases get searched through in order to determine relevant phrases (markable candidates) for the actual coreference resolution task. Semantic information is incorporated in order to compute the relevancy of a markable candidate. All relevant markables identified among all markable candidates subsequently serve as anaphor and antecedent candidates for a possible coreferent relation.

3. **Coreference resolution**
   Each of the relevant markables serves as a potential anaphor. All preceding markables in the text are considered as candidate antecedents. A set of predefined coreference resolution rules is applied to each anaphor – candidate antecedent pair in order to denote a coreferent relation between these two markables. The necessary semantic information to perform this task is provided by the MMTx program respectively the UMLS including its Metathesaurus and Semantic Network.

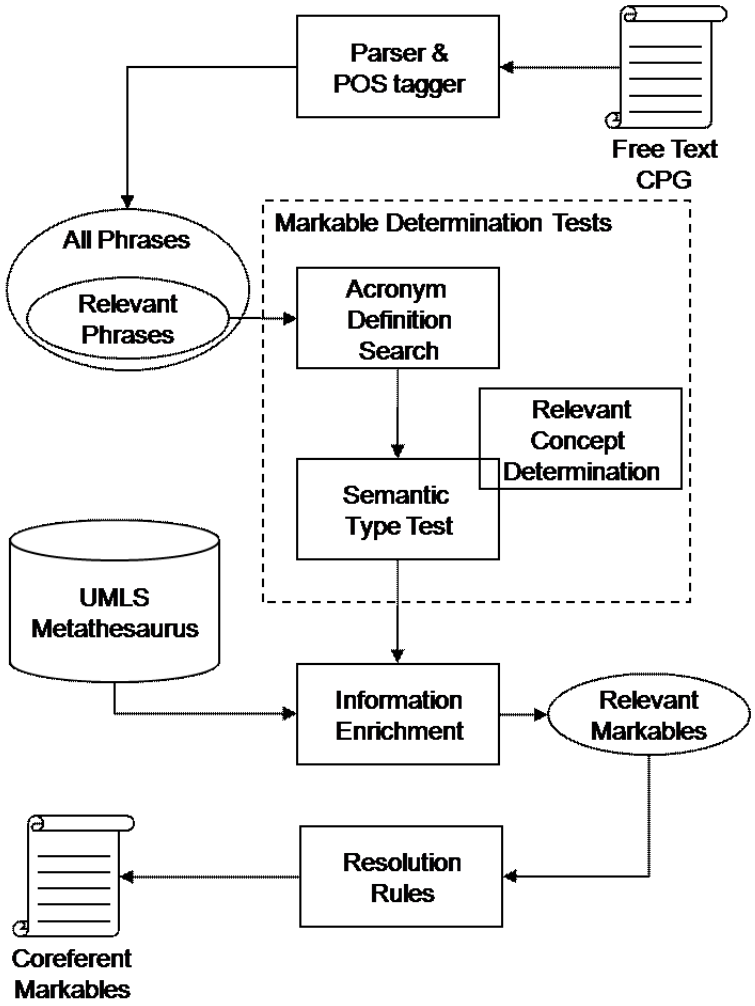

Figure 11: Schematically illustration of the coreference resolution algorithm

The three modules form a NLP chain and therefore perform one after another (as illustrated in Figure 11). The output from the first module is the input for the second and its output the input for the final third module.

5.2.1 Phrase Detection

In our coreference resolution algorithm the functionality provided by the MMTx program is used to process the input text to the phrase level and to map the identified noun phrases and prepositional phrases to the best matching UMLS concepts. This process is schematically illustrated in Figure 12.

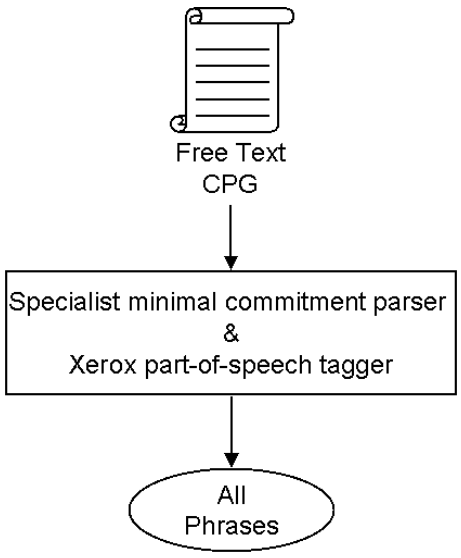

Figure 12: Schematically illustration of the phrase detection step

After tokenization the input text is submitted to a specialist minimal commitment parser [McCray et al., 1994] that relies on the syntactic information in the UMLS Specialist Lexicon. Part-of-speech ambiguities are resolved using the Xerox part-of-speech tagger [Cutting et al., 1992]. Each word found during tokenization gets labeled with its corresponding syntactic (part of speech) category. Finally, syntactically related words identified in the input text are combined to non-overlapping phrases.

MMTx distinguishes between several different types of phrases depending on the syntactic categories of the words they contain:

- **Noun phrase**
  "A phrase whose head is a noun (or a pronoun), optionally accompanied by a set of modifiers. Functionally, a noun phrase may be defined as any category which can bear some grammatical relation within a sentence, such as subject, direct object, indirect object or oblique object." [Trask, 1993].

- **Prepositional phrase**
  "A phrase consisting of a preposition and a noun phrase serving as its object." [Trask, 1993]. The MMTx program distinguishes between general and three specific prepositional phrases that either start with the preposition "by", "of", or "to".

- **Adjective phrase**
  "A phrase with an adjective as its head. Adjectival phrases may occur as pre- or postmodifiers to a noun, or as predicatives (predicate adjectives) to a verb." [Trask, 1993].

- **Adverb phrase**
  "A linguistic term for a single adverb or a group of more than one word operating adverbially, when viewed in terms of their syntactic function. An adverbial phrase can modify a verb phrase, an adjectival phrase or an entire clause." [Trask, 1993].

- **Verb phrase**
  "A phrase composed of the predicative elements of a sentence. It functions in

providing information about the subject of the sentence." [Trask, 1993]. The MMTx program distinguishes between general and two specific verb phrases that either contains a form of the verb "be" or "have".

- **Conjunction phrase**
  A phrase that consists of a conjunction.

- **Unknown phrase**
  A phrase that cannot be assigned to one of the above types.

During the phrase detection stage the identified phrases are grouped in a three-dimensional list. The first dimension is defined by the sections in a given CPG document, the second dimension is defined by the sentences inside a section, and the third dimension is defined by the actual phrases.

As specified above, the main goal of MMTx is to map (bio)medical text to UMLS Metathesaurus concepts. Therefore, all noun phrases and prepositional phrases identified in the input text are mapped to the best matching concept or set of concepts. A Metathesaurus concept is the union set of all synonyms, views and alternative names of the same (bio)medical concept derived from all of the different terminologies that form the UMLS Metathesaurus. Each concept is identified by a concept unique identifier (CUI).

Those concepts play an important role in the following markable determination and coreference resolution stages.

5.2.2 Relevant Markable Determination

Among all the different types of phrases the MMTx program is capable to detect in a natural language text, only some are important for our coreference resolution algorithm and therefore considered possible relevant or markable candidates. The main focus lies on noun phrases and prepositional phrases since they serve as subject or object in a sentence. Following this, the set of all noun and prepositional phrases determined by the MMTx program in a given CPG text is considered markable candidates for our coreference resolution algorithm.

The markable determination stage (see Figure 13) of our coreference resolution approach takes every member of the markable candidate set as an input in order to identify the relevant markables. Therefore, each candidate markable is subject to an exact and deeper internal analysis that consists of several tests. If a candidate markable passes a test, additional semantic and syntactic information required for the following coreference resolution step is derived from the UMLS Metathesaurus. Both, the markable determination tests and the information enrichment step will be explained in the following sections. Furthermore, the headword as well as possible existing modifiers (nouns and adjectives) of the noun or prepositional phrase is extracted.

The output of the markable determination stage is the set of relevant markables, enriched with semantic and syntactic information, that serve as anaphor and antecedent candidates in a possible coreferent relation.

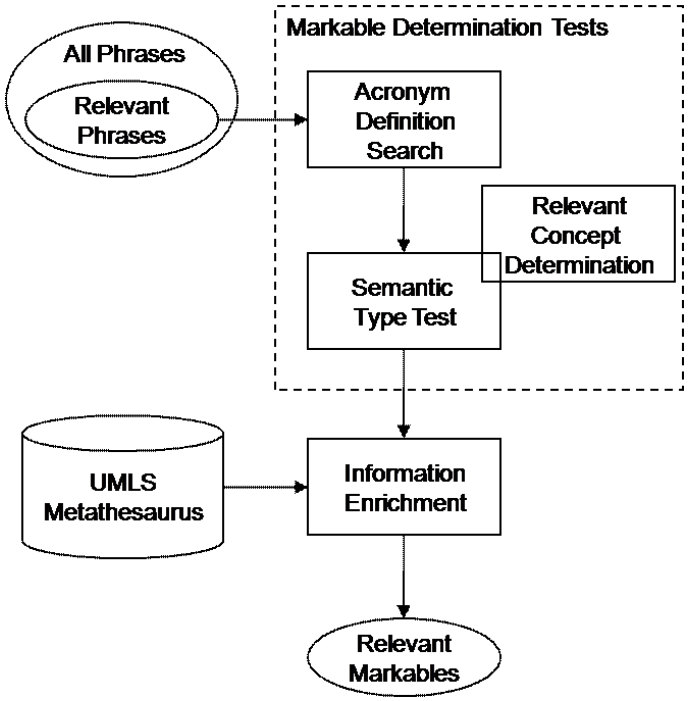

Figure 13: Schematically illustration of the markable determination step.

Acronym Definition Search

This test is used in order to search for acronym definitions in an input text. An acronym is defined as the abbreviation formed mostly using the initial components in a phrase. Acronyms are frequently used in clinical documents in order to shorten medical terms. Usually, the long form of the medical term immediately introduces its acronym that is put in parentheses at the first appearance in the text. From there on only the acronym form is used by the author.

In order to detect acronyms in the input text our algorithm searches for textual elements put in parentheses. If a parenthesized text is found the direct prior phrase is regarded as the reference expression and the acronym determination test is performed.

In the following example from one of our CPG target texts our algorithm detects two occurrences of acronyms "performance status" – "PS" and "Eastern Cooperative Oncology Group" – "ECOG":

> "When selecting patients for systemic chemotherapy, performance status (PS) at the time of diagnosis should be used because it is a consistent prognostic factor for survival. Patients with a PS of Eastern Cooperative Oncology Group (ECOG) 0 or 1 should be offered chemotherapy."

While analyzing the input text the acronym determination step detects the textual element "(PS)" put in parentheses. Following the acronym resolution rule the direct prior phrase "performance status" is regarded as the reference expression. In order to confirm the existence of a valid acronym a string containing the initial letters of the words in the referring expression is build and compared with the supposed acronym expression. If both strings match the found acronym is saved in an acronym list for future comparisons. The next time the expression "PS" appears in the text this acronym list is consulted in order to

distinguish that this phrase is connected to an acronym definition somewhere prior in the text.

Relevant Concept Determination

As mentioned above MMTx maps (bio)medical text to the best matching concept or set of concepts. In the majority of cases however, MMTx does not provide only one, but several Metathesaurus concepts for one phrase. If so, it is important to dictate an unambiguous semantic representation for the further processing of the markable candidates. Therefore, the relevant concept of each markable has to be determined firstly.

Since the headword is usually the most important part of a phrase, a concept concerning the headword is most likely the relevant one. If there is more than one concept concerning the headword, further processing is required.

For the following markable phrase

> "… the fourth most frequently diagnosed cancer ..."

MMTx provides the following 5 sets of Metathesaurus concepts:

(1) fourth	C0205438 Fourth
(2) most	C0205393 Most
(3) frequently	C0332183 Frequent
(4) diagnosed	C0011900 Diagnosis
(5) cancer	C0998265 Cancer Genus
	C0006826 Malignant Neoplasms
	C1547140 Specialty Type – cancer

Furthermore, MMTx denotes "cancer" as the headword of the phrase. Following the rule that the relevant concept of a phrase most likely concerns the headword, the relevant concept in this example must be a member of the fifth set that contains three candidates.

The final decision about the correct relevant concept has to be made during the further processing steps.

Semantic Type Test

In this test the semantic relevance of the markable candidates is analyzed. Furthermore, in case of ambiguous relevant concepts this test helps to select the correct relevant concept among the set of possible candidates.

A markable candidate is considered semantically relevant if its relevant concept, or at least one of its relevant concept candidates (as in the above example), is asserted to one of the semantic types defined in the relevant semantic type set (see Appendix).

Our coreference resolution algorithm retrieves this information through a hand crafted configuration file that includes a listing of predefined relevant semantic types. The selection of these specific semantic types among all existing possibilities was processed manually taking into account an analysis of the most frequent appearances of semantic types in relevant markables identified within our target CPG documents.

If there is no concept that fulfills this requirement, the markable candidate gets dismissed from the relevant markable set. If only one concept exists that fulfills this requirement, the markable candidate is ruled as relevant markable and the concept is considered as relevant. If there still two or more concepts that fulfill the requirement, the markable candidate is ruled relevant, but the relevant concept determination requires further processing. In such a case the concept with the highest MMTx mapping score is selected.

In the above example, among the three remaining relevant concept candidates only one concept is semantically relevant.[3]

C0998265 Cancer Genus	Invertebrate (T009) of Living Beings (LIVB)
C0006826 Malignant Neoplasms	Neoplastic Process (T191) of Disorders (DISO)
C1547140 Specialty Type – cancer	Biomedical Occupation or Discipline (T091) of Occupations (OCCU)

"C0006826 Malignant Neoplasms" is determined as relevant concept. Since there is at least one concept that is semantically relevant, the markable passes the semantic type test and is considered as relevant markable.

Information Enrichment

For all noun or prepositional phrase markable candidates, that pass the entire previous tests, additional semantic information is gathered from the UMLS.

The UMLS Metathesaurus emulates a tree structure in order to represent associative and hierarchical relationships between its concepts. This fact is especially important for the resolution of bridging coreference. In order to identify hypernym/hyponym relations between two concepts in the Metathesaurus tree, the parent and child nodes of the concept have to be identified. Our algorithm searches the Metathesaurus for all entries that either hold a parent or child relation (rel) to the CUI of a specific concept:

- A parent node can be identified by the relation code **PAR** ("has parent relationship in a Metathesaurus source vocabulary") or by the relation code **RB** ("has a broader relationship").

- A child node can be identified by the relation code **CHD** ("has child relationship in a Metathesaurus source vocabulary") or by the relation code **RN** ("has a narrower relationship").

Finally, a set of all found CUIs that either represents a parent or child node is added to the markable under investigation.

The following example illustrates some of the parent and child nodes derived during the semantic information derivation process for the (bio)medical term "anemia" (UMLS Release 2006AA):

---

[3] **red** … not a relevant semantic type
green … a relevant semantic type
(see Appendix "A1 - Relevant semantic type set" for more information)

"anemia" → CUI = C0002871

cui = C0002871 AND rel = "PAR" (35 entries)

Hematologic Diseases (C0018939)
Blood and Lymphatic Disorders (C0851353)
Red blood cell disorder (C0221016)

cui = C0002871 AND rel = "RB" (12 entries)

blood disorder (C0018939)
Red blood cell disorder (C0221016)
Blood and Lymphatic Disorders (C0851353)

cui = C0002871 AND rel = "CHD" (161 entries)

Anaemia of chronic disorder (C0002873)
Anemia due to blood loss (C0948824)
Sickle cell anemia (C0002895)

cui = C0002871 AND rel = "RN" (121 entries)

Hemoglobin very low (C0474527)
sideroblastic anemia (C0002896)
Anemia of mother, with delivery (C0156844)

5.2.3 Coreference Resolution

This is the final stage of our coreference resolution approach. After filtering out the irrelevant markables among the set of all possible phrases in the CPG documents and deriving necessary semantic and syntactic information for the remaining relevant ones, a set of predefined coreference resolution rules is applied to two relevant markables in order to answer two questions:

1. Does a coreferent relation hold between these two markables?
2. If yes, what type of coreference is it?

Our approach broadly differs between two types of coreferent relations, sortal and pronominal coreference, depending on the type of markables involved. As already stated above the developed algorithm shall be able to treat both types, but since pronominal coreference is not very frequent in (bio)medical texts this process was not massively investigated.

Consequently, we focus on the resolution of sortal coreference. This type of coreference either holds between two noun phrases, a noun or a prepositional phrase, or two prepositional phrases.

Types of Sortal Coreference

During literature research in combination with the analysis of several CPG documents that we consequently use as input texts for the training phase of our coreference resolution system we identified three different types of coreference that frequently appear in these documents. This classification differs from the one presented in corresponding literature, but since our approach relies heavily on the information provided by the UMLS and especially the Metathesaurus we made them suitable for the use in our developed coreference resolution algorithm. From our point of view these three identified coreference types are relevant mainly for this specific domain of discourse and for the use in subsequent projects that also deal with CPGs:

- **Acronym definition coreference**
  Acronyms are frequently used in clinical documents in order to shorten medical terms. An acronym is an abbreviation formed mostly using the initial components of the original phrase.
  We define an acronym definition coreference if a coreferent relation holds between two terms in an input text whereas one expression is the long form of a medical term and the second is its acronym that is formed strictly by the long form's initial letters. Furthermore, the following three conditions have to be met:

  1. The long form of the medical term must be immediately followed by its acronym.

  2. The abbreviated expression has to be put in parenthesis.

  3. The abbreviated expression must only consist of the initial letters of the lexical items (words) of the long form of the medical term.

  In the following example all of the three conditions are met. Consequently, our algorithm would identify an acronym definition coreference that holds between the two expressions "non-small cell lung cancer" and "(NSCLS)".

  "For patients with stage I and II non-small cell lung cancer (NSCLC), surgery to remove the NSCLC is the treatment of choice."

- **Acronym coreference**
  We define a coreferent relation that holds between two terms in an input text as a pure acronym coreference, if the following conditions are met:

  1. Both textual expressions have must have the same headword.

  2. The headword has to be an abbreviated expression as defined above. This means that an acronym definition coreference must hold between two phrases, the long form of the acronym and one with the same headword somewhere before in the text.

  3. The term must not be associated with a Metathesaurus concept.

The next example shows an acronym coreference as defined for our approach that holds between the two terms "(CT)" and "of CT". The term "(CT)" is an abbreviated form of the expression "computed tomography". This means that an acronym definition coreference holds between those two terms. In the next sentence we find the phrase "of CT" with the same headword as the phrase "(CT)". Furthermore, there exists no concept for "CT" in the Metathesaurus. Considering all of these factors, our algorithm would detect an acronym coreference relationship.

"Evaluation with preoperative computed tomography (CT) scanning of selected patients …
Many series have reported the utility of CT in detection of liver metastases …"

- **Hypernym/Hyponym coreference**
  A hypernym/hyponym coreference exists, if a coreferent relation holds between a more general expression (hypernym) and a more specific expression (hyponym). In our approach we distinguish between two different types of this special coreference type:

  1. In the first scenario ("type A") a hypernmy/hyponym coreference holds between two medical expressions if the relevant concepts of these phrases are in a direct or indirect parent-child or broader-narrower relationship in the UMLS Metathesaurus. A direct relationship exists when the two relevant concepts are directly connected, whereas an indirect relationship is characterized by the existence of an intermediate level (in our approach only one intermediate level is possible).
     In the next sentence for example, the phrases "surgery" (C0543467) as the more specific term and "the treatment" (C0087111) as the more general term hold a hypernym/hyponym coreference, because their relevant concepts are in an indirect parent-child relationship in the Metathesaurus.
     → C0087111 is **PAR** of C0679624 and C0543467 is **CHD** of C0679624

     "For patients with stage I and II non-small cell lung cancer (NSCLC), surgery to remove the NSCLC is the treatment of choice."

  2. The second type ("type B") of hypernym/hyponym coreference holds between two medical phrases that as a prerequisite have to share the same headword. Additionally, only if there exists one or more modifiers in one phrase and no modifier in the other phrase our algorithm rules this constellation as hypernym/hyponym coreference. In the next example both phrases have the same headword "anemia", but only one phrase has a modifier "chemotherapy-associated".

     "The use of epoetin is recommended as a treatment option for patients with chemotherapy-associated anemia and a hemoglobin concentration that has declined to a level <10 g/dL. Red blood cell transfusion is also a treatment option depending upon the severity of anemia or clinical circumstances."

In order to identify such coreferent relations within the target texts our algorithm uses two information sources:

1. Semantic and syntactic information
2. Coreference resolution rules

Semantic and Morphological Information

This kind of information is required for each markable. The coreference resolution rules take it as an input in order to confirm or refuse coreferent relations.

In our approach the morphological information is presented in the input text itself, whereas semantic information is provided by the MMTx program respectively the UMLS and stored at the relevant markables. Hence, each relevant markable holds the following type of information:

- **Headword**
  The headword of the phrase represented by the current markable.

- **Modifiers**
  A list that consists of all nouns, adjectives, and adverbs that serve as modifiers for the headword of the current markable.

- **Relevant Metathesaurus concept**
  The Metathesaurus concept representing the headword of the current markable. If the headword is related to more than one concept in the Metathesaurus, the relevant concept has already been selected during the previous processing steps.

- **Parent concepts**
  A list of all Metathesaurus concepts that either holds a parent or broader relationship to the relevant concept of the current markable.

- **Child concepts**
  A list of all Metathesaurus concepts that either holds a child or narrower relationship to the relevant concept of the current markable.

- **Semantic types**
  A list of all semantic types from the Semantic Network that are related to the relevant concept of the current markable.

- **Section number**
  A number identifying the section (text environment) of the input text the current markable appears in.

Coreference Resolution Rules

Our algorithm uses several crafted rules in order to identify sortal coreference relations in the CPG documents. All relevant markables identified in the previous steps are considered as potential anaphors. A possible coreference relation is resolved by applying the coreferent

resolution rules to a potential anaphor and to all members of its antecedent candidates set. This set contains of all relevant markables that appear prior within the text.

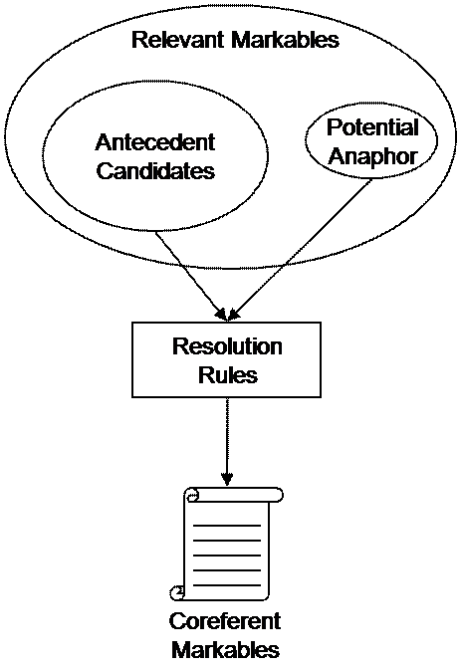

Figure 14: Schematically illustration of the coreference resolution step

Each rule uses parts of the above presented morphological, semantic and/or syntactic information of a potential anaphor and an antecedent candidate as input.

For each of the three coreference types distinct resolution rules had to be identified:

- **Acronym_Definition**
  In order to resolve acronym_definition coreference our algorithm uses the following rules to dismiss unlikely anaphor-antecedent pairs. At first, if the potential anaphor is not put in parenthesis and not the immediate successor of the antecedent candidate, the phrase pair under investigation cannot be ruled as coreferent. Additionally, if the potential anaphor phrase is not formed by the initial letters of the words of the antecedent candidate phrase our approach also does not rule the two phrases as acronym_definition coreferent.
  In some cases it might be possible that the abbreviated expresion is not connected to any UMLS Metathesaurus concept. In other words, the abbreviated medical term does not exist in the Metathesaurus. Since we are interested in the resolution of all acronym coreference occurrences in a medical input text, our approach does not take Metathesaurus concepts into account in the resolution process. Instead, we only rely on the morphological information presented in the text itself.

- **Acronym**
  The resolution of pure acronym coreference, as defined in our approach, is mainly based on morphological and “nonexistent” semantic information in the UMLS Metathesaurus. At first, if the two phrases under investigation do not share the same headword the pair gets dismissed. Secondly, if the shared headword is not an acronym, i.e. it must be the headword of the anaphor phrase of an acronym_definition corefernce pair somewhere prior in the input text our algorithm

also rules the anaphor-acronym pair as not coreferent. Finally, only if the headword of the two phrases is not connected to a Metathesaurus concept our approach approves an acronym coreference between the antecedent candidate and the potential anaphor.

- **Hypernym_Hyponym**
  As mentioned above, hypernym_hyponym coreference exists between a more general and a more specific term within an input text. As defined for our approach such a special constellation holds between two medical expressions if:

  a) The relevant concepts of the two terms are in a direct or indirect parent-child or broader-narrower relationship in the UMLS Metathesaurus.

  b) Both phrases share the same headword, but phrase has one or more modifiers and the other one has none.

  As a constraint we are only interested in hypernym_hyponym coreference that holds between two markables that are located inside a range of plus/minus one section within the target input text. Therefore, we have to at first apply a rule that only investigates two markables that fulfill this prerequisite.

  The identification of the first type requires semantic information derived from the UMLS Metathesaurus, whereas the resolution of the second one relies solely on morphological information that can be found directly within the input text. Therefore, our hand crafted resolution rules for the determination of hypernym_hyponym coreference have to follow two different strategies:

  1) In order to state a hypernym_hyponym coreference “type A” our algorithm has to check if a direct or indirect parent-child or broader-narrower relationship exists between their relevant concepts in the Metathesaurus. At first we investigate a potential direct relationship. Therefore, our algorithm takes the parent concepts and the child concepts of the antecedent candidate and check whether the relevant concept of the potential anaphor is a member of one of these two sets. Thereupon, the same determination is performed for the relevant concept of the antecedent candidate and the parent and child concepts of the potential anaphor.
     If no direct connection can be found, our approach tries to elicit a potential indirect relationship. Therefore, each parent concept of the potential anaphor is compared with every child concept of the antecedent candidate and each parent concept of the antecedent candidate with every child concept of the potential anaphor. An indirect relationship is identified, if a match is found in one of these determinations.

  2) For the resolution of hypernym_hyponym coreference “type B” our algorithm at first has to compare the headwords of the two markables under investigation. As mentioned above, a prerequisite for an existing coreference relation is the fact that both phrases, the antecedent candidate as well as the potential anaphor, share the same headword. Only if this is the case, we investigate the modifiers of the phrases. The information about the existence/nonexistence of modifiers helps to determine hypernym_hyponym coreference “type B” as defined for our

approach. Only if one medical term has one or more modifiers and by contrast the other phrase has none our approach rules such constellation as hypernym/hyponym coreference

In most cases these rules provide an adequate amount of information in order to determine the correct antecedent candidate for a potential anaphor. Nevertheless, it is still possible that our algorithm identifies more than one suitable antecedent candidate for one anaphor. In such cases we apply a closest first preference rule that selects the closest antecedent candidate phrase as the correct one.

6 PERFORMANCE EVALUATION

This final chapter presents the performance measuring process for our coreference resolution algorithm.

Like any other NLP system, a coreference resolution system needs to run some kind of evaluation procedure in order to measure its performance and accuracy. These results can consequently also be compared with results provided by similar systems or humans in order to improve its resolution capability.

In means of performance evaluation we consider our algorithm as an isolated system. This gives us the possibility to measure its coreference resolution capability with respect to a so-called “gold standard” template, which is a predefined benchmark that is considered as ideal or absolute correct.

6.1 Training

Before the actual performance of a coreference resolution approach can be measured the system has to be trained on some target texts that belong to the specific domain of discourse it should operate in. The goal of training is to improve the resolution capability of the approach. For this task we chose the following guidelines as training texts:

- Chemotherapeutic management of stage IV non-small cell lung cancer [Socinski at al., 2003]
- Chronic cough due to lung tumours [Kvale, 2006]
- Singapore cervical cancer [SMOH, 2004]
- Use of epoetin in patients with cancer [Rizzo et al., 2002]

The training phase can be seen as an iterative process. On each of these texts the coreference resolution process was performed several times. After each iteration cycle the output was analysed. We intensively investigated the results of each resolution round in order to identify incorrectly resolved or missing coreference pairs. The resolution rules applied in the next round were adapted according to the findings made during this analysis. With the help of this approach it was possible to significantly improve the resolution capacity of our coreference resolution algorithm.

6.2 Evaluation Process

The general performance evaluation process for our coreference resolution algorithm is illustrated in Figure 15. It basically includes two units that process the same input information in order to create equally assembled result sets that contents can be subsequently compared by a third unit. This so-called scoring procedure leads to a report that indicates the overall as well as some special defined performance measure of the approach.

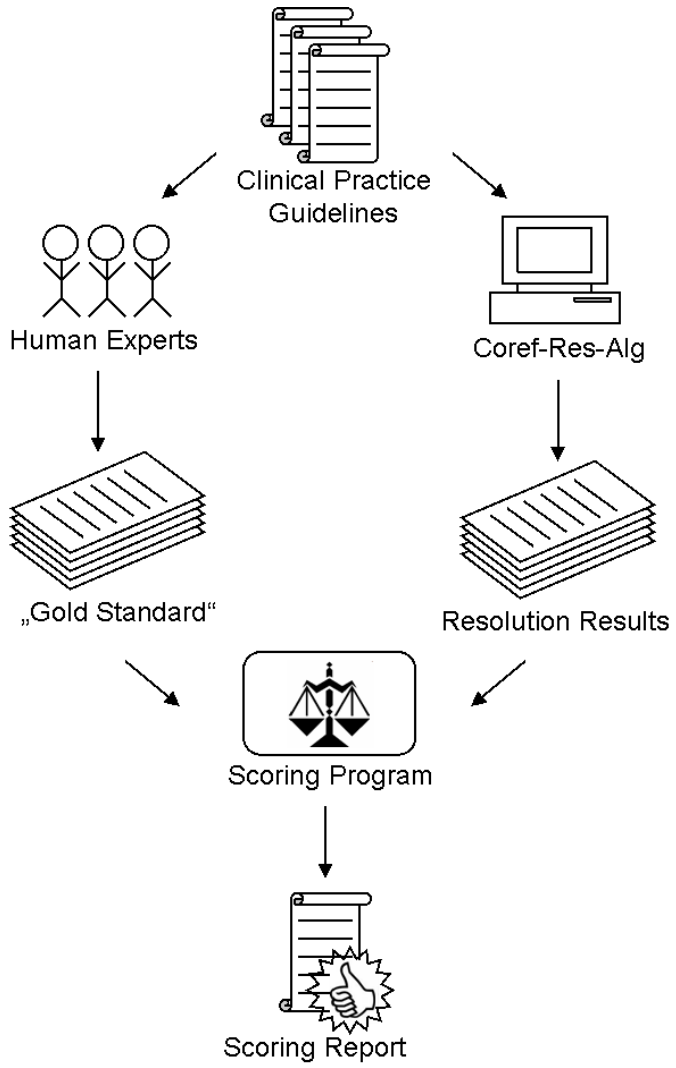

Figure 15: High level performance evaluation process. Adapted according to **[Lehnert et al., 1994]**

The input for this task is one or a set of target texts. In most cases these texts belong to a specific domain of discourse that is subject of the actual coreference resolution process. In this particular case the analysed documents are CPG documents that server as input for human experts as well as for our coreference resolution algorithm.

For each target CPG document the following steps have to be performed:

1. **Gold standard creation**
   One or a team of human experts create a template that is regarded as definitive. This so-called gold standard includes all possible coreferences that exist in one CPG document. The creation of a cold standard is a very expensive and labor-intensive process, depending on the size of the target text, but it provides the possibility to infinitely repeat the performance measure process.

2. **Coreference resolution**
   The to be evaluated coreference resolution algorithm is executed in order to detect as much coreferent antecedent–anaphor pairs as possible within a CPG document. Therefore, the three main tasks (i.e., phrase detection, relevant markable determination, and coreference resolution) are applied one after another. This

process results in a set of phrase pairs that are considered as coreferent according to a certain resolution strategy.

3. **Scoring**
   The actual performance evaluation is executed in the final step. A scoring program aims to compare the values in the gold standard with the antecedent–anaphor pairs resolved by the coreference resolution algorithm. Therefore, several scoring rules are applied to these two information sources in order to identify correctly, not correctly and partially correctly resolved coreference appearances.

6.3 Gold Standard Creation

As mentioned above a gold standard is a benchmark used for comparison that is considered as ideal or absolute correct. Since the accuracy of the whole evaluation process relies solely on the correctness of the gold standard, it is absolutely important that its creation is performed by one or preferable several experts in the corresponding area featured with the maximum level of background knowledge.

Unfortunately, due to expense limitations the creation of the gold standard templates used in the evaluation of our coreference resolution approach was performed by only one human expert. In order to reduce the risk of wrongly defined coreference pairs in the gold standard templates the creating expert was obligated to use the online version of the UMLS [USNLM, 2008] as an additional information source.

The creation of the gold standards itself can be divided in several steps. Source of the process are the original CPG documents. In a first stage these texts have to be split to phrase level. This is done with the help of the MapFace program [Gschwandtner et al., 2008], which uses MMTX in order to tokenize text into sections, sentences, and phrases. Additionally, it maps the noun phrases in the input text to the best matching UMLS concept or set of concepts [Divita, 2005]. After this automated tokenization the human expert is able to modify the results within the MapFace editor in order to correct eventually wrong annotated phrases. The output of this stage is used to automatically create an XML-file that serves as the actual source of the gold standard creation process. It includes the original CPG text divided in sections, sentences, and phrases. Additionally, each of these elements is accordingly numbered.

The concrete task for the gold standard creator is to identify valid antecedent-anaphor pairs. Therefore, the information presented in the XML-file is used to describe these tuples. Such a description consists of the information necessary to unambiguously identify the two phrases involved as well as the type of coreference. Each phrase is identified by its section, sentence, and phrase number. Furthermore, the phrase string itself is presented. An identified coreference relation looks like this:

```
<coreference_relation>
      <string1 section="1" sentence="2" phrase="1">Ovarian cancer</string1>
      <string2 section="1" sentence="2" phrase="7">
                a non-epithelial tumor</string2>
      <relation type="HYPERNYM_HYPONYM"/>
</coreference_relation>
```

All coreferent phrase pairs together are saved in a new XML-file that is the real gold standard, which gets subsequently used in the scoring program.

Our coreference resolution system creates similar XML files as output of the resolution process. The identified coreference relations are represented in the same way. This makes it easy to compare the values and to process the performance evaluation.

6.4 Scoring Program

The overall system performance of our algorithm is measured with the help of a scoring program. It compares the values presented in the two XML files, the gold standard and the one created by our coreference resolution algorithm, in order to calculate two performance metrics:

- **Recall**
  "The recall score measures the ratio of correct information extracted from the texts against all the available information present in the texts" [Lehnert et al., 1994].

- **Precision**
  "The precision score measures the ratio of correct information that was extracted against all the information that was extracted" [Lehnert et al., 1994].

For this calculation process the scoring algorithm firstly computes the value of the following predefined variables [Lehnert et al., 1994]:

- **POS (possible)**
  The total number of coreference relations according to the gold standard template.

- **ACT (actual)**
  The number of coreference relations identified by our coreference resolution algorithm. (= COR + PAR + INC)

- **COR (correct)**
  The number of correct coreference relations identified by our coreference resolution algorithm.

- **INC (incorrect)**
  The number of incorrect coreference relations identified by our coreference resolution algorithm.

- **MIS (missing)**
  The number of coreference relations erroneously not identified by our coreference resolution algorithm.

- **PAR (partial)**
  The number of partially correct coreference relations identified by our coreference resolution algorithm. A partially correct identified pair exists, if a coreference relation was indicated by our algorithm, but the coreference type is not correct.

With the help of these variables it is possible to compute recall and precision values for each target CPG document. Recall (= (COR + PAR) / POS) describes the ratio of correct identified coreference relations against all possible coreference relations as defined in the gold standard or in other words how good the system is able identify the investigated information. By contrast, the precision (= (COR + PAR) / ACT) score shows the ratio of correct identified coreference values against all identified coreference values or in other words how good the system is in identifying not relevant information.

The overall performance of the approach is the average recall and precision score over all target CPG documents.

6.5 Evaluation Results

We now present the evaluation results of our coreference resolution algorithms in order to measure its performance and accuracy. As mentioned above after a training phase with several rather short guideline texts we evaluate the performance of our approach by accomplishing a quantitative evaluation procedure. A scoring program compares the coreferences resolved by the algorithm with a gold standard template created by human experts that is regarded as definitive or absolute correct. For this process we use the following three guidelines developed by the Scottish Intercollegiate Guidelines Network (SIGN)[4]. Out of these guidelines we selected six representative chapters on which we measured the performance of our algorithm:

- SIGN 67: Management of colorectal cancer [SIGN, 2003a]
  - Chapter 6: Diagnosis
  - Chapter 7: Surgery
- SIGN 68: Dyspepsia [SIGN, 2003b]
  - Chapter 3: Management of uncomplicated dyspepsia
  - Chapter 4: H. pylori tests
  - Chapter 5: Management of functional dyspepsia
- SIGN 69: Management of obesity in children and young people [SIGN, 2003c]
  - Chapter 5: Treatment

In Table 9 we present the achieved recall and precision score for each of the six SIGN guideline chapters. For calculation of recall and precision we use the two formulas presented above.

In our analysis we furthermore distinguish between a common and a separate performance measure for the three types of coreference that are in scope of this work. This gives us the chance to subsequently investigate the strengths and weaknesses of our algorithm.

The overall achieved evaluated scores are **84,96%** in recall and **68,49%** in precision.

During further analysis of the result especially of the missing or erroneously resolved coreferences we aimed reason these numbers by categorizing them. We were able to identify three sources of mistakes that are described below.

[4] http://www.sign.ac.uk/ (last assessed: March 12, 2009)

Table 9: Evaluation results

		POS	ACT	COR	INC	MIS	PAR	REC (%)	PRE (%)
SIGN 67.6	Acronym Definition	2	2	2	0	0	0	100,00%	100,00%
	Acronym	0	1	0	1	0	0	---	0,00%
	Hypernym/Hyponym	28	32	23	9	5	0	82,14%	71,86%
	Overall	30	35	25	10	5	0	**83,34%**	**71,43%**
SIGN 67.7	Acronym Definition	1	1	1	0	0	0	100,00%	100,00%
	Acronym	3	2	2	0	1	0	66,67%	100,00%
	Hypernym/Hyponym	77	96	64	32	13	0	83,12%	66,67%
	Overall	81	99	67	32	14	0	**82,72%**	**67,68%**
SIGN 68.3	Acronym Definition	0	0	0	0	0	0	---	---
	Acronym	0	0	0	0	0	0	---	---
	Hypernym/Hyponym	48	66	46	20	2	0	95,84%	69,70%
	Overall	48	66	46	20	2	0	**95,84%**	**69,70%**
SIGN 68.4	Acronym Definition	1	0	0	0	1	0	0,00%	---
	Acronym	0	0	0	0	0	0	---	---
	Hypernym/Hyponym	12	14	11	3	1	0	91,67%	78,57%
	Overall	13	14	11	3	2	0	**84,62%**	**78,57%**
SIGN 68.5	Acronym Definition	2	0	0	0	2	0	0,00%	---
	Acronym	3	0	0	0	3	0	0,00%	---
	Hypernym/Hyponym	59	79	59	20	0	0	100,00%	74,68%
	Overall	64	79	59	20	5	0	**92,19%**	**74,68%**
SIGN 69.5	Acronym Definition	0	0	0	0	0	0	---	---
	Acronym	0	0	0	0	0	0	---	---
	Hypernym/Hyponym	30	37	18	19	12	0	60,00%	48,65%
	Overall	30	37	18	19	12	0	**60,00%**	**48,65%**
Overall results		**266**	**330**	**226**	**104**	**40**	**0**	**84,96%**	**68,49%**

6.5.1 Mistakes Caused by Incorrect Information Produced by MMTx

A potential source for incorrectly resolved or missed coreferences is the information derived from the MMTx. The MMTx should correctly parse the input text to phrase level and subsequently correctly map medical expressions to UMLS Metathesaurus concepts. We identified two potential error scenarios caused by the MMTx:

1. In case that the input text is not correctly parsed on phrase level our algorithm is most likely to miss coreference relations defined in the gold standard. On the other hand it is of course also possible that the creator of the gold standard caused by mistake or by a lack of knowledge tags a phrase wrongly during the gold standard creation phase.
2. In case that a medical term is not mapped to the correct concept our algorithm will miss a coreference defined in the gold standard. On the other hand if the knowledge of the creator of the gold standard is not sufficient he or she might fail to identify a potential coreference relation that subsequently will be resolved by the resolution algorithm. We often focus such a problem in cause of acronym definitions. For example the phrase “performance status” is assigned to the correct concept, whereas in contrast its acronym “PS” is mapped to two Metathesaurus concepts that describe different terms.

6.5.2 Coreference Relations Missed by Our Resolution Rules

Each coreference relation defined in the gold standard, which is not resolved by our algorithm decreases the achieved recall score. The recall score of 84,96% achieved by our coreference resolution approach is absolutely competitive compared with the numbers presented by other systems. Nevertheless, during intensive analysis of the resolution results we tried to identify several reasons why an existing coreference was missed by our algorithm.

Insufficient Acronym Detection Algorithm

Taking a closer look to the performance numbers one might notice that our algorithm focuses some problems when it comes to resolve acronym_definiton coreference. Only three out of six possible occurrences could be determined correctly.
By definition an acronym_definition coreference holds between a long form of a medical term and its acronym that is formed strictly by the long form's initial letters as in "non-small cell lung cancer" – "(NSCLC)".

In several cases the acronym is not formed by the initial letters only, such as in "into histamine receptor antagonists" – "(H2RAS)". Our algorithm is not capable of identifying such constellations. Missing such acronym definitions leads reduces the recall score. As a consequence of a missed acronym_definition coreference, the entire related acronym coreferences were also missed by the algorithm. This also leads to a significant reduce in the recall score reached by our coreference resolution approach.

Resolution Rules not Compliant with Gold Standard

- It often happens that there are multiple suitable antecedent candidates available for one potential anaphor. In such cases we defined to apply a preference rule that selects the closest antecedent candidate phrase as the correct one. Several gold standard documents however did not only determine one coreference relation (closest antecedent–anaphor). Instead they included multiple relations, one for each suitable antecedent candidate and the anaphor. Since our algorithm is configured to only resolve the coreference relation that includes the closest antecedent candidate all of the other coreference pairs are missed like in the following example:

```
<coreference_relation id="13">
    <string1 section="9" sentence="2" phrase="8">with upper GI
    endoscopy</string1>
    <string2 section="9" sentence="5" phrase="9">as
    endoscopy</string2>
    <relation type="HYPERNYM_HYPONYM"></relation>
</coreference_relation>

<coreference_relation>
    <string1 section="9" sentence="3" phrase="14">symptomatic
    treatment</string1>
    <string2 section="9" sentence="5" phrase="9">as
    endoscopy</string2>
    <relation type="HYPERNYM_HYPONYM"/>
</coreference_relation>
```

If the gold standard defines a coreference relation between “as endoscopy” (section="9" sentence="5" phrase="9">) and “with upper GI endoscopy” (section="9" sentence="2" phrase="8”) as well as with “symptomatic treatment” (section="9" sentence="3" phrase="14”) our algorithm will only resolve the second one since “symptomatic treatment” is the closest antecedent candidate. The relation including “with upper GI endoscopy” will be missed.

- Another problem that is connected with resolution rules that are not compliant with the gold standard is illustrated in the following example.

```
<coreference_relation>
  <string1 section="22" sentence="1" phrase="14">
  early upper GI endoscopy</string1>
  <string2 section="23" sentence="1" phrase="9">
  of upper GI endoscopy</string2>
  <relation type="HYPERNYM_HYPONYM"/>
</coreference_relation>
```

According to the defined resolution rules two phrases that share the same headword hold a hypernym_hyponym coreference if one phrase has no modifier and the other one has one or more. In the example this is not the case (3 vs. 4 modifiers), but the creator of the gold standard ruled the two terms as coreferent. Obviously, this will be missed by our algorithm.

- Our algorithm also faces a problem with abstract modifiers. The following coreference relation defined in the gold standard is missed because the word “most” is ruled as a significant modifier. Consequently, the resolution rule identifies one modifier for each phrase and misses this coreference relation.

```
<coreference_relation>
  <string1 section="14" sentence="1" phrase="6">most children</string1>
  <string2 section="14" sentence="2" phrase="9">obese children</string2>
  <relation type="HYPERNYM_HYPONYM"/>
</coreference_relation>
```

6.5.3 Erroneously Resolved Coreference Relations

If our resolution algorithm determines a coreference resolution that is not defined in the gold standard, the precision score is decreased. During the testing phase we achieved a score of 68,48%. The reason for erroneously resolved coreference relations are mainly caused by the complexity of the UMLS Metathesaurus and the enormous number of relations that are defined between its concepts that are sometimes not comprehensible for a human being, even also not for a medical expert. Especially, in cases of an indirect hypernym_hyponym coreference as illustrated in the following example:

Colon (and some rectal) cancers may be excised by polypectomy at colonoscopy (polyp cancers), and cohort studies indicate that such lesions do not require further surgery unless there is histopathological evidence of tumour at the margin (incomplete excision), lymphovascular invasion or the invasive tumour is poorly differentiated.

In the above sentence our algorithm identifies an indirect hypernym_hyponym coreference between the phrases “further surgery” (C0543467) as the more general term and “incomplete excision” (C0728940) as the more specific one.

“Surgery” (C0543467) is parent of “Type of surgical procedure” (C0679638), which is parent of “Excision” (C0728940).

Such and even more complex indirect hypernym_hyponym relations are often resolved by our algorithm even though they do not exist in the gold standard. This of course reduces the achieved precision score.

7 CONCLUSION

In this final chapter we shortly summarize our developed coreference resolution approach and take a look at future enhancement that can help to improve the overall performance of the algorithm.

7.1 Summary

Clinical practice guidelines (CPGs) are medical documents presenting up-to-date and state-of-the-art knowledge to various practitioners. To support their automatic application in clinical tasks they have to be transferred in a computer-interpretable form. This is a difficult task to achieve. To accomplish this process it is first of all necessary to correctly interpret the discourse of the text. Coreference detection and resolution is one of the key tasks in this process.

A coreference relation is a certain linguistic structure that holds between two textual expressions (anaphor and a preceding antecedent) whereas both are related to the same referent in the real world. Such a proposition can be frequently observed in natural language text corpora since a human author tries to avoid word repetition by using a variety of noun phrases that describe the same object.

In this work we present a computerized coreference resolution approach that is capable to detect and resolve distinct coreference relations in medical documents. More precisely we focus on the resolution of three different types of coreference in CPG texts:

- **Acronym Definition coreference**
  This type of coreference is defined as one that holds between two terms whereas one expression is the long form of a medical term and the second is its acronym that is formed strictly by the long form's initial letters.

- **Acronym coreference**
  This type of coreference is defined as one that holds between two terms whereas both phrases share the same headword. As a constraint this headword has to be an acronym.

- **Hypernm/Hyponym coreference**
  This type of coreference is defined as one that holds between a more general expression (hypernym) and a more specific expression (hyponym).

The resolution strategy of our algorithm relies on different kinds of background knowledge, but mainly on the Unified Medical Language System (UMLS) that supplies the required domain dependent semantic information via its large (bio)medical repository, the UMLS Metathesaurus, and its Semantic Network. Furthermore, we extensively use the functionality provided by the MetaMap Transfer (MMTx) program that allows an analysis of the input text on a syntactic level and the mapping of (bio)medical text to UMLS Metathesaurus concepts.

Our developed knowledge-based approach can be basically divided into three main modules:

1. **Phrase detection**
   At first we apply the MMTx program that tokenizes and parses the input text in order to determine all existing phrases. All noun phrases and prepositional phrases identified in the input text are mapped to the best matching UMLS concept or set of concepts.
2. **Relevant markable determination**
   In the second step all existing phrases get searched through in order to determine relevant phrases (markable candidates) for the actual coreference resolution task. A relevant phrase is either a noun or prepositional phrase that is mapped to a UMLS concept with a relevant semantic type. In order to compute the relevancy of a markable candidate, semantic information from the UMLS Metathesaurus as well as information the Semantic Network is incorporated. All relevant markables identified among all markable candidates subsequently serve as anaphor and antecedent candidates for a possible coreferent relation.
3. **Coreference resolution**
   Each of the relevant markables serves as a potential anaphor. All preceding markables in the text are considered as candidate antecedents. A set of predefined coreference resolution rules that uses semantic information collected during the previous steps as well as morphological information derived directly from the input text is applied to each anaphor – candidate antecedent pair in order to denote the likelihood of an existing coreferent relation between these two markables.

The performance measures that are computed with the help of recall and precision are the most significant metrics in order to evaluate the accuracy of a NLP system. With the help of these benchmarks it is possible to compare the resolution capabilities of two or more approaches. Furthermore, we applied a scoring program in order to measure the efficiency of our system with respect to so-called “gold standard” templates, which are handcrafted benchmarks that are considered as ideal or absolute correct.

During the testing stage, where we performed coreference resolution on six chapters of three guidelines developed by the Scottish Intercollegiate Guidelines Network (SIGN), our algorithm achieved overall scores of 84,96% in recall and 68,49% in precision. The reduction in recall can be explained with a higher complexity of the test guidelines compared with the documents used in the training stage. Furthermore, the gold standard was sometimes not as accurate as desired. The missing precision score mostly depends on the complexity of the UMLS Metathesaurus and the enormous number of relations that are defined between its concepts that are sometimes not comprehensible for human experts.

Nevertheless, from our point of view these are promising results that form an important basis for further automated processing of CPG documents.

7.2 Future Work

In future we aim to improve the performance, recall as well as precision, of our coreference resolution approach. Therefore, we analyzed the erroneously resolved or missing coreferences in order to distinguish why these mistakes occurred.

Several acronym definition coreferences and consequently all related acronym coreferences were missed by our approach because of a too primitive acronym detection algorithm that is capable of identifying acronyms that are strictly formed by the initial letters of the long form only. An acronym detection algorithm that is trained on more guideline texts and consequently equipped with a higher number of acronym detection rules can help to improve the future resolution capability of the approach.

In some cases the algorithm missed to resolved coreference relations where one phrase included abstract modifiers. They were ruled as significant modifiers and therefore a potential hypernym/hyponym was not resolved. In order to improve our algorithm, some kind of syntactic information should be incorporated with the goal to identify abstract modifiers and subsequently exclude them from the information applied in the resolution process.

Another improvement to our approach would be a complex synonym detection algorithm that is able to identify a synonym relation between two phrases such as “dyspepsia patient” and “patient with dyspepsia”. Providing this functionality can also help to raise the recall and precision score of our coreference resolution algorithm.

Bibliography

[Aronson, 2001] A.R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the AMIA Symposium*, pages 17–21, 2001

[Aronson, 2006] A.R. Aronson. MetaMap: Mapping Text to the UMLS Metathesaurus. Technical Report, Lister Hill National Center for Biomedical Communications, National Library of Medicine, 2006.

[Bagga, 1998] Amit Bagga. Evaluation of Coreferences and Coreference Resolution Systems. In *Proceedings of the First Language Resource and Evaluation Conference*, 1998.

[Baldwin, 1997] Breck Baldwin. CogNIAC: High precision coreference with limited knowledge and linguistic resources. In *Proceedings of the ACL/EACL Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 38–45, Madrid (Spain), 1997.

[Black, 1999] Paul E. Black, Algorithms and Theory of Computation Handbook, CRC Press LLC, 1999, in Dictionary of Algorithms and Data Structures (online), U.S. National Institute of Standards and Technology. Available from: http://www.nist.gov/dads/HTML/longestCommonSubsequence.html (assessed: 03.07.2008)

[Cardie and Wagstaff, 1999] Claire Cardie and Kiri Wagstaff. Noun Phrase Coreference as Clustering. In *Proceedings of the Joint Conference on Empirical Methods in NLP and Very Large Corpora*, 1999.

[Castano et al., 2002] J. Castano, J. Zhang and J. Pustejovsky. Anaphora resolution in biomedical literature. *In Proceedings of the International Symposium on Reference Resolution*, Alicante, Spain, 2002.

[Cunningham, 1999] H. Cunningham. Information Extraction - A Users Guide. Research Memo CS-97-02, University of Sheffield, Sheffield, 1997.

[Cutting et al., 1992] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A practical part-of speech tagger. In *Proceeding of the Third Conference on Applied Natural Language Processing*. 1992.

[Denber, 1998] M. Denber. Automatic Resolution of Anaphora in English. Eastman Kodak Co., Imaging Science Division, 1998.

[Dimitrov, 2002] Martina Dimitrov. A Light-weight Approach to Coreference Resolution for Named Entities in Text. Master Thesis, University of Sofia, 2002.

[Divita, 2005] Guy Divita. MMTX-API Documentation. http://mmtx.nlm.nih.gov, Lister Hill National Center for Biomedical Communications, National Library of Medicine, 2005

[Eiken, 2005] U. Eiken. Corpus-based Semantic Categorization for Anaphora Resolution. Master Thesis, University of Bergen, 2005.

[Feldbaum, 1998] C. Feldbaum. WordNet: An Electronical Lexical Database. The MIT Press, Cambridge, MA. 1998.

[Field and Lohr, 1990] MJ Field, KN Lohr (Eds). Clinical Practice Guidelines: Directions for a New Program. Institute of Medicine, Washington, DC: National Academy Press, 1990.

[Garera and Yarowsky, 2006] Nikesh Garera and David Yarowsky. Resolving and generating definite anaphora by modeling hypernymy using unlabeled corpora. In *Proceedings of the Conference on Natural Language Learning (CoNLL)*, 2006.

[Gschwandtner et al., 2008] Theresia Gschwandtner, Katharina Kaiser and Silvia Miksch. MapFace - A Graphical Editor to Support the Semantic Annotation of Medical Text. In *Proceedings of the Junior Scientist Conference 2008 (JSC'08)*, pages 91–92, 2008.

[Hearst, 1992] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, pages 539–545, 1992.

[Hobbs, 1978] Jerry Hobbs. Resolving pronoun references. *Lingua* **44**(4):311–338, 1978.

[Hoste, 2005] Véronique Hoste. Optimization Issues in Machine Learning of Coreference Resolution. PhD thesis, 2005.

[Humphreys et al., 1998] BL Humphreys, DA Lindberg, HM Schoolman, GO Barnett. The unified medical language system: An informatics research collaboration. *Journal of the American Medical Informatics Association (JAMIA)*, **5**(1):1–11, 1998.

[Kvale, 2006] P. Kvale. Chronic cough due to lung tumors: ACCP evidence-based clinical practice guidelines. *Chest*, **129**(1): 147–153, 2006.

[Kibble and van Deemter, 1999] Rodger Kibble and Kees van Deemter. Coreference Annotation – Whither? In *Proceeding of 2nd International Conference on Language Resources and Evaluation*, Athens, Greece, 1999.

[Lappin and Leass, 1994] S. Lappin and H. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, **20**(4): 535–561, 1994.

[Lehnert et al., 1994] W. Lehnert, C. Cardie, D. Fisher, J. McCarthy, E. Riloff and S. Soderland. Evaluating an Information Extraction System. *Journal of Integrated Computer-Aided Engineering*, **1**(6): 453–472, 1994.

[Lin and Liang, 2004] Yu-Hsiang Lin and Tyne Liang. Pronominal and sortal anaphora resolution for biomedical literature. In *Proceeding of the conference on Chinese computational linguistics (ROCLING XVI)*, 2004.

[McCray et al., 1994] A.T. McCray, S. Srinivasan, and A.C. Browne. Lexical methods for managing variation in biomedical terminologies. In *Proceeding of the 18th Annual*

Symposium on Computer Application in Medical Care (SCAMC), pages 235–239. 1994.

[Mitkov, 1998] Ruslan Mitkov. Robust pronoun resolution with limited knowledge. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-1998/ACL-1998)*, pages 869–875. 1998

[Mitkov, 1999] Ruslan Mitkov. Anaphora Resolution: The State of the Art. Working paper, University of Wolverhampton, 1999.

[Mitkov, 2003] Ruslan Mitkov. Anaphora Resolution. Chapter 14 in Mitkov (ed): The Oxford Handbook of Computational Linguistics. Oxford University Press, pages 266–283, 2003.

[Munoz and Palomar, 2001] Rafael Munoz and Manuel Palomar. Semantic-driven Algorithm for Definite Description Resolution. In *Proceedings of Recent Advances in Natural Language Processing (RANLP-2001)*, pages 180–186, 2001

[Ng and Cardie, 2002a] Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, 2002.

[Ng and Cardie, 2002b] Vincent Ng and Claire Cardie. Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

[Poesio and Vieira, 1998] Massimo Poesio and Renata Vieira. A Corpus-based Investigation of Definite Description Use. *Computational Linguistics* **24**(2):183–216, 1998.

[Quinlan, 1993] John Ross Quinlan. C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco, CA. 1993.

[Rindflesch and Fiszman, 2003] Thomas C. Rindflesch and Marcelo Fizman. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. *Journal of Biomedical Informatics* **36**(6):462–477, 2003.

[Rizzo et al., 2002] Rizzo, J., Lichtin, A., Woolf, S., Seidenfeld, J., Bennett, C., Cella, D., Djulbegovic, B., Goode, M., Jakubowski, A., Lee, S., Miller, C., Rarick, M., Regan, D., Browman, G., and Gordon, M. Use of epoetin in patients with cancer: evidence-based clinical practice guidelines of the american society of clinical oncology and the american society of hematology. *Blood*, **100**(7):2003–2020, 2002.

[SIGN, 2003a] Scottish Intercollegiate Guidelines Network (SIGN). Management of colorectal cancer. SIGN publication 67, Scottish Intercollegiate Guideline Network (SIGN), Edinburgh (Scotland), 2003.

[SIGN, 2003b] Scottish Intercollegiate Guidelines Network (SIGN). Dyspepsia. SIGN publication 68, Scottish Intercollegiate Guideline Network (SIGN), Edinburgh (Scotland), 2003.

[SIGN, 2003c] Scottish Intercollegiate Guidelines Network (SIGN). Management of obesity in children and young people. SIGN publication 69, Scottish Intercollegiate Guideline Network (SIGN), Edinburgh (Scotland), 2003.

[SMOH, 2004] Singapore Ministry of Health (SMOH). Cervical cancer. Clinical practice guideline, Singapore Ministry of Health, Singapore, 2004.

[Snow et al., 2005] Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. Learning Syntactic Patterns for Automatic Hypernym Discovery. *Advances in Neural Information Processing Systems* **17**:1297-1304, 2005.

[Socinski at al., 2003] Mark A. Socinski, David E. Morris, Gregory. A. Masters and Rogerio Lilenbaum. Chemoterapeutic Management of stage IV non-small cell lung cancer. *Chest* **123**:226S-243S, 2003.

[Soon et al., 2001] Wee M. Soon, Hwee T. Ng, and Daniel C. Y. Lim. A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Journal of Computational Linguistics* **27**(4):521–544, 2001.

[Strube et al., 2002] Michael Strube, Stefan Rapp, and Christoph Müller. The influence of minimum edit distance on reference resolution. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002)*, pages 312–319, 2002.

[Trask, 1993] Robert Lawrence Trask. A Dictionary of Grammatical Terms in Linguistics. London & New York: Routledge, 1993.

[Torii and Vijay-Shanker, 2007] Manabu Torii and K. Vijay-Shanker. Sortal anaphora resolution in Medline abstracts. *Journal of Computational Intelligence* **23**(1):15–27, 2007.

[UMLS, 2006] UMLS Overview. Tutorial, http://www.nlm.nih.gov/research/umls/pdf/AMIA_T12_2006_UMLS.pdf, (assessed: January 2009)

[USNLM, 2008] United States National Library of Medicine: UMLS Knowledge Sources, April Release 2008AA Documentation. http://www.nlm.nih.gov/research/umls/umlsdoc.html (assessed: 21.05.2008)

[van Deemter and Kibble, 2000] Kees van Deemter and Rodger Kibble. Coreference in MUC and Related Annotation Schemes. *Journal of Computational Linguistics* **26**(2):629–637, 2000.

[Versley, 2007] Yannick Versley. Antecedent Selection Techniques for High-Recall Coreference Resolution. In *Proceedings of EMNLP/CoNLL*, 2007.

[Vieira and Teufel, 1997] Renata Vieira and Simone Teufel. Towards Resolution of Bridging Descriptions. In *Proceeding of the ACL Students Session*, Association for Computational Linguistics, 1997.

[Vieira et al., 2003] Renata Vieira, Caroline Gasperin, Rodrigo Goulart. From manual to automatic annotation of coreference. In *Proceedings of the International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization*, Venice (Italy), 2003.

[Yang et al., 2004] Xiaofeng Yang, Jian Su, Guodong Zhou and Chew Lim Tan. An NP-Cluster Based Approach to Coreference Resolution. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 226–233, 2004.

[Yang et al., 2005] Xiaofeng Yang, Guodong Zhou, Jian Su and Chew Lim Tan. Improving Noun Phrase Coreference Resolution by Matching Strings. In *Proceedings of the First International Joint Conference on Natural Language Processing – IJCNLP 2004*, Hainan Island, China, pages 22–31, Springer Verlag, 2005.

Appendix

A1 - Relevant semantic type set

T023|Body Part, Organ, or Organ Component
T031|Body Substance
T033|Finding
T034|Laboratory or Test Result
T046|Pathologic Function
T047|Disease or Syndrome
T058|Health Care Activity
T059|Laboratory Procedure
T060|Diagnostic Procedure
T061|Therapeutic or Preventive Procedure
T074|Medical Device
T093|Health Care Related Organization
T094|Professional Society
T095|Self-help or Relief Organization
T096|Group
T097|Professional or Occupational Group
T098|Population Group
T099|Family Group
T100|Age Group
T101|Patient or Disabled Group
T110|Steroid
T121|Pharmacologic Substance
T125|Hormone
T126|Enzyme
T127|Vitamin
T129|Immunologic Factor
T184|Sign or Symptom
T191|Neoplastic Process
T195|Antibiotic
T200|Clinical Drug
T201|Clinical Attribute
T203|Drug Delivery Device