知 能 と 複 雑 系 128−5
（２００２．５．２３）

文脈情報を用いた医学用語分類

山田 寛康　　新保 仁　　松本 裕治
{hiroya-y,shimbo,matsu}@is.aist-nara.ac.jp
奈良先端科学技術大学院大学 情報科学研究科

本論文では, 自然言語処理技術と機械学習の手法を用いて, 英語医学生物学論文要旨に含まれる医学用語の意味クラス分類を行う. 医学用語の出現個所の前後の文脈情報, 用語の内部情報, 統語解析を行って抽出した単語間の依存関係を素性として用い, 各々, およびそれらの組み合わせの有効性について評価する. MEDLINE アブストラクトを対象に MeSH Tree の最上位ノードを意味クラスとみなして行った実験の結果は, 単語間の依存関係が単純な前後の単語列と比較してより有効な素性となる可能性を示している.

Context-Based Classification of Medical Terms

Hiroyasu Yamada　　Masashi Shimbo　　Yuji Matsumoto
{hiroya-y,shimbo,matsu}@is.aist-nara.ac.jp
Graduate School of Information Science, Nara Institute of Science and Technology

We investigate a practical method of classifying technical terms from the abstracts of medical and biological papers, with a main objective of identifying a set of features relevant to the task. The features considered are: (1) spelling of a term, (2) words around the occurrence of a term, and (3) syntactic dependency of a term with surrounding words. We evaluated the effectiveness of these features in a task of classifying terms in the abstracts from the MEDLINE database, in which target classes were determined in accordance with the first five top-level nodes of the MeSH tree. The results prove the dependency feature works more effectively compared with the sequence of words around terms.

1 Introduction

The ability to cope with technical terms is essential for natural language processing (NLP) systems dealing with scientific and technical documents. Since a majority of these terms are not in general-purpose dictionaries, domain-specific lexicons are often used in combination. It is still unrealistic to expect the lexicons to enumerate all technical terms, because in the active fields of research such as biology and medicine, new terms are produced on a daily basis. Another difficulty in dealing with technical terms is that they are often polysemous; even if terms are recognized with the help of the lexicons, the meaning of each occurrence of the terms must be identified.

Robust techniques are thus required for (1) recognizing technical terms, and for (2) identifying the semantic class of those terms. The first task was tackled by several researchers, and some useful linguistic properties common to technical terminology have been identified. Moreover, recent advance in statistical NLP techniques allows the extraction of compound terms at a practical level of accuracy. By contrast, semantic categorization of technical terms have attracted fewer researchers, mainly because the task is more involved and requires extensive expert knowledge to correctly evaluate the results.

In this paper, we construct an experimental system for identifying the semantic class of biological and medical terms using the state-of-the-art NLP and machine learning techniques. Our objectives with this system is not only to evaluate the applicability of these techniques, but also to examine the effectiveness of new features extracted from syntactic dependency structure within sentences. Several other features are considered as well, such as the spelling of the terms and the words occurring around the terms. We also exploit publicly available resources as much as possible to avoid costly annotation of corpora by human experts.

2 Background

Although there have been some earlier attempts (e.g., [3]) using handcrafted patterns and rules to identify the class of technical terms, the non-negligible

cost of constructing and maintaining such patterns has since shifted the focus of the research area to automatically acquiring the classification rules from large annotated corpora using supervised machine learning methods. The latter approach assumes that most of the relevant features are domain-independent, yet class- and domain-specific characteristics can be automatically extracted from these features. The work along this approach includes Collier et al. [1], Gouhara et al. [4] and Yamada et al. [12], as well as the present paper.

There are three factors dominating the performance of terminological classifiers constructed with the supervised learning approach: (1) the size and quality of training corpora, (2) the choice of the leaning algorithm, and (3) the choice of the features used for classification. Below, we review how these factors have been addressed in the literature, as well as our own approach.

2.1 Size and quality of training corpora

Previous work in the area [1, 4, 12] used a relatively small number of examples due to the difficulty in constructing a large corpus of text with high-quality annotations. The corpora used in the work consisted merely of 35–100 abstracts containing 1500–3300 technical terms. Moreover, Yamada et al., who employed two human experts to annotate the same set of abstracts in the MEDLINE database to evaluate the quality of the corpus, observed about 20% disagreement rate of the annotated tags between the two annotators[1]. Part of this disagreement comes from large cross-over in vocabulary of each semantic classes, yet it reveals that the classification task is non-trivial even for human experts.

The size and the quality of annotated corpora are thus non-negligible practical factors for supervised learning approach. In this paper, the difficulty of constructing a training corpus is alleviated by the use of existing thesaurus.

2.2 Learning algorithms

Several machine learning algorithms have been applied for terminology classification. Collier et al. [1] used Hidden Markov Models; Gouhara et al. [4] used decision trees with co-training; and most recently, Yamada et al. [12] used Support Vector Machines (SVMs) to deal with high-dimensional feature space incurred by the use of abundant information on spelling, parts-of-speech, and substrings. Compared with Yamada et al., the former two researchers used smaller number of features, due to the limited scalability of the learning algorithms used.

Following Yamada et al., the present paper uses SVMs, which are known to perform well in the presence of many features as in our formulation of the problem.

2.3 Choice of the features

Both phrase-internal information and extra-phrase, or *contextual*, information has been used for classification of technical terminology. Phrase-internal information includes features such as character types and parts-of-speech of constituent words. The effectiveness of these features has been demonstrated in [1] and [12]. As to the contextual features, use of bigram or trigram sequence of words surrounding the terms is popular. However, fixed-length sequences are problematic in that how far we should look beyond its surroundings are actually situation-dependent. For instance, Sentences (1) and (2) below, both retrieved from MEDLINE database, are the examples in which the bigram word sequence fails to capture words that could possibly help in determining the class of terms.

Both the azide-insensitive and azide-sensitive components of F1-ATPase activity are equally inhibited by *labelling the enzyme with 7-chloro-4-nitrobenzofurazan*, by binding the natural inhibitor protein, or by cold denaturation of the enzyme.	(1)

Results suggest that E. chaffeensis infections are common in free-ranging coyotes in Oklahoma and that these wild canids could play a role in the *epidemiology of human monocytotropic ehrlichiosis*.	(2)

In Sentence (1), the bigram word sequence feature conveys only the information on two words preceding the term “7-chloro-4-nitrobenzofurazan,” namely, “enzyme” and “with.” They hardly avail to elicit the relationship between the term and “enzyme” because the information on verb “labeling” is missing. Similarly, in Sentence (2), there are three words between the term “ehrlichiosis” and the key word “epidemiology” which strongly suggests that the term is the name

[1] A similar disagreement rate has also been reported by Tateisi et al. [10]

of a disease.

Making sequence length larger (e.g., 4) solves the problem in the above examples, but it does not come without cost; it would indeed provide the classifier with richer information, but it would also result in data sparseness in a high dimensional feature space, making learning with a small number of examples extremely difficult. It is hence desirable to use a context feature more adaptive and flexible than fixed-length sequences.

3 Syntactic dependency structure as a feature for classification

One way to overcome the inflexibility of fixed-length context features is to utilize the dependency structure of words within a sentence. It allows us to make selective use of information on distant words, without making the feature space too sparse. Such a structure can be detected in multiple ways, but in this paper we extract it from the parse trees of sentences. We will sketch how this is done with an illustration in Figure 1, which depicts a partial parse tree near the occurrence of "7-chloro-4-nitrobenzofurazan" in Sentence (1).

In the parse tree, each parent-child relation signifies an application of a context-free production rule of the form $X \rightarrow Y_1, \ldots, Y_n$, where X is a non-terminal symbol (denoting its syntactic categories such as NP, VP and PP) of the parent node, and $Y_1, \ldots, Y_n$ are the symbols of the children. A node is labeled not only with a symbol, but also with a *head word*. For a terminal node, it is the lexical entry of the node (shown in italics in the figure); for a non-terminal node, it is inherited from one of its children (shown in parentheses). If a node X has two or more children $Y_1, \ldots, Y_n$, $n \geq 2$, the so-called "head rule[2]" associated with the production rule $X \rightarrow Y_1, \ldots, Y_n$ determines a child (called *head constituent*) Y_i from which X inherits the head word. In the figure, bold arrows depict how head words are inherited; e.g., the bold arrow from NN to PP shows that NN is the head constituent[3] of production rule $\mathrm{PP} \rightarrow \mathrm{IN}, \mathrm{NN}$.

When a parse tree is available, dependency structure can be extracted by recursively merging every head constituent node with its parent (i.e., by merging every parent-child pair connected with bold arrow in the figure), and marking the merged node with the same label as the head constituent. Then, in the resulting tree, a parent-child pair denotes a dependency of the head word of the child on that of the parent.

Applying this procedure to the tree in Figure 1, we can see that the preposition "with" depends on "7-chrolo-4-nitrobenzofurazan," the determinant "the" depends on "enzyme," and both "7-chrolo-4-nitrobenzofurazan" and "enzyme" depend on "labelling." Hence, by collecting the words that depend on and those depended by the term of interest, we can extract dependency information relevant to the term. This allows us, for instance, using the verb "labelling" as a feature for "7-chrolo-4-nitrobenzofurazan" in Sentence (1); and in Sentence (2), since "epidemiology" is the head word of the noun phrase containing the term "ehrlichiosis," the dependency of the term on "epidemiology" can be extracted with this procedure as well.

4 Experiments

To evaluate the effectiveness of the features obtained from the dependency information extracted from parse trees, we applied it along with other features to the task of identifying the semantic classes of technical terms in the MEDLINE abstracts.

4.1 Experimental setting

The experimental setting is described below.

Classes The target semantic classes were determined in accordance with the first five top-level nodes of the 2002 MeSH Tree. They are (A) Anatomy, (B) Organisms, (C) Diseases, (D) Chemicals and Drugs, and (E) Analytical, Diagnostic and Therapeutic Techniques and Equipments.

Data Sets A corpus of abstracts was obtained in the following way. First, 15000 terms from the above classes in the MeSH Tree were randomly sampled. Next, using these terms as query keywords, we retrieved 216404 abstracts from MEDLINE, and then resampled 1200 abstracts for each class from the set. Removing duplicates from the resampled collection resulted in a corpus of 5842 distinct abstracts.

In the corpus, 7531 terms belonging to exactly one of the classes (A) to (E) were identified and used as

[2]We used a slightly modified version of the head rules used by Collins [2].

[3]This is one of the major modification we made to the head rules found in [2], in which IN instead of NN is the head consistuent of the production $\mathrm{PP} \rightarrow \mathrm{IN}, \mathrm{NN}$.

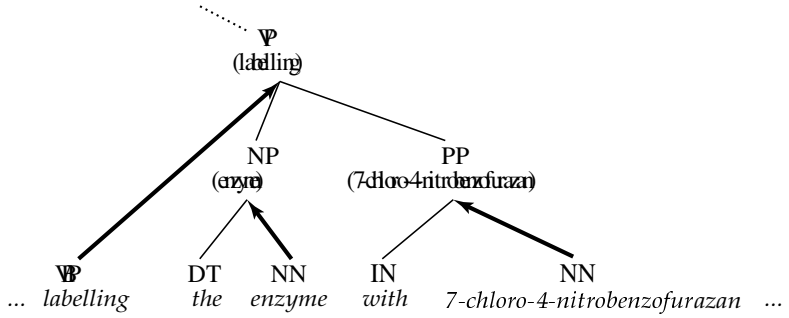

Figure 1: Parse tree of a verb phrase containing "7-chloro-4-nitrobenzofurazen." Bold arrows signify head word inheritance, and parenthesized words the inherited head words.

Table 1: Number of examples for each class

Class	# of examples
A	4571
B	2811
C	5004
D	7335
E	4101
Total	23822

the target terms. This yielded a total of 23822 distinct examples. The number of examples for each class is shown in Table 1.

Features The types of features used by the classifiers were as follows. In addition to the ones using only one of these feature sets, we constructed the classifiers with various combinations of the feature sets as well.

- *Suffix features* — the suffix strings of the head words of target terms. The head word of a target term is determined by the same head rule as described in Section 3. We used the suffixes of lengths 3 and 4.
- *Bigram word sequence features* — the surface and the parts-of-speech (POS) of words in the bigram sequences preceding and succeeding target terms. To obtain the POS, every sentence in the corpus containing one or more technical terms was fed to Nakagawa et al.'s POS tagger[4] [8].
- *Dependency features* — the words on which a target term depends, and the words which the term is depended on, together with their corresponding POS. To obtain these features, the output of the POS tagger was further fed to Yamada and Matsumoto's bottom-up parser [13], under the constraint that technical terms occuring in the sentence should be labeled as either NN (noun) or NP (noun phrase)[5]. The output parse trees were then used for extracting above features with the method of Section 3.

Algorithm Given a set of examples and a combination of features, we constructed an SVMs for each class (A) to (E). The examples whose target terms fall into other four classes were used as negative instances. In all cases, the SVMs used a linear kernel with a fixed soft margin parameter of $C = 1$.

Evaluation We conducted five-fold cross validation with the data set. The examples were partitioned into five sets so that no target terms appear in two sets, and so that each set contains a nearly equal number of distinct target terms. This partitioning scheme avoids a term to appear both in training and test sets during cross validation; since we make use of spelling (suffix) information as features, simply partitioning examples into five sets of equal size at random would make the problem much easier. As a result, the number of examples (occurrences of terms) in each set is not uniform, because some of these terms occur more than once in the abstracts. Table 2 shows the numbers of terms in each set.

[4]The POS tagger performs well even in the presence of unknown words, with the accuracy of 87% for unknown words, and 96% overall in the Penn TreeBank [6].

[5]This constraint reflects the observation by Justeson and Katz [5] that a vast majority of the occurrences of technical terms are noun phrases.

Table 2: Number of terms and examples in each cross-validation set.

Set ID	# of terms	# of examples
1	1507	5236 (22.0%)
2	1506	4361 (18.3%)
3	1506	4844 (20.3%)
4	1506	4715 (19.8%)
5	1506	4666 (19.6%)
Mean	1506.2	4764.4 (20.0%)
Total	7531	23822 (100.0%)

4.2 Results

Under the setting described above, two experiments were conducted.

The first experiment compares the performance of the types of contextual features. Table 3 shows the performance of two classifiers, each using only one of the dependency or bigram word sequence features. The result clearly shows the superiority of dependency information over bigram sequences.

In the next experiment, we combined the contextual features with the phrase-internal suffix features. The performance of classifiers with various feature combinations is listed in Table 4. As a base line, the performance of the classifier using only the suffix features is also included in the table.

The classifier using all of the dependency, bigram sequence and suffix features performed best, but was only slightly ahead of the one with dependency and suffixes. Both of these outperformed the classifiers not using dependency information in most of the classes. Even in a few cases in which the latter surpassed the former, the difference was not significant. However, the performance advantage of dependency over bigram sequences was much smaller than the one observed in the previous experiment in which these features were used alone.

5 Summary and future directions

We have constructed a system for terminological classification in biological and medical papers. Motivated by practical considerations, the system takes advantage of state-of-the-art natural language processing and machine learning techniques, as well as publicly available resources. We have further evaluated the performance of the system over different set of features. Although more thorough experiments are desirable, the experimental results of Section 4 suggest the effectiveness of syntactic dependency information as features for classification.

The future research directions include:

- Classification into more detailed sub-categories. We used only the descriptors on the top-level nodes of the MeSH Tree Structure as semantic categories. It should be necessary to evaluate the performance of our methods in the tasks of classification into more detailed sub-categories.

- Measurement of performance in disambiguating multi-class terms. We trained classifiers only with terms whose class could be uniquely determined according to the MeSH Tree, and excluded multi-class terms from consideration. It would be interesting to apply the classifier trained this way to disambiguate the meaning of each occurrence of the multi-class terms in the corpus.

- Utility of information on multiple occurrences of terms. Justeson and Katz argued that when an entity is referred to by a terminological noun phrase and is rementioned subsequently, it is more likely that the full noun phrase is used intact. This property suggests that when a term is used more than once within an abstract, it is likely that the referent entity and hence its semantic class is unique in the abstract. Collier et al. [1] report that an improvement of 2.3% in F-score was achieved by a similar post-processing.

Acknowledgments. We are grateful to Taku Kudo and Tetsuji Nakagawa for providing us with their part-of-speech tagger and SVM programs. This research was supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan under Grant-in-Aid for Scientific Research on Priority Areas (B) no. 759.

References

[1] N. Collier, C. Nobata, and J. Tsujii. Extracting the names of gens and gene products with a Hidden Markov Model. In *Proceedings of COLING'2000*, pages 201–207, 2000.

[2] M. Collins. *Head-driven statistical models for natural language processing*. Phd dissertation, University of Pennsylvania, 1999.

Table 3: Performance with different contextual information. All the numbers are means over five cross validation trials.

	Dependency			Bigram Sequence		
Class	Precision	Recall	F-score	Precision	Recall	F-score
A	**0.802**	**0.464**	**0.587**	0.553	0.088	0.152
B	**0.792**	**0.324**	**0.459**	0.696	0.197	0.306
C	**0.829**	**0.522**	**0.640**	0.662	0.199	0.306
D	**0.767**	**0.456**	**0.571**	0.644	0.253	0.362
E	**0.739**	**0.369**	**0.491**	0.564	0.165	0.254

Table 4: Performance with various feature combinations. P: precision, R: recall, F: F-score

	Dependency + Sequence + Suffix			Dependency + Suffix			Sequence + Suffix			Suffix only		
Class	P	R	F	P	R	F	P	R	F	P	R	F
A	0.914	**0.707**	**0.794**	0.913	0.703	0.791	0.912	0.703	0.790	**0.916**	0.693	0.786
B	**0.878**	**0.548**	**0.674**	0.787	0.546	0.640	0.842	0.545	0.658	0.792	0.516	0.619
C	0.916	0.841	0.876	**0.920**	0.839	**0.877**	0.906	**0.842**	0.872	0.906	0.829	0.864
D	**0.857**	0.848	**0.851**	0.837	**0.861**	0.848	0.849	0.850	0.848	0.819	0.838	0.827
E	**0.895**	0.685	0.771	0.867	**0.693**	0.766	0.887	0.690	**0.772**	0.853	0.700	0.765

[3] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. Towards information extraction: identifying protein names from biological papers. In *Proceedings of the Third Pacific Symposium on Biocomputing (PSB'98),*, pages 707–718, Maui, Hawaii, USA, 1998.

[4] H. Gouhara, T. Miyata, and Y. Matsumoto. Extraction and classification of medical and biological technical terms. IPSJ SIG Note 2000-NL-135-6, Information Processing Society of Japan, 2000. In Japanese.

[5] J. S. Justeson and S. M. Katz. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27, 1995.

[6] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1994.

[7] D. Maynard and S. Ananiadou. TRUCKS: a model for automatic multi-word term recognition. *Journal of Natural Language Processing*, 8(1):101–125, 2001.

[8] T. Nakagawa, T. Kudo, and Y. Matsumoto. Unknown word guessing with support vector machines. IPSJ SIG Note NL-141-13, Information Processing Society of Japan, 2001. In Japanese.

[9] K.-Y. Su, M.-W. Wu, and J.-S. Chang. A corpus-based approach to automatic compound extraction. In *Proceedings of the 32nd Annual Meetings of the Association for Computational Linguistics*, 1994.

[10] Y. Tateisi, T. Ohta, N. Collier, C. Nobata, and J. Tsujii. Building annotated corpus from biomedical research papers. In *Proceedings of the COLING'2000 Workshop on Semantic Annotation and Intelligent Content*, pages 28–34, 2000.

[11] U.S. National Library of Medicine. MeSH: Medical Subject Headings. http://www.ncbi.nlm.nih.gov/mesh/.

[12] H. Yamada, T. Kudo, and Y. Matsumoto. Using substrings for technical term extraction and classification. IPSJ SIG Note NL-140-11, Information Processing Society of Japan, 2000. In Japanese.

[13] H. Yamada and Y. Matsumoto. Deterministic bottom-up parsing with support vector machines. IPSJ SIG Notes 2002-NL-149, Information Processing Society of Japan, 2002. In Japanese (To appear).