Automatic term and relation extraction for medical question answering system

Fahmi, Ismail

Chapter 1

Introduction

The outcome of any serious research can only be to make two questions grow where only one grew before.
— Thorstein Veblen (1857 - 1929)

1.1 The Future of Search

"Question answering is the future of enterprise search," says Matthew Glotzbach, head of products for Google Enterprise (Grimes, 2007). This statement shows that Internet technology is getting closer to serving our daily life, and is expected to deliver answers like in the way we communicate every day. Consider, for examples, some hottest queries in the form of questions reported by the Google Trends[1] on 17 February 2008 below:

(1) a. what does kitt stand for,
b. what type of fuel does k.i.t.t. run on,
c. who is mike traceurs father,
d. what do the letters k.i.t.t. stand for,
e. what color does k.i.t.t. morph into outside the casino.

Above *questions* are among the 20 hottest *keywords* submitted by Google's users who were so curious about the history and peculiarities of the talking car, KITT, after watching the new Knight Rider on NBC that night. So, *what does KITT stand for?* Answer: *Knight Industries Two Thousand*.

Answers to these questions are usually short, direct, often factoid answers, and should not be a list of web pages that might contain the answer. At the present time, we have to dig into web pages to find the answer by ourselves. But in the future, this job should be delivered to machines that provide question answering services.

In the Computational Linguistics community, the Question Answering (QA) task has obtained a lot of attention. Since its inception as a track in the eighth Text REtrieval Conference (TREC) in 1999, and until TREC-15 (2006), QA systems were designed to return answers themselves, instead of returning documents that contain answers (Dang et al., 2006). Early editions of the track

[1]Google Trends, http://google.com/trends?hl=en (Retrieved: 17 February 2008).

aimed to answer *factoic* questions. Their answers are facts and usually short (e.g., *Where did the 6th annual meeting of Indonesia-Malaysia forest experts take place?* TREC 1999). Then, other types of questions were added, such as definition questions (e.g., *What is an atom?* TREC 2001), list questions (e.g., *List the names of chewing gums.* TREC 2003), and finally the task was extended with interactive QA (e.g, *Who was the first Imam of the Shiite sect of Islam? Where is his tomb?* TREC 2006).

Unfortunately, although it is one of the most frequent search topics on search engines (Zweigenbaum, 2003), health was not targeted by the above conferences (Wedgwood, 2005). They were focused on general questions, and were aimed at paving the way to the development of better QA methods and systems in general. And health information regarding the medical area, how urgent is the need for medical QA systems?

The need for systems is motivated by the estimation that the half-life of medical knowledge is about 7 years, or 2 years for biomedical knowledge (Zweigenbaum, 2003). This estimation is based on the vast development of new knowledge in this field. Ely et al. (1999) found that physicians spent less than 2 minutes on average seeking an answer to a question. Furthermore, a study revealed that a healthcare provider needs an average of more than 30 minutes to find answers from the PubMed information retrieval system (Hersh et al., 2002). As a consequence, this causes a new problem to doctors, as mentioned by Alper et al. (2001) as well as by Ely et al. (2002) that, "Doctors are overwhelmed by the amount of information available, yet they often cannot answer their questions about specific clinical problems." The vast amount of information which needs to be searched in limited time is currently a challenge faced by doctors.

Besides the health care professionals, there are patients that often look for medical information in the Internet before or after seeing their doctors. They want to know more about their condition by accessing various sources. Because of this, sometimes they are better informed than the doctor on the topic related to their diseases or symptoms (Zweigenbaum, 2003).

In these respects, the need for QA systems in the medical domain is obvious. The systems are very important both for the health care professionals and for the patients. Due to the limited time resources especially of doctors, these systems should be able to provide short but correct and trusted answers to their medical questions.

1.2 Medical Question Answering

What kinds of questions arise in a medical domain? Answers to this question will tell us whether we can use the same approaches in open-domain QA or whether we should approach the problems differently.

Based on an extensive survey on 100 clinical questions from family doctors, Ely et al. (2000) concluded that "Clinical questions in primary care can be categorized into a limited number of generic types." The first category is diagnosis question, asking for causes or interpretation of clinical findings (21.3%), for example:

(2) a. What is the cause of `symptom x`?
b. What is the cause of `physical finding x`?

c. What is the cause of `test finding x`?
d. Can `drug x` cause (adverse) `finding y`?

The next category are questions that are related to tests or procedures (11.8%), e.g., *What is the best test in* `situation x`*?*; treatments (9.3%), e.g., *Should I change the dose of* `drug x` *(in* `situation y`*)?*; and management of findings (4.8%), e.g., *How should I manage* `finding x`*?*

For these kinds of questions, Niu et al. (2003) showed that the current approaches to the open-domain QA are not adequate. Relational information in open-domain QA contains relations between named entities. However, they observed that "the types of named entities in the medical area are different." Unlike relational information in the open-domain QA, most of the medical named entities are in the form of terms or concepts of the following types: diseases, symptoms, medicines, treatments, and so on. To give a correct answer, a medical QA system should be able to identify the types of terms in a relation.

In the present thesis, we aim at extracting medical relational information that will become potential answers for a QA system developed in the IMIX QA project.[2] We investigated the use of syntactic and semantic dependency relations in extracting potential medical answers. Our main research question is:

> To what extent can term and relation extraction techniques contribute to medical question answering?

In the next sections, we describe several issues related to the extraction of medical relational information. These consist of extracting and labeling medical terms, and extracting medical term relations. For each issue, we introduce a sub-research question.

1.3 Term Extraction

How can a term be identified from text? A term is different from a general word, with respect to the meaning it designates. Consider, for example, the first paragraph of the HIV topic in the Wikipedia[3] below:

> *Human immunodeficiency virus* (*HIV*) is a *retrovirus* that can lead to *acquired immunodeficiency syndrome* (*AIDS*), a condition in *humans* in which the *immune system* begins to fail, leading to life-threatening opportunistic infections. Previous names for the *virus* include *human T-lymphotropic virus-III* (*HTLV-III*), *lymphadenopathy-associated virus* (*LAV*), and *AIDS-associated retrovirus* (*ARV*).

This paragraph contains 14 keywords (in italic font), which are marked in the Wikipedia as links or bold font. These keywords are considered as terms or concepts because they convey special meanings in a medical domain. They are different from other words in that paragraph, such as *previous*, *names*, or *begin* that have no special meaning in this domain.

[2]The project is funded by NWO (Netherlands Organization for Scientific Research), as part of the IMIX programme (Interactive Multimodal Information eXtraction).

[3]http://en.wikipedia.org/wiki/HIV (Retrieved: January 14, 2008)

Since in most cases such marking is not available in text, several approaches have been made to recognize terms in text. They can be classified into three categories, namely linguistic, statistical, and hybrid approaches. The last approach is a combination of the previous two approaches. In the next subsections, we discuss the linguistic and statistical approaches, and additionally the use of external knowledge in identifying medical terms.

1.3.1 Linguistic Approaches

A linguistic approach relies on linguistic information, such as part-of-speech (PoS) information, to identify terms. Since terms can consist of only a limited number of PoS tags (Wright and Budin, 1997), such as noun and adjective, we can extract all nouns from text to become candidate terms.

As an illustration to this approach, consider the following PoS-tagged part of the above paragraph:

```
Human/NNP immunodeficiency/NN virus/NN (/LBR HIV/NNP )/RBR
is/VBZ a/DT retrovirus/NN that/WDT can/MD lead/VB to/TO
acquired/VBN immunodeficiency/NN syndrome/NN (/LBR AIDS/NNP
)/RBR ...
```

The meanings of the PoS tags are: `NNP` (singular proper noun), `NN` (singular noun), `LBR` (left bracket), `RBR` (right bracket), `DT` (determiner), `WDT` (wh-determiner), `MD` (modal), `TO` (to), `VBN` (verb past participle), `VB` (verb base form), and so on. There are several tags for nouns, such as `NNP` and `NN`; and for our purpose in the next processing they can be simplified to `N`.

Dagan and Church (1994) used a simple regular expression filter to identify terms, in which they defined a term as a sequence of one or more nouns (`N+`). When this filter is applied to that PoS-tagged text, we will get the following candidate terms that matched with this filter:

- *Human immunodeficiency virus* (`N N N`)
- *HIV* (`N`)
- *retrovirus* (`N`)
- *immunodeficiency syndrome* (`N N`)
- *AIDS* (`N`)

The extracted candidate terms seem to be true terms. However, there is an incomplete candidate, that is *immunodeficiency syndrome*. The verb *acquired* is missing in this candidate because it was tagged as `VBN` (verb). Despite this imprecision problem, this example shows that it is possible to automatically extract terms from free text with the help of linguistic information, especially PoS tags.

A considerable amount of literature has been published on the linguistic method for term extraction. Dagan and Church (1994), Justeson and Katz (1995), Frantzi and Ananiadou (1997) used PoS tag filters; Ananiadou (1994) used a combination of PoS tags, lexical, and morphological information; Jacquemin

and Tzoukermann (1999), Hippisley et al. (2005) used the head-modifier relation of a multi-word term; and Morgan et al. (2003), Shen et al. (2003) used orthographic information to recognize gene names.

As most of the studies are for English while our intended application is in Dutch, we investigate the characteristics of Dutch terms and present our methods and results to answer the first research question below:

Research question #1 Which linguistic knowledge is most useful for recognizing terms in Dutch text?

1.3.2 Statistical Approaches

In some cases, linguistic information may lead to the extraction of nouns or noun phrases that have no specific meaning for a particular domain. To illustrate this problem, consider the following PoS-tagged sentence from the previous paragraph:

```
Previous/JJ names/NNS for/IN the/DT virus/NN include/VBP
human/JJ T-lymphotropic/NNP virus-III/NN (/LBR HTLV-III/NNP
)/RBR ,/, lymphadenopathy-associated/JJ virus/NN (/LBR
LAV/NNP )/RBR ,/, and/CC AIDS-associated/NNP retrovirus/NN
(/LBR ARV/NNP )/RBR ./.
```

The meaning of the PoS tags are: `JJ` (adjective), `NNS` (plural noun), `IN` (preposition), `CC` (coordinating conjunction), and `VBP` (verb non 3rd ps. sing. present).

When the previous linguistic filter is applied again to this PoS-tagged text, we will get the following candidate terms:

- *names* (N)
- *virus* (N)
- *T-lymphotropic virus-III* (N N)
- *HTLV-III* (N)
- *LAV* (N)
- *AIDS-associated retrovirus* (N N)
- *ARV* (N)

Most of them seem to be medical terms, except the word *names*. This word has no special meaning for the medical domain, and can be found in any general text. It is different from the term *T-lymphotropic virus-III*, for example, whose occurrence in other text, such as in a computer domain, is not likely.

Several statistical methods have been proposed to filter out candidate terms that are not relevant to a particular domain (Dagan and Church, 1994; Justeson and Katz, 1995; Shen et al., 2003; Nakagawa and Mori, 2003; Eumeridou et al., 2004). The simplest one is *frequency*. In this method, a candidate term is considered as a relevant term for a particular domain, if its frequency of occurrence in its domain text is relatively high compared to the frequency of occurrence of other candidates. Other methods, that are based on word co-occurrence, such as

log-likelihood (Dunning, 1993), *mutual information*(Church and Hanks, 1989), *dice* (Dice, 1945), and so on, are targeted to extract multi-word terms. These methods compute the stickiness or the collocation strength of words constructing a multi-word unit.

Besides the standard measures, several new methods have been proposed. For example, *C-value* (Frantzi and Ananiadou, 1996) uses frequency of occurrence, term nesting, and term lengths, and *NC-value* adds contextual information to the previous measure (Frantzi and Ananiadou, 1997).

In this thesis, we compare the performance of the various statistical methods in ranking candidate multi-word terms. Furthermore, we seek an answer to the second research question below:

Research question #2 What statistical approach to multi-word term extraction is most successful?

1.3.3 Using External Knowledge

The previous two approaches use information within text and its linguistic properties (PoS tags) solely. No external information is involved. However, in some domains, such as a medical domain, there is terminological information available. This terminology contains a list of terms relevant to that domain, which is not part of text being processed, and can be used to improve the performance of the linguistic and statistical approaches.

Several attempts have been made to exploit external knowledge that is usually in the form of a thesaurus, dictionary, or general corpus (Lauriston, 1996; Maynard, 2000; Vivaldi and Rodríguez, 2001; Fukushige and Noguchi, 2001; Xu et al., 2000; Drouin, 2003). For example, Maynard (2000) used the Unified Medical Language System (UMLS) Metathesaurus and Semantic Network to obtain semantic knowledge of a term. This information is added to the *NC-value*, which is aimed at increasing the scores of extracted terms that are semantically related to a medical domain.

In the present thesis, we work on Dutch medical corpora. Although health is the richest domain in terms of its terminological resources, the proportion of Dutch terms in the biggest medical terminology resource, i.e. UMLS, is very low. In this resoruce less then 0.5% of all Dutch and English terms are in Dutch.

Motivated by the benefits that can be contributed by external knowledge and by the limited resources for non-English terms, especially Dutch, we seek answers to the third research question below:

Research question #3 How can we use existing multilingual terminological resources for extracting non-English terms, especially in Dutch, and how can the resources be used in statistical and linguistic approaches?

1.4 Term Variations

The main purpose of a QA system is to allow users to ask their own natural-language questions, and to deliver exact answers extracted from a document collection. A problem may occur during that extraction, where answers may not be found in the documents simply because a term used in a question to designate a concept is different from the term used in the documents to designate the

same concept. For example, a user prefers using the term *luchtpijp* 'windpipe' (a popular term) in their questions, while the documents use the term *trachea* (a technical term) to refer to a tube that carries air to the lung. A QA system with simple term matching will miss the answers of the questions.

This problem has been addressed in Ferret et al. (2000) and Dowdall et al. (2003) as the problem of recognizing term variations and finding answers containing term variants in documents. There are several types of variations recognized in the references, such as, exact word or lemma (e.g., *drug* and *drugs*), synonyms (e.g., *navel* and *umbilicus*), and composition (e.g., *syndrome of Tietze* and *Tietze's syndrome*).

In the present thesis, we attempt to investigate the problem in Dutch medical questions and documents. Thus, our fourth research question related to this purpose is:

Research question #4 What are the types of term variations that occur frequently in Dutch medical questions, and how can we recognize some of the variation types from documents in Dutch?

1.5 Semantic Types

The importance of semantic types in answering medical questions has been discussed in Niu and Hirst (2004). In their medical QA system, a specific query format is developed that corresponds to the basic elements of semantic types in a question. This format is based on questions, that arose in a clinical teaching unit, as shown in the following example:

> "*In a patient with a suspected MI does thrombolysis decrease the risk of death if it is administered 10 hours after the onset of chest pain?*" (Niu and Hirst, 2004).

A query to their system contains information about patient's condition (problem), an intervention, and a clinical outcome. Each piece of information corresponds to a semantic type, such as *disease* (*MI*), *treatment* (*thrombolysis*), and *outcome* (*death*). In the above example, the relation between all of these terms constructed a treatment scenario.

To answer the question, the system should understand the relational information contained in text. And because terms in text also construct a relational scenario, such as a treatment scenario or a diagnosis scenario, therefore, the system needs to know the semantic types of the terms that are involved in the relations. This requirement conforms with the observation in Leroy and Chen (2005) that if the name of a gene and a disease co-occur, it is almost certain that there is a scenario, i.e. causal relation, between the two.

The above question is a complex question that might be difficult for an open-domain QA system to answer. However, the observation could also be applied to simpler questions as have been collected previously in Ely et al. (2000). Consider the following question:

(3) What is the cause of AIDS?

This question is asking for a name or a term that causes a *disease* (*AIDS*). If we look for all semantic types that may cause *diseases* (or labelled as *Pathologic*

Function) in the UMLS Semantic Network,[4] we will get the following types: *Bacterium, Fungus, Invertebrate, Manufactured Object, Rickettsia or Chlamydia, Substance*, and *Virus*. Thus, we know that the answer should be a term of one of these types. We know from the first sentence of the previous HIV topic page that *HIV* can lead to *AIDS*. Since *HIV* is labeled as *Virus* in the UMLS, it matches with one of the types in the UMLS Semantic Network above. Thus, *HIV* becomes the potential answer to that question.

The use of UMLS to classify medical terms has been reported in several studies (Bodenreider, 2000; Spasic et al., 2003; Niu and Hirst, 2004). This method will only be successful if terms can be mapped to UMLS concepts. For non-English terms, such as Dutch terms, which are less covered by the UMLS, this method is challenging.

In this thesis, our corpora contain a majority of Dutch terms, although some of the terms are in English, Latin, and Greek. We classify the terms based on semantic information in the UMLS. And considering that Dutch is one of minority languages in UMLS, we seek an answer to the fifth research question below:

Research question #5 How can we use the UMLS resources to classify non-English terms, especially Dutch terms?

1.6 Semantic Relations

How can a relationship between two terms be identified from text? This is the next question that we have to deal with in extracting relational information for a medical QA. To understand the problem, consider the following sentence:

(4) AIDS is caused by the retrovirus HIV.

The sentence contains two terms, namely *AIDS* and *the retrovirus HIV* or *HIV*. By applying a lexical analysis, looking at the words in the sentence, we know that the terms are connected by a relation, which is *is caused by*. Thus, a piece of relational information that can be extracted from the sentence is: *AIDS* [*is_caused_by*] *HIV*; if we change the passive voice into active, we will get *HIV* [*causes*] *AIDS*.

Serban et al. (2007) use a similar method to construct the relation pattern contained in a medical guideline sentence. They filter detected n-grams, i.e., a fixed sequence of words, using relevant semantic relations from an ontology, and replace the n-grams with corresponding conceptual patterns. For example, given an n-gram {*should receive*}, they replace it with the pattern {*recommendation_operator, treatment*}.

Difficulties arise, however, when we apply the n-gram method or the lexical method to extracting patterns from Dutch sentences. Unlike English, Dutch has a high degree of word order variation, and therefore surface word order will be less effective in capturing relevant patterns. Consider the following examples of Dutch sentences that express the same pattern *word_veroorzaak_door* 'is_caused_by':

[4]The **Semantic Network** consists of (1) a set of broad subject categories, or **Semantic Types**, that provide a consistent categorization of all concepts represented in the UMLS Metathesaurus, and (2) a set of useful and important relationships, or **Semantic Relations**, that exist between Semantic Types. (http://www.nlm.nih.gov/research/umls/meta3.html).

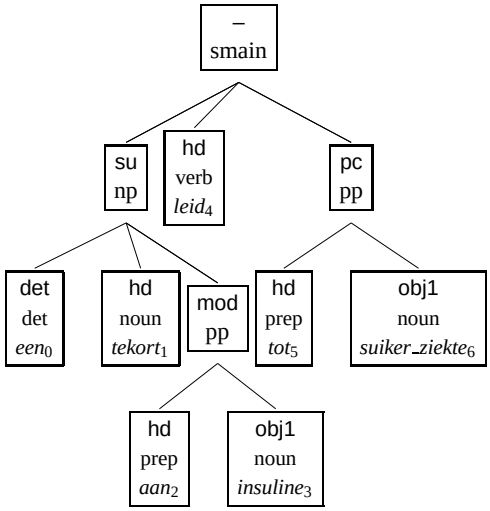

Figure 1.1: A dependency tree of the sentence *Een tekort aan insuline leidt tot suikerziekte* ('A shortage of insulin leads to diabetes').

(5) a. *AIDS word veroorzaakt door het retrovirus HIV*
'AIDS is caused by the retrovirus HIV'.
b. *Nachtblindheid wordt meestal veroorzaakt door een tekort aan vitamine A*
'Night blindness is usually caused by a lack of vitamin A'.
c. *Echte griep of influenza is een ziekte die veroorzaakt wordt door het influenzavirus*
'Real flu or influenza is a disease that is caused by the influenza virus'.
d. *Buiktyfus is een geheel andere (darm) ziekte , die door Salmonella bacteriën wordt veroorzaakt*
'Typhoid is a whole other (intestine) disease, which is caused by Salmonella bacterie'.
e. *Brucellose bij mensen wordt met name door brucella melitensis veroorzaakt*
'Brucellosis in humans is particularly caused by Brucella melitensis'.

These examples obviously show that the n-gram method will suffer from the word order variation. The separated words constructing the pattern are difficult to capture with this method. Therefore, a deeper linguistic analysis, such as dependency parse trees, is required to extract relations from Dutch text. In a dependency tree, every word is linked hierarchically to the other words it depends on, and this makes it possible to capture distant words of a pattern.

The use of dependency parse trees for this task has been addressed in several investigations (Rinaldi et al., 2006; Fundel et al., 2007; Katrenko and Adriaans, 2007). In general, the method can be described as the following. Given a dependency parse tree of a sentence, for example as shown in Figure 1.1, we define a starting point (X) and an end point (Y) of a relation. In this example,

we know that X is *Een tekort aan insuline* 'A shortage of insulin' and Y is *suikerziekte* 'diabetes'. The relation between X and Y is obtained by extracting paths, e.g., the `hd` (head) nodes, in the dependency tree that lead from X to Y; Thus, the extracted pattern is *leid tot* 'lead to'.

Rinaldi et al. (2006); Fundel et al. (2007) extract term relations based on manually defined patterns. For example, Fundel et al. (2007) use three rules for extracting protein-to-protein relations, namely '*A activates B*', '*Activation of A by B*', and '*Interaction between A and B*'. They only extract relations that match with these hand-written rules. The drawback of this method is that unseen relations cannot be identified. To solve this problem, Katrenko and Adriaans (2007) learn protein-protein interactions from dependency trees using a machine learning technique. They use a least common subsumer node of two nodes A and B as one of its features. However, it needs a dataset that has been labelled with proteins and their interactions.

In our QA project, the dataset was only labeled with relation types contained in its sentences. No labeling is present which explicitly identifies the arguments of the relation, and furthermore, it is not guaranteed that suitable arguments for a relation can actually be found within the sentence.

Besides that, there are a number of relation types that should be investigated, namely *causes, has_definition, diagnoses, occurs, prevents, has_symptom*, and *treats*. Creating relation patterns manually for each of these relation types is time consuming, and moreover, it is prone to missing unseen relations. We assume that we do not have any prior knowledge of what relation patterns are found with each of the relation types.

Proceeding from that condition, we use a two-step approach in extracting relations: (1) learning relation patterns automatically from training data, and (2) applying the patterns to extract relations from unlabeled text. While pursuing these tasks, we seek answers to the sixth research question below:

Research question #6 How can we learn relation patterns from dependency trees and use them to extract relations from text?

1.7 Claims

The following list provides the major conclusions that we can claim from our experiments to answer the defined research questions:

1. Part-of-speech information is the best linguistic knowledge for the extraction of candidate medical terms in Dutch text (Chapter 3).

2. One can use either Log-likelihood or χ^2 as the most promising method to detect association strenght of a multi-word term. While for extracting single-word terms, one can use a corpus comparison method, which is relatively simple and easy to implement (Chapter 3).

3. An existing multilingual terminology is useful for identifying new multi-word terms of a particular language, as long as there is word overlap between terms in both languages. And depending on the language pair, several methods can be used to increase the overlap, such as using stemming and translation (Chapter 3).

4. Synonym, abbreviation, and grammar-based variation are three types of term variation that frequently occur in Dutch medical questions (Chapter 4).

   - A very small seed list is enough for extracting synonyms through pattern learning and tuple extraction iterations.
   - One needs a heuristic that matches letters and their corresponding words to extract abbreviation pairs.

5. Thanks to translation, head words, surface length, and frequency, we can use a multilingual terminology containing terms and their labels to classify unseen terms of a particular language (Chapter 5).

6. In a dependency tree, a clausal node of category main clause, verb-initial main clause, subordinate clause, or infinitive clause is a starting point to extract a medical relationship from text, assuming that it contains subject and object labeled as medical terms (Chapter 6).

7. And for the main research question: Term and semi-automatic relation extraction techniques have improved the performance of our QA system. We also found that semantic labeling helps selecting potential answers (Chapter 8).

1.8 Chapter Overview

This book consists of 9 chapters, starting with this chapter, the *Introduction*. The next chapter elaborates problems in our QA system, particularly in extracting potential answers for medical questions, which leads to the necessity of extracting terms and their relationships from text. After describing the goals of our work, we explain the theoretical foundation of terminology and analyze previous approaches to extracting terms, detecting term variations, and extracting term relations. Then, we introduce our approach to these issues, and describe our expected contributions.

Chapter 3 reports our experiments on *Extracting Terms from Text*. We start this chapter by comparing two linguistic filters, namely a PoS-tag filter and a syntactic filter. We also compare eight statistical methods to get the most successful one in extracting multi-word terms. Based on these comparisons, we choose the PoS-tag filter for the linguistic approach, and χ^2 for the statistical approach. On the basis of these results, we describe our contribution to improving the statistical method, by combining the unithood and termhood values of multi-word terms. Then, we report the implementation of that method, which is called *Association and Domain Significance* (*ADS*), in extracting multi-word terms from Dutch text with the help of a multilingual terminology (UMLS). Finally, we describe our approach to extracting single-word terms. Some of the work in this chapter was presented in the RANLP workshop on Multi-source Multilingual Information Extraction and Summarization (Fahmi et al., 2007a).

Chapter 4 describes our method to detect *Term Variations*, one of important issues in a medical QA system, which uses one of the previously reported approaches in literature. The method, based on a lexico-syntactic approach, is

tested by extracting medical term variations from medical corpora. The results are evaluated manually and errors are analyzed.

Chapter 5 describes our approach to *Term Labeling*. We use a multilingual terminology (i.e. UMLS) as a knowledge source to classify non-English terms, especially Dutch. This method is based on a heuristic that utilizes syntactic information, translation, surface length, and frequency to find the best label from sample terms retrieved from the terminology. We evaluate the contribution of these parameters by comparing our labeling results to the labeling found in the corpus.

Chapter 6 explains our approach to *Learning and Extracting Relations*, that consists of two steps. The goal of the first step is to get a set of relation patterns for each of the seven relation types, from a set of relation-labelled sentences. We describe our method in extracting, weighting and filtering the patterns. Since the performance of the patterns can only be evaluated when they are used in extracting term relations, and this will be reported in the next chapter, therefore, in this chapter, we evaluate the patterns by comparing the extracted patterns across relation types. We analyze their overlap based on the assigned relation types and then discuss how we could improve the distinctiveness of relation patterns for an individual relation type. The second step is the extraction of relations from text. This step uses the relation patterns extracted earlier to extract new relations from unlabeled data. We describe our method in ranking and grouping the resulting relations. As for the evaluation, first we extract relations from a set of labelled sentences to get precision and recall figures for each of the relation types, and second we extract relations from an unlabeled medical dataset. We report their performance, discuss the benefits and disadvantages of dependency relations for this task and evaluate the role of semantic labels in extracting relations. We also discuss how we can extract relations from unseen patterns having a limited set of training data to improve recall.

Chapter 7 *Identifying Definitional Sentences Using Machine Learning* is another attempt to classify sentences to a particular relation type which will support the generation of training sentences for the relation extraction. The correctly classified sentences are expected to contain definition relations, which is required for learning definition relation patterns in chapter 6. We compare several machine learning methods with several feature configurations, and then discuss the results. The work described in this chapter was presented in the EACL workshop on Learning Structured Information in Natural Language Applications (Fahmi and Bouma, 2006).

Chapter 8 tests our extracted relations in the previous chapter by *Evaluating with the Question Answering System*. We describe our method in collecting test questions, and report the answering results which are provided by Joost, our QA system. For the evaluation, we measure the performance of Joost when it is supplied with answers from our method compared to when it is supplied with answers from a hand-written-pattern-based method.

Chapter 9, the final chapter, wraps up conclusions drawn from each chapter with general discussion and conclusions. It closes this thesis with some suggestions on future work.