Finding High-Frequent Synonyms of A Domain-Specific Verb in English Sub-Language of MEDLINE Abstracts Using WordNet

Chun Xiao and Dietmar Rösner

Institut für Wissens – und Sprachverarbeitung,
Universität Magdeburg
39106 Magdeburg, Germany
Email: xiao@iws.cs.uni-magdeburg.de, roesner@iws.cs.uni-magdeburg.de

Abstract. The task of binary relation extraction in IE [3] is based mainly on high-frequent verbs and patterns. During the extraction of a specific relation from MEDLINE[1] English abstracts, it is noticed that besides the high-frequent verb itself which represents the specific relation, some other word forms, such as the nominal and adjective forms of this verb, as well as its synonyms, also play a very important role. Because of the characteristics of the sub-language in MEDLINE abstracts, the synonym information of the verb can not be obtained directly from a lexicon such as WordNet[2] [1]. In this paper, an approach which makes use of both corpus information and WordNet synonym set (WN-synset) information is proposed to find out the synonyms of a domain-specific verb in a sub-language. Given a golden standard synonym list obtained from the test corpus, the recall of this approach achieved 60% under the condition that the precision is 100%. The verbs corresponding to the 60% recall cover 93.05% of all occurrences of verbs in the golden standard synonym list.

1 Introduction

The rapid growth of the size of digital databases inspired the research on automatic information extraction (IE) instead of the traditional manual IE. With the development of natural language processing techniques, more and more tools and resources are available, which leads to fruitful applications in the IE domain. Recent years the IE in biomedical domain has been also very well researched, particularly the task of named entity (NE) recognition. Moreover, relation extraction and event extraction have been also investigated.

Relation extraction is a main task of IE, as defined in the Message Understanding Conferences (MUCs) [3]. In recent years, the extraction of protein-protein interactions in biomedical articles and abstracts are reported in many works such as [2,4,5,6,7]. In this work, the relations to be extracted are binary ones, and the frequently occurring verbs as well as patterns are used in order to construct the template elements of the relations which will be extracted.

[1] PubMed offers free access to MEDLINE, with links to participating on-line journals and other related databases, available at http://www.ncbi.nlm.nih.gov/PubMed/

[2] http://www.cogsci.princeton.edu/~wn/index.shtml

Petr Sojka, Karel Pala, Pavel Smrž, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2004, Proceedings, pp. 242–247.

From the most frequent domain-specific verbs[3] in biomedical texts, we can learn the most frequent relations in this domain. From a test corpus with 800 MEDLINE abstracts extracted from the *GENIA Corpus V3.0p*[4], we can see that "induce", "mediate", "affect", and etc. are the most frequent domain-specific verbs in MEDLINE abstracts. Those high-frequent domain-specific verbs can be semantically categorized. For instance, the verbs such as "activate", "associate", and "interact" were used as the key verbs in extracting the protein-protein interactions in [2,4]. Theoretically, even given a complete lexicon which contains all the lexical entries, the categorization of the verbs in a corpus could still not be solved perfectly, if additional contextual cues are not available. Because many words are polysemous, i.e. they have more than one semantic interpretation, contextual information is necessary for disambiguation. In fact, we do not have such a perfect lexicon, even WordNet, therefore the situation is much more difficult.

In our experiment, we aimed to extract the inhibitory relation in MEDLINE abstracts, since this relation is one of the basic relations in the biomedical domain[5]. This work is based on some previous works such as NE recognition, part of speech tagging, even shallow or full parsing, etc. A very fundamental problem in this relation extraction task is how to choose the proper high-frequent verbs that represent an inhibitory relation.

Obviously the synonyms of the verb "inhibit" have to be taken into account, according to the synonym information provided by a lexicon such as WordNet. But the vocabulary of the sub-language of MEDLINE abstracts seems quite different compared to the general English[6]. Many of the synonyms of the verb "inhibit" provided by WordNet (Version 1.7.1) do not occur even once in the 800-abstract test corpus, such as "subdue", "conquer", etc. Some of these synonyms occur only with a very limited frequency, e.g. "confine" occurs only once in the test corpus. Instead, what can be found in the test corpus are verbs such as "block", "prevent", and so on, as example 1 shows. They are not in the synonym list of "inhibit" in WordNet but provide cues of an inhibitory relation.

Example 1. *Aspirin appeared to* **prevent** *VCAM-1 transcription, since it dose-dependently inhibited induction of VCAM-1 mRNA by TNF.*

Following shows the occurrences of some WordNet synonyms (WN-synonyms) of "inhibit", as well as some non-WordNet synonyms (nonWN-synonyms) in the 800-abstract test corpus.

- **WN-synonyms** suppress (69), limit (16), restrict (5)
- **nonWN-synonyms** block (124), reduce (119), prevent (53)

In addition, we found although the nominal forms of *"inhibit"* are more frequent than the verb forms, the verb "inhibit" occurs quite frequently in the test corpus. It is different from the familiarity description of "inhibit" in WordNet, which says *"inhibit used as a verb is*

[3] Actually, the domain-specific verbs should not include the general verbs independent of the domain in the scientific papers, such as "analyze", "indicate", "observe", and so on. Spasić et al. [8] also discussed this problem.

[4] GENIA project, available at `http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/`

[5] In some relation extraction works, inhibitory relation is treated as a kind of protein-protein interaction.

[6] WordNet is regarded as a semantic lexicon for general English, since its sources are quite broad [1].

rare". And we found that the "estimated frequency" in WordNet differs from that in the sub-language of MEDLINE abstracts. For instance, in WordNet, "restrain" is more frequent than "limit", but in the test MEDLINE abstract corpus, the situation is just reversed. This indicates that the expressions in the sub-language of MEDLINE abstracts are quite domain-specific.

This paper proposes an approach in order to find out these synonyms in the sub-language. It is constructed as follows: section 2 describes the approach of finding the synonyms of a verb in the sub-language of MEDLINE abstract, and, section 3 presents the result and discussion.

2 Finding Out Synonyms in Sub-language Corpus

Definition: Keyword, Core Word, and Language Unit In this experiment, let *keyword* denote a word whose base form is "inhibit", while *core word* denotes the verb "inhibit". For example, "inhibitory" and "inhibition" are both keywords in this experiment. A *language unit* may be a sentence, several sentences, or a paragraph, even several paragraphs, which expresses the same semantic topic.

In order to find out the synonyms of the core word, with the help of WordNet information, the corpus information is also considered. In this test the verbs which occur around a keyword in the text of an abstract are examined.

This idea comes from the assumption that the synonyms of a verb, which have very close semantic relation with its corresponding keyword, have a likelihood to co-occur in the same language unit with the keyword than with other words. Note that in our approach only the localization of all the verbs around the keyword is considered. Other information such as the sentence boundaries and sentence structures, are not considered yet, although they must be very useful in some other corpora. Because in MEDLINE abstract corpus, each abstract consists of only one paragraph, namely several sentences[7], and each abstract either has only one topic, or the topics in an abstract are dependent on each other, then the whole abstract can be treated as a language unit. The vocabulary of a language unit is limited heavily by the topic(s), which means it is very likely that the vocabulary consists of words that have close semantic relations to each other in a language unit. Namely, the vocabulary in the same language unit can be more probably grouped into fewer synonym or antonym sets. Moreover, with the localization of a keyword, the verbs around the keyword may be limited semantically to have semantic relations (synonyms or antonyms) with the keyword[8].

2.1 Method and Resources Used in The Experiment

Golden Standard List (S_G) for Evaluation At first a synonym list of the verb "inhibit" is obtained by counting the frequencies of each verb in a manually produced 50-synonym list in the test corpus, based on WN-synset information, and choosing the ones with more than 6 occurrences. By this process a 10-word synonym list is obtained, which is used in the following work as a golden standard list S_G. In S_G only 3 verbs come directly from the WN-synset of "inhibit", but the rest 7 verbs come from its hypernyms and the synonyms'

[7] For the 800-abstract test corpus, each abstract consists of 8.41 sentences in the average, excluding the title of each abstract.

[8] Because of the restriction of the pages, an example here is omitted.

synonyms. This golden standard list provides the standard to evaluate this approach.

Expansion of Synonym List (S_i): Learning Synonym Information from WordNet In order to make use of the WN-synset information, the synonyms of each word which is a synonym of "inhibit" are considered in order to improve the coverage of synonyms in the MEDLINE abstract corpus. Let S_i ($i > 0$) be the expanded WN-synset word list, it can be obtained in the following way: at first the synonym list of "inhibit" is expanded by adding all synonyms of this verb, the list contains 16 items by then, which is symbolized as S_1. Furthermore, S_1 can be also expanded by adding all synonyms of each verb in the list, the list is then expanded to be a 94-item one, i.e. S_2. If we want to enhance the recall, we can just expand this synonym list by recursively adding the complete synonym list of each word in this list again, and go on. But at the same time the misleading information will grow in an exponential way.

Verb List (V_j) from the Test Corpus: Collecting Verb Candidates (S_g) We can get a set of verbs (V_j) which are chosen from the test corpus around a keyword in the window size of j ($j > 0$), with the corresponding frequencies from the test corpus. The list provides the corpus information in our experiment. In the 800-abstract test corpus, for example, there are total 318 verbs around the keyword in a searching window of size 2. In these 318 verbs, the occurrences of 23 verbs are $\geq$ 26 times. It is quite surprising that in these 23 words, 9 of them are synonyms or antonyms of the verb "inhibit", including the verb itself. The expanded synset lists S_i ($i > 0$) are used to give synonym information of the high-frequent verbs around a keyword. If a high-frequent verb around a keyword or one of the synonyms of this high-frequent verb is in this synonym list, it will be added to the learnt synonym candidate list S_g.

Expansion of Misleading Verb List ($STOP_k$): Learning Misleading Information from Genre Analysis of Corpus and WordNet Because the sub-language in MEDLINE abstracts quite often uses the verbs to construct the whole abstracts structure, such as "suggest", "indicate", "show", and so on, they should be excluded from S_g. An initialized stop-word list $STOP_0$ is given with 15 such verbs (including several antonyms of "inhibit") in this experiment. However, the necessary expansion of the stop-word list $STOP_k$ ($k \geq 0$) is carried out also in a similar way as the expansion of S_i. If a verb v, $v \in V_j$ and $v \in STOP_k$, then $S_g = S_g - \{v\}$.

Balance between Recall and Precision This approach is a bidirectional one. That is, in one direction the positive synonym information is expanded according to WN-synsets, or the searching windows are enlarged, so that the recall will be improved but the precision will be impaired; in the other direction, the stop-word list is also expanded in order to improve the precision, meanwhile the recall will be impaired. Therefore, the balance between recall and precision is also very important. That means, the expansions of both the synonym list and the stop-word list are limited. For instance, in this experiment, the synonym list has been expanded for maximal 4 times (S_i, $i = 1...4$), whereas the stop-word list has been expanded only once ($STOP_1$). In addition, by only focusing on the relative high-frequent words in this experiment, the work of evaluating recall and precision is much simplified.

3 Result and Discussion

This approach makes use of three kind of sources. One is the synonyms information of the verb "inhibit" obtained independently from any corpus but from a lexicon (WordNet). The second is the frequencies of verbs around a keyword, which depends closely on the corpus. The last is the information of unlikely verbs, which depends partly on the verb "inhibit" itself, i.e. its antonyms, and partly also on the corpus, i.e. the verbs for the construction of MEDLINE abstracts.

Table 1. Recall (R_j) and precision (P_j) on synonym list S_i ($i = 1, ..., 4$), in searching window with window size j ($j = 1, ..., 5$). The word frequency limit in this table is ≥ 15 in the test corpus, with an expanded stop-word list of 256 items (first part of this table) and 1512 items (second part of this table), respectively.

256	R_1	P_1	R_2	P_2	R_3	P_3	R_4	P_4	R_5	P_5
S_1	20%	100%	40%	100%	40%	100%	40%	100%	40%	100%
S_2	30%	100%	**60%**	**100%**	60%	100%	60%	100%	60%	100%
S_3	30%	100%	60%	100%	60%	100%	60%	85.71%	60%	85.71%
S_4	30%	100%	60%	85.71%	60%	85.71%	60%	75%	60%	75%

1512	R_1	P_1	R_2	P_2	R_3	P_3	R_4	P_4	R_5	P_5
S_1	10%	100%	20%	100%	20%	100%	20%	100%	20%	100%
S_2	10%	100%	**30%**	**100%**	30%	100%	30%	100%	30%	100%
S_3	10%	100%	30%	100%	30%	100%	30%	100%	30%	100%
S_4	10%	100%	30%	75%	30%	75%	30%	75%	30%	75%

Look at the data with 256 stop words in Table 1, with the increase of expansion of both synonym and stop-word lists, the recall comes to 60% stably, in which only 33.4% comes directly from the WN-synset of "inhibit". And in the test corpus, the verbs corresponding to the 60% recall cover 93.05% of all occurrences of verbs in the golden standard list, this means that this approach finds out the most frequent synonyms of "inhibit" in the test corpus. It also indicates that these high-frequent synonyms distribute mainly in ± 2 positions around a keyword. Note that here *position* refers to a verb chunk around a keyword. In comparison to the data with 1512 stop words, the data with 256 stop words indicate when the stop-list is too large, it causes the decrease of recall sharply. Then the stop-word list should not be expanded too much so that the intersection of $STOP_k (k > 0)$ and $S_i (i > 0)$ can be minimized.

By this approach, it should be possible to semantically classify the high-frequent domain-specific verbs in MEDLINE abstracts for further IE tasks. However, this approach is limited to be applied in MEDLINE abstract corpus. Second, the core word occurring in the test corpus should not be too sparse. In case that the core word occurs with a low frequency in the test corpus, its synonyms with high frequencies should be considered instead. Since this approach focuses only on the high-frequent verbs in the corpus, the recall is rather moderate. In future work it will be investigated how syntactic cues and information from phrase patterns could improve the recall.

References

1. Christiane Fellbaum: WordNet: An Electronic Lexical Database. The MIT Press. Cambridge. Massachusetts (1998) Foreword, xv–xxii, Chapter 1, 23–46.
2. T. Sekimizu, H. S. Park and J. Tsujii: Identifying the Interaction between Genes and Gene Products Based on Frequently Seen Verbs in Medline Abstracts. Genome Informatics, Universal Academy Press (1998).
3. Jim Cowie, Yorick Wilks: Information Extraction. Handbook of Natural Language Processing (2000) 241–160.
4. J. Thomas, D. Milward, C. Ouzounis, S. Pulman, M. Carroll: Automatic extraction of protein interactions fromscientific abstracts. In: The Pacific Symposium on Biocomputing'2000, Hawaii (2000) 541–551.
5. A. Yakushiji, Y. Tateisi, Y. Miyao, J. Tsujii: Event Extraction from Biomedical Papers Using a Full Parser. In: The Pacific Symposium on Biocomputing. (2001) 6:408–419.
6. R. Gaizauskas, K. Humphreys, G. Demetriou: Information Extraction from Biological Science Jounal Articles: Enzyme Interactions and Protein Structures. In: The Workshop Chemical Data Analysis in The Large: The Challange of The Automation Age. (2001).
7. T. Ono, H. Hishigaki, A. Tanigami, T. Takagi: Automated extraction of information on protein-protein interactions from the biological literature. Bioinformatics, 17(2)(2001) 155–161.
8. I. Spasić, G. Nenadić, and S. Ananiadou: Using Domain-Specific Verbs for Term Classification. In: The ACL 2003 Workshop on NLP in Biomedicine. Sapporo, Japan, (2003) 17–24.