Automatic term and relation extraction for medical question answering system

Fahmi, Ismail

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2009

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license. More information can be found on the University of Groningen website: https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment.

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 9

Conclusions and Future Prospects

You cannot acquire experience by making experiments.
You cannot create experience.
You must undergo it.
—Albert Camus (1913 - 1960)

9.1 Research Questions

The work reported in this thesis is motivated by the main research question below:

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

In the course of experiments conducted to answer the question, we raised the following six sub-questions:

1. Which linguistic knowledge is most useful for recognizing terms in Dutch text?
2. What statistical approach to multi-word term extraction is most successful?
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?
4. What are the types of term variation that occur frequently in Dutch medical questions, and how can we recognize some of the variation types from documents in Dutch?
5. How can we use the UMLS resources to classify non-English terms, especially Dutch terms?
6. How can we learn relation patterns from dependency trees and use them to extract relations from text?
We provide conclusions to those research questions in Section 9.2, and explain some future prospects in Section 9.3. Finally, we summarize our contributions in Section 9.4.

9.2 Conclusions

A term is defined in ISO-704 (2000) as “... a designation consisting of one or more words representing a general concept in a special language.” This function as a concept designator in a specialized vocabulary distinguishes it from words. In order to in general extract terms, we need to know how a term is formatted. There are three methods of term formation, i.e., using existing resources, modifying existing resources, and creating new linguistic entities. The second method is the most productive one, which creates new terms through affixation, compounding, creating phrasal terms, conversion (a verb used as a noun) and compression (abbreviation, acronym, etc). These processes involve linguistic knowledge, which can be different according to the domain language. Since we want to know the most useful linguistic knowledge for recognizing Dutch terms, we compared part-of-speech (PoS) tag and syntactic filters on Dutch text. With regard to research question #1, in Section 3.4 we concluded that the PoS-tag filter by Justeson and Katz (1995) recognized more true terms with a higher precision value compared to the syntactic filter which extracted noun phrases. The PoS-tag filter misses some terms within longer terms, and also fails to extract terms containing coordinations. The syntactic filter on the other hand, suffers from attachment errors (typically involving coordination and PP-modifiers) and misses sub-phrases which do not correspond to a full NP.

In many applications, linguistic filtering is the first step toward extracting terms. The next phase is to rank extracted candidate terms according to their statistical values, which place the most relevant terms on the top. To know the most successful statistical method, we conducted a set of experiments which compare eight statistical methods commonly used to measure the strength of association of bigram word strings. With regard to research question #2, in Section 3.5 we found that $\chi^2$ and Dice, both information theoretic-based measures, outperform all other methods. And particularly for frequency-based measures, we found that Log-likelihood is superior. Which measure we should choose for improvement? According to the formulae, Dice will only give a non-zero score if its numerator has a value, while $\chi^2$ will give a non-zero score as long as both parts of the term’s bigram are non-zero. Since this characteristic of $\chi^2$ is important for our improvement strategy, we choose this method for the next experiments.

The goal of term extraction is often not so much the creation of a new list of domain specific terms, but rather the (semi-)automatic extension of an existing term list. Assuming the availability of such a list of known terms in a particular domain, we attempt to improve the previously selected statistical method with the help of the list. In our medical domain, the most comprehensive existing list of terms is UMLS (Unified Medical Language System), in which most of the terms are in English and few in Dutch. We want to know how we can use such existing terms for the improvement. For this purpose, in Section 3.6 we developed a new scoring formula that uses a set of known terms to improve the performance of $\chi^2$. The formula combines two essential characteristics of a term,
9.2. Conclusions

namely association significance, which is measured based on the occurrence of candidate terms in a corpus, and domain centrality, which is measured based on overlaps found in candidate terms and known terms. To increase the overlap we apply stemming to candidate terms and known terms. In Section 3.7 we used that formula, which we called ADS (Association and Domain Significant), together with the UMLS as the source of known terms to rank new terms. We found that ADS_{lex}, which uses lexical matching and stemming, significantly improves the performance of $\chi^2$ alone. For multilingual terms, we believe that the degree of overlapping can also be increased with translation. Thus, with regard to research question #3, we conclude that an existing multilingual terminology is useful for identifying new multi-word terms of a particular language, as long as there is at least one word that overlaps in two terms of both languages. And depending on the language pair, several methods can be used to increase the overlap, such as using stemming and translation.

One of the important issues related to terms in a medical QA system is the detection of term variation, which is defined as alternative names for a concept. This issue arises due to, among other things, the different practices in the designation of medical terms among expert and general users. In Section 4.2 we collected a large number of medical questions posted in the Internet. With respect to research question #4, we found three term variation types which frequently occur in the medical questions, i.e., synonyms, abbreviations, and grammar-based variation. We investigated methods to extract the first two variation types. The grammar-based variation has been studied in Jacquemin (2001) and the FASTR tool was developed to detect terms and variation for French and English.

Our method to extract synonyms is adapted from the DIPRE method (Brin, 1999), and we incorporate syntactic information into its patterns as in Hearst (1992) and Pustejovsky et al. (2001). The syntactic information (PoS tags) is used mainly for detecting terms occurring in the corpus. The method consists of four steps run iteratively, which require a small seed list of synonym pairs in the very first iteration. Given the seed list and a parsed corpus, first we search the occurrence of the seeds in the corpus to generate a set of patterns. In the second step, we use the patterns to extract new synonym pairs from the corpus, and in third step we evaluate the extracted synonym pairs by computing their semantic compatibility based on their occurrence in the Web pages. In the last step, we use synonym pairs above a threshold as a seed list for the next iteration. We compared two phrasal patterns, i.e., ‘$x \text{ or } y$’ ($P_{or}$) and ‘$x \text{ NEAR } y$’ ($P_{NEAR}$), to measure semantic compatibility scores of the candidate synonym pairs. In Section 4.5 we found that the first phrasal pattern is superior. The second phrasal pattern, which has been used in Lin et al. (2003) to evaluate the compatibility of synonymous words, is not suitable for evaluating the compatibility of synonymous terms. It is because two terms co-occurring near each other often do not convey a similar meaning (synonym) but they may construct another relation type, such as a cause-effect relation or a location-disease relation.

To extract abbreviation pairs, our method consists of two extraction steps. First, apply a set of regular expressions to identify the occurrence of abbreviation short forms, and apply a PoS-tag filter to identify terms (abbreviation long forms) occurring before or after the short forms. Second, for each short form and long form pair, determine using a set of pattern-matching rules whether the pair is a valid abbreviation pair. We add a translation procedure in the rules
to detect abbreviations whose long form is of a different language. This rule is not considered in previous work. We compared this rule-based method with the two phrasal patterns. In Section 4.6 we found that the rule-based method outperforms the \( P_{or} \) and \( P_{NEAR} \) patterns. This high performance is caused by the fact that abbreviations and their long forms can be easily recognized only from their formations. On the other hand, the abbreviation pairs may contain new abbreviations which do not occur in the Web pages. Thus, the last two patterns may suffer from low recall. When we combined the rule-based method with \( P_{or} \), we achieved a higher F-measure value, which is contributed by the addition of abbreviation pairs which do not match with the rules.

An off-line medical QA system requires relation tables which contain medical relation instances. These instances can be expressed in text by rather general linguistic patterns, such as \( X \) may lead to \( Y \) or \( X \) occurs in \( Y \). Such patterns can nevertheless be used to extract medical relations with high accuracy if we require that both \( X \) and \( Y \) are medical terms, and if we apply the restriction that \( X \) and \( Y \) have to be terms that belong to a given class, e.g., a Virus and a Disease, respectively. To assign a class to a medical term, we index English and Dutch terms including their Semantic Types in UMLS, and retrieve the Semantic Types that match with the medical term. With respect to research question #5, in Chapter 5 we found that head words, surface length, frequency, and translation are pieces of information which are useful to match a new term with similar terms in UMLS, a multilingual terminology. Terms which have the same head word as the query term tend to share the same label with the query term. This conclusion is also supported by our finding that the precision of labeling terms tagged with adjective is low, since these terms usually do not have any function as head words. For multi-word terms, the number of words (surface length) is a good indicator for finding similar terms. Our evaluation showed that most of the multi-word terms in the corpus are new terms, of which 29.4% are labeled through exact translation.

There are 7 relation tables generated from the corpus, namely, causes, has_symptom, has_definition, occurs, treats, prevents, and diagnoses. Each of these tables is used by our QA system to answer a corresponding question type, by returning relation instances that match with the term in a question. Relation instances are extracted from text using semi-automatically-learned relation patterns. During these processes, we use dependency relation information to get the relation between the arguments of each relationship. With this dependency information, the relation between a predicate and its potentially distant arguments can be captured. This relation detection method is promising for our task because sentences in Dutch medical text tend to be long and complex. To make sure that we only extract meaningful and relevant relation instances for a particular relation type, as mentioned above, we require both arguments to be medical terms and that they belong to particular semantic types. With regard to research question #6, we can use a clausal node of category main clause, verb-initial main clause, subordinate clause, or infinitive clause as a starting point to extract a medical term relationship from text, which contains subject and object labeled as medical terms. Our experiments in Chapter 6 showed that the precision of our method is relatively high for most of the relation types, but recall varies. The method performs reasonably well for the has_definition and causes relation types, and performs less well for the diagnoses relation. We found that our relation patterns cannot distinguish causes from has_symptom relation type.
very well, because these relation types share some dependency patterns. An important source of the relation extraction errors was due to coreference. Our estimation indicates approximately 9% of the relation candidates in our corpus contain pronominal or definite NPs that needs anaphoric interpretation. An obvious next step would be to apply coreference resolution to medical terms, so as to obtain a full interpretation of the term, and a term which can be used for concept classification.

Our approach to learning relation patterns in Chapter 6 uses sentences that have been labeled manually with relation types. However, the distribution of labeled sentences according to relation types in our training corpus is not uniform. Some relation types have many labeled sentences, while the others have fewer labeled sentences. As an alternative to the manual approach, we can automatically increase the number of labeled sentences by using a supervised machine learning approach. In Chapter 7 we compared three learning methods, i.e., naïve Bayes, maximum entropy (ME), and support vector machine (SVM), in a task to classify sentences into definition and non-definition classes. It turned out that ME outperforms the other methods on a set of features which consists of bag-of-words, bigrams, syntactic properties, and sentence position.

After answering all of the six sub-questions, we now return to our main research question. In Chapter 8 we evaluated the extracted relation tables using our medical QA system, by answering 58 medical questions from three question types, namely has definition, causes, and has symptom. We compared an experiment setting which uses the relation tables extracted using our semi-automatic pattern learning method, with another setting which uses relation tables extracted using a manual learning method or an information retrieval (IR) method. In general, the manual method and the pattern learning method outperform the baseline (IR method), where they correctly answered more questions. And compared to the manual method, our pattern learning method clearly increases the performance of the QA system by correctly answering 11 more questions, or by about 37% of improvement. We found that relation tables generated using our method have a larger coverage and return higher precision answers compared to those generated using the manual method.

The effect of using semantic information on the performance of the QA system is clearly shown in Section 8.3.4. All of the correct answers are from dependency triples whose arguments have semantic labels. In general, about 75.6% of the answers are from dependency triples where both arguments have semantic labels. Only 31.3% and 27.8% of the answers to the has definition and causes questions, respectively, are from dependency triples where only one of the arguments has a semantic label. This finding suggests that we can improve the precision of the relation tables by taking into account dependency triples of the first and second semantic levels, without harming the recall of the correct answers.

Term variation is vital for a medical QA system since a medical term may occur in the corpus and in the questions with various forms. Two forms of variations, i.e., synonym and abbreviation, have been used in this experiment, and have shown to increase the performance of the QA system, especially in answering questions containing term variation.

Among the three question types, only the has definition question type receives incomplete answers. With respect to MRR scores, the performance of the three extraction methods for this relation type is the lowest, while for the
causes relation type is the best. This finding clearly suggests that a different approach to extract has_definition relations is needed. The \( x \text{ is } y \) pattern, which is very dominant in generating relation instances for the has_definition tables, is obviously not enough to detect correct and complete definitions. Having definitional sentences identified using the Maximum Entropy classifier with a set of features consisting of bag-of-words, bigrams, and syntactic properties (note that we did not use sentence position since the document formats of the training and testing corpora are different), we investigate whether the sentences can improve the performance of the QA system on this particular question type. In Section 8.5 we found that ME has classified most of the sentences the QA system needed to get relevant answers as definition, which results in more definition questions being answered correctly.

9.3 Future Prospects

If we return to the general workflow of our experiments in Section 2.4, we will find two applications that will use the output of the term and relation extraction processes, namely Ontology Population and Question Answering. In this thesis, we only describe our work on using the extracted relation tables to answer medical questions in our Question Answering system.

Considering that our evaluation on the QA system was intended as a preliminary and an exploratory work, which is aimed at finding an alternative approach to extracting relation tables from text for our QA system, in this work we only used a small set of medical questions to quickly see the contribution of the relation tables. To get a better picture of the result, we believe that a thorough evaluation using a larger number of medical questions is necessary. We do not carry out such an evaluation because it will require a lot of work and is out of the scope of this thesis. We think that a bigger project like IMIX or a bigger forum such as CLEF will be an appropriate place to conduct such an evaluation. In these evaluation forums, we can assign more than one individual annotator to evaluate returned answers for less biased performance measures.

We have extracted 7 types of medical relation tables and 7 types of medical questions, but only three of them, i.e., has_definition, causes, and has_symptom, have been used in the QA system. It is because we want to get a comparison with the existing method, the manually-extracted pattern method, where in this case only these three types of relation tables were generated. Therefore, in the future further work is needed to evaluate the extracted relation tables on all of the question types. This work should include extending the question analysis module of the QA system for new question types, i.e., occurs, treats, prevents, and diagnoses.

Besides its application on the QA system, our work is also suitable for the other application, Ontology Population. Gruber (2008) defines ontology:

“In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application.”
Following the definition, in an ontology building task we organize classes and their properties by means of relationships, while in an ontology population task we fill an existing ontology by populating instances of existing classes. The last task is similar to an Information Extraction task, where we extract relation triples or instances of particular classes and relationships from text. Our work in this thesis is exactly applicable to this task. The output of our work, i.e., terms, labels (classes), variations (which contain is a relations), and relations (including the 7 relation types), provides necessary information needed to add new instances into a medical ontology. This ontology can be derived, for example, from UMLS. If this is the case, our work can be used to enrich a part of the UMLS terminology, e.g., the Dutch part, with new concepts and relationships found in text.

If from relation tables we can retrieve relevant answers for a QA task, from an ontology we can retrieve concepts and their existing relationships, and infer new information expressed implicitly within the structure of the ontology. Examples of this application can be found in the semantic web field, where users search for information objects (e.g., disease names, persons, and events) and their relationships (e.g., causes, invents, involved in [event]) to get new insight of the objects. In the SWHi (Semantic Web for History) application, our other project (Fahmi et al., 2007b, 2008a), user can search for historical concepts, present the concepts in various ways (timeline, geographical location, and concept network map), and deduce new information (Fahmi et al., 2008b) thanks to properties and relations of the concepts in its ontology.

9.4 Contributions

The main contributions of this thesis are:

1. A comparison of two sources of linguistic information, i.e., part-of-speech and syntactic information, aimed at extracting candidate terms from text (Chapter 3).

2. A comparison of eight association measures and a corpus comparison method aimed at ranking candidate multi-word terms (Chapter 3).

3. A new statistical method called ADS (Association and Domain Significance), which combines association significance and domain centrality characteristics of terms, aimed at ranking candidate multi-word terms (Chapter 3).

4. An application of the ADS measure using a multilingual terminology (UMLS) as a source of known terms (Chapter 3).

5. An evaluation of a corpus comparison technique compared to frequency in ranking candidate single-word terms (Chapter 3).

6. Types of term variation frequently found in medical questions (Chapter 4).

7. An application of the DIPRE method combined with a linguistic filter to identify synonyms (Chapter 4).
8. A method to extract abbreviation pairs, which takes into account multilingual abbreviation long forms (Chapter 4).

9. A method to label Dutch medical terms using a multilingual medical terminology (UMLS) (Chapter 5).

10. A method to extract relation patterns and relation triples which involves linguistic and semantic information (Chapter 6).

11. A comparison of three learning algorithms, i.e., naive Bayes, Maximum Entropy, and Support Vector Machine (SVM), aimed at detecting definitional sentences (Chapter 7).

12. An application of the Joost QA system to answer three types of medical questions, i.e., \textit{has\_definition}, \textit{causes}, and \textit{has\_symptom} (Chapter 8).

13. A comparison of three answer sources, i.e., manually pattern-generated tables, semi-automatically pattern-generated tables, and an IR (Information Retrieval) system, aimed at answering medical questions with Joost QA system (Chapter 8).