Automatic Resolution of Ambiguous Abbreviations in Biomedical Texts using Support Vector Machines and One Sense Per Discourse Hypothesis

Zhonghua YU
Department of Computer Science, University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo,
113-0033, Japan
Department of Computer Science, Sichuan University, Chengdu,
610064, China
zhyu@ is.s.u-tokyo.ac.jp

Yoshimasa TSURUOKA
CREST, JST (Japan Science and Technology Corporation)
Department of Computer Science, University of Tokyo,
Hongo 7-3-1, Bunkyo-ku, Tokyo,
113-0033, Japan
tsuruoka@logos.t.u-tokyo.ac.jp

Jun'ichi TSUJII
Department of Computer Science, University of Tokyo,
Hongo 7-3-1, Bunkyo-ku, Tokyo,
113-0033, Japan
CREST, JST (Japan Science and Technology Corporation)
tsujii@ is.s.u-tokyo.ac.jp

ABSTRACT
We present an algorithm to disambiguate abbreviations in Medline abstracts using Support Vector Machines (SVM) and *one sense per discourse* hypothesis. In contrast to other work using SVM for natural language disambiguation which always depend on handcrafted training and testing data, the algorithm provided here automatically extracts the training and testing data through searching long form of abbreviation in the texts and using one sense per discourse hypothesis. In the phase of testing, we also use this hypothesis to unify the outputs of the classifier via majority voting. The results obtained in our experiments demonstrate that SVM is a promising technique for abbreviation disambiguation and using majority voting in the phase of testing can improve the accuracy from 82.35% to 84.31%.

Keywords
Abbreviation Disambiguation, Support Vector Machines, One Sense Per Discourse

1. INTRODUCTION
With the widespread use of computers in the biomedical area a vast amount of data with different forms is being generated. Extracting valuable information from the data become more and more important.

Nowadays various approaches to automatically extract information (knowledge) from the data are provided and developed. For example, Data Mining and Knowledge Discovery from Database (DM and KDD) methods find their own applications when extracting information from structured data, and at the same time for semi structured data such as Web Pages the methods based on structure of the pages are proposed.

Natural language is one of the most important forms for information representation in a variety of areas, particularly in biomedicine. For example, in the biomedical area a natural language text collection named Medline contains over 11-million citations indexing major medical, biological, and other life sciences journals (Medline, 2003), and the citation information often contains the abstracts of citied articles. Therefore, extracting information from biomedical texts attracts attention of many researchers from the areas of natural language processing and biomedicine. The concerned problems include named entities recognition, extracting protein-protein relations, and abbreviation-expansion pair extraction.

In biomedicine many named entities are expressed by abbreviation, therefore understanding abbreviation is a necessary component for information extraction from biomedical texts. However, understanding abbreviation is a difficult task because abbreviations in the area are highly ambiguous (Liu et al., 2002a) and a few are paired with their own long form in the texts (Yu et al., 2002).

In this paper our focus is on the problem of abbreviation disambiguation in medical texts. Like in (Chang et al., 2002), here abbreviation is defined as a string that is a shortened form of a sequence of words. The sequence of words is called *long form* of abbreviation, and abbreviation disambiguation means to select the correct long form for every abbreviation occurrence in text depending on context of the occurrence. We provide a machine learning method for resolving ambiguous abbreviations in Medline abstracts. The method is based on SVM (Support Vector Machines) and one sense per discourse hypothesis (Gale et al., 1992). This paper is organized in the following manner: the approaches to automatically generate training and testing data are described in section 2, testing conditions and the obtained result are presented in section 3, discussion about related works are in section 4, and finally future work and conclusions are provided in section 5.

2. METHOD

2.1 Support Vector Machines (SVM)

SVM is a state of the art supervised machine-learning technique proposed by Vapnik et al. (Cortes and Vapnik, 1995) and is based on Structured Risk Minimum Principle. By the principle, when training a classification model, the aim of the learner is to optimize not just the error rate on the training data set, but also the ability of the model for prediction, and the ability depends on concept VC-dimension. Following the Structured Risk Minimum Principle, training a SVM is summed up as finding optimal classifying hyperplane that has the largest margin. The margin is defined as the distance from the hyper plane to the closest training examples. The SVM is being applied in many areas such as text classification (Joachims, 2001), word sense disambiguation (Cabezas et al., 2001), and has showed many advantages over the other supervised machine-learning methods.

2.2 One long form Per Abstract for an Abbreviation Hypothesis

One Sense Per Discourse hypothesis was introduced by Gale, Church and Yarowsky. In (Gale et al., 1992) Gale, Church and Yarowsky reported that 94% occurrences of ambiguous words from the same discourse have the same meaning. For abbreviation, analogically, when considering its sense as its long form, we can observe and assume that when an abbreviation has different occurrences within an abstract of the Medline, all of the occurrences have the same long form. We can call this hypothesis one long form per abstract for an abbreviation.

2.3 Majority Voting for One Sense Per Discourse

In this paper one sense per discourse hypothesis is used not only in the training phase, as in the previous work (Liu et al., 2002b), but also in the testing phase. By the hypothesis, all the occurences of an abbreviation in an abstract must have the same long form, however, it is possible that the SVM classifier predicts different long forms for the ocuurences of an abbreviation. Therefore, a postprocessor is needed to unify the outputs returned by the SVM classifier. In other words, all the outputs for a particular abbreviation within an abstract must be the same. For this purpose we adopt a majority voting method for imposing one sense per discourse constraint. It chooses the majority long form as the long form for all the occurrences of the abbreviation. If there is tie, the process of correction is canceled for the abbreviation and the prediction produced by each SVM is used.

2.4 Automatically extracting training and testing data

Like other supervised machine learning algorithms, SVMs require a labeled data set as its training data. For abbreviation disambiguation task, in particular, the data set contains vectors and each vector is a description of an abbreviation occurrence. The vector has the form ($Feature_1$, $Feature_2$, …, $Feature_n$, Label), where Label represents which long form is being used in this occurrence of the abbreviation, and the $Feature_1$, $Feature_2$, …, $Feature_n$ describe the context of the abbreviation occurrence.

When an abbreviation is not well-known in the area, its occurrences are always paired with its long forms. This characteristic makes annotation of the labeled data set easy, because we can think an occurrence of a long form as an occurrence of the abbreviation with the long form sense. And using the One Long Form per Abstract hypothesis, we can further increase the data set assuming all the occurrences of the abbreviation from the same abstract have the same long form. The features we have used here, as in many of other related works, are local context, specifically, left 2 words and right 2 words from each of abbreviation occurrence. The algorithm proposed here to automatically extract training and testing data is given in Figure 1.

3. TESTING and RESULTS

3.1 Abbreviations and Long Forms disambiguated

For comparing performance of the method provided here with previous work (Liu et al., 2002b; Pakhomov, 2002), we have chosen two groups of abbreviations to disambiguate. These abbreviations and their long forms are given in Table 1 (Set A, chosen from (Pakhomov, 2002)) and Table 2 (Set B, chosen from (Liu et al., 2002b)).

3.2 Training and testing data

The data used for experimentation in this paper comes from Medline (Medline, 2003). We downloaded and collected abstracts from Medline by querying for the long forms of the abbreviations to be disambiguated. A set of vectors is automatically extracted from the collected abstracts through our algorithm described in Figure 1. An example of automatically extracted vector is shown in Figure 2, and the numbers of the vectors extracted for each of the abbreviations are given in Table 1 and Table 2.

The extracted set of the vectors for each abbreviation is split into training and testing data sets at random in the 80%-20% fashion. The training data set is used to train a SVM model and the testing data set is for examining performance of the trained SVM. We used LIBSVM package for our purpose because it provides C++ source code and supports multi-class setting. In the experiment we have chosen RBF kernel, and the parameter γ was set to 0.001. For further details of LIBSVM, see http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

3.3 Experiments using Words Plus Position Features

In this experiment before extracting words in the context all stop words and symbols in the sentence were filtered out because stop words and symbols are not important to disambiguate abbreviations. And the features used were the neighboring words with their positions in the context. This means that for word "gene", there may be four different features L_2=gene, L_1=gene, R_1=gene and R_2=gene depending on its relative position to the abbreviation. The feature values used were binary. Table 3 is the accuracy of the experiment for the abbreviations in Set A in this condition. The results show that the long forms for the abbreviations in Set A have been predicted correctly 82.20% on average.

3.4 Experiments using Bag-of-Word Features

As in the experiment using word plus position features, before extracting words in the context all stop words and symbols in the

Input:

(1) A set of Medline abstracts (SMA)

(2) Abbreviation dictionary containing the abbreviations and their long forms disambiguated (AD)

Output:

Set of vectors in form (Feature$_1$, Feature$_2$, …, Feature$_n$, A, LF), where A is an abbreviation, and LF is one of the long forms of the abbreviation A (every vector represents an occurrence of the abbreviation A in context (Feature$_1$, Feature$_2$, …, Feature$_n$,) and the abbreviation A in the context has the sense (long form) LF)

Algorithm:

```
FOR (X ∈SMA)  DO   //X is an abstract in the set SMA
{   FOR (A ∈AD)  DO
   //A is an abbreviation in the dictionary AD
    { IF ((a long forms of abbreviation A is found in X)
        AND  (the found long form is LF)
        AND (no other long form of A is also found in X))
      {   FOR (each of the occurrences of A in X) DO
        {  Generate vector (Feature1, Feature2, …, Featuren,
                                    A, LF),
           where Feature1, Feature2, …, Featuren, is the
           context of the occurrence;
        }
      }
}
```

Figure 1. Algorithm to extract training and testing data

sentence were filtered out. But the features used are the neighboring words without the information of their positions in the context (bag of words). This means that for word "gene", there may be only one feature "gene". The feature values used were also binary. Table 4 is the accuracy of the experiment for the abbreviations in Set A in this condition. On average the long forms for the abbreviations in Set A have been predicted correctly 82.35%, which is about the same as in experiment using word plus position feature.

3.5 Experiments using Majority Voting Method

As in the experiments mentioned above, before extracting words in the context all stop words and symbols in the sentence were filtered out. And the features used were extracted words without their positions (bag of words) in the context. The feature values were also binary. But when split the data set into training and testing data set at random, the split unit was every abstract, not vector, as done in the first two experiments. Therefore, all vectors of an abstract were not split into the different data sets (training data set and testing data set). After SVM prediction, majority voting for one sense per discourse is used to check and correct the prediction produced by the SVM classifier. The results obtained in this setting is given in Table 5 for Set B and Table 6 for Set A. The results in Table 6 show that, on average the long forms for the abbreviations in Set A have been predicted correctly 84.31%, which is considerably higher than the results obtained in the experiment using bag of words features and the experiment using word plus position features for the same abbreviations.

Table 1. Set A - Abbreviations and long forms (chosen from (Pakhomov, 2002)) and number of vectors extracted

Abbreviation	Long Form (#Vectors)
INF	infective (916), infant (1567), inferior (3332), interferon (3388), infusion (4636), infected (4956), infection (5248)
PN	periarteritis nodosa (3640), positional nystagmus (457), parenteral nutrition (3808), polyneuritis (967), pyelonephritis (4156), polyneuropathy (2387), peripheral nerve (1939), peripheral neuropathy (877), polyarteritis nodosa (6337), pneumonia (6492), penicillin (3540)
BD	bundle (4872), twice a day(1158), band (6260)
PA	periarteritis (3086), plasma aldosterone (5404), pantothenic acid (903), physician assistant (359), pernicious anemia (1473), paranoia (883), pyruvic acid (17), panic attack (492), paternal aunt (35), procainamide (3322), pulmonary artery (285), pathology (3865), polyarthritis (236), pseudomonas aeruginosa (368), polyarteritis (5584)
NR	nonresponder (2019), nonreactive (2940), no report (57), no response (407), nerve root (1965), nurse (3975), no refill (3), no recurrence (336), no radiation (19), normal range (1688)
RA	right atrial (3923), rheumatic arthritis (162), refractory anemia (2661), right atrium (2153), right arm (1822), radioactive (3139), renal artery (4358), rheumatoid arthritis (11305)

From the results, we can obtain the following conclusions:

(1) Positions of the feature words are not very important for the abbreviation disambiguation. This may be caused by data sparse specific for natural language texts.

(2) Using hypothesis that all occurrences of an abbreviation within an abstract have the same long form can improve accuracy of the abbreviation disambiguation. But because of short size of abstract, occurrences of an abbreviation within an abstract is little; therefore the improvement is not very obvious. If the method proposed here is used to disambiguate abbreviations in the journal papers, the accuracy will be better because of the more occurrences of an abbreviation in a paper.

Table 2. Set B - Abbreviations and long forms (chosen from (Liu et al., 2002b)) and number of vectors extracted

Abbreviation	Long Form (#Vectors)
ACE	antegrade colonic enema (44), adrenocortial extract (0), amsacrine cytarabine etoposide (2), doxorubicin cyclophosphamide etoposide (12), doxorubicin cyclophosphamide (113), angiotensin converting enzyme (1131), acetylcholinesterase (805)
APC	activated protein c (875), aphidicholin (12), atrial premature complexes (44), adenomatous polyposis coli (1008), antigen presenting cells (412)
ASP	aspartate (1023), aspartylglycine (73), aspartic acid (238), asparaginase (1103), antisocial personality (379), ankylosing spondylitis (752)
BSA	bovine serum albumin (969), body surface area (723)
CSF	competence and sporulation factor (0), cytostatic factor (387), colony stimulating factors (66), cerebrospinal fluid (1331)
EMG	electromyogram (1134), electromyographs (85), electromyography (365), exomphalos macroglossia gigantism (45)
IBD	irritable bowel syndrome (678), inflammatory bowel diseases (289)
MAS	macandrew alcoholism scale (86), meconium aspiration syndrome (685), mccune albright syndrome (550)
RSV	rous sarcoma virus (857), respiratory syncytial virus (2089)
VCR	vanadyl ribonucleoside complex (6), videocassette recorder (22), vincristine (950)

4. RELATED WORK

Liu et al. provided an algorithm to automatically extract training and testing data from Medline abstracts for abbreviation disambiguation in the biomedical area (Liu et al., 2002b). The algorithm is based on semantic relationships in UMLS (UMLS, 2000) and one sense per discourse hypothesis. To compare our algorithm with the algorithm provided in (Liu et al., 2002b) we have chosen some abbreviations for our experiment (see table 2). We have not chosen the other abbreviations because there are not long forms for the other abbreviations in (Liu et al., 2002b), and the long forms of abbreviation PVC given in (Liu et al., 2002b) are not found in Medline through querying for the long forms. Comparing our algorithm with the algorithm proposed in (Liu et al., 2002b), we have the following discussions about the work in (Liu et al., 2002b):

(1) The task of abbreviation disambiguation is thought as a classification problem, and a supervised classification algorithm is used to predict a probable long form of an abbreviation, this is the same as the method provided here;

Marked bradykinin-induced tissue plasminogen activator release in patients with heart failure maintained on long-term angiotensin-converting enzyme inhibitor therapy.
OBJECTIVES: The aim of the present study was to assess the contribution of angiotensin-converting enzyme (ACE) inhibitor therapy to bradykinin-induced tissue-type plasminogen activator (t-PA) release in patients with heart failure (HF) secondary to ischemic heart disease. BACKGROUND: Bradykinin is a potent endothelial cell stimulant that causes vasodilatation and t-PA release. In large-scale clinical trials, ACE inhibitor therapy prevents ischemic events.
METHODS: Nine patients with symptomatic HF were evaluated on two occasions: during and following seven-day withdrawal of long-term ACE inhibitor therapy. Forearm blood flow was measured using bilateral venous occlusion plethysmography. Intrabrachial bradykinin (30 to 300 pmol/min), substance P (2 to 8 pmol/min), and sodium nitroprusside (1 to 4 pmol/min) were infused, and venous blood samples were withdrawn from both forearms for estimation of fibrinolytic variables. RESULTS: On both study days, bradykinin and substance P caused dose-dependent vasodilatation and release of t-PA from the infused forearm ($p < 0.05$ by analysis of variance [ANOVA]). Long-term ACE inhibitor therapy caused an increase in forearm vasodilatation ($p < 0.05$ by ANOVA) and t-PA release ($p < 0.001$ by ANOVA) during bradykinin, but not substance P, infusion. Maximal local plasma t-PA activity concentrations approached 100 IU/ml, and maximal forearm protein release was approximately 4.5 microg/min.
CONCLUSIONS: Long-term ACE inhibitor therapy augments bradykinin-induced peripheral vasodilatation and local t-PA release in patients with HF due to ischemic heart disease. Local plasma t-PA activity concentrations approached those seen during systemic thrombolytic therapy for acute myocardial infarction. This may contribute to the primary mechanism of the anti-ischemic effects associated with long-term ACE inhibitor therapy.

The following are vectors extracted from this abstract for the long form "angiotensin converting enzyme" of the abbreviation "ACE".

L2= contribution, L1= of, R1= inhibitor, R2= therapy

L2= trials, L1=",", R1= inhibitor, R2= therapy

L2= of, L1= long-term, R1= inhibitor, R2= therapy

L2=":", L1=Long-term, R1=inhibitor, R2=therapy

L2= with, L1= long-term, R1= inhibitor, R2= therapy

Figure 2. Example of automatically extracted vectors using the algorithm described in Figure 1

(2) The supervised classification algorithm is the Naïve Bayes algorithm, but ours is SVM;

(3) Labeled training data is extracted automatically using semantic relationships between concepts obtained from UMLS, and for this reason the algorithm strongly depends on handcrafted knowledge base, but the algorithm proposed here does not depend on any handcrafted information

Table 3. Result of experiment using word plus position features for Set A

Abbreviation	Accuracy
INF	77.28%
PN	79.14%
BD	91.28%
PA	77.24%
NR	84.65%
RA	83.58%
Average	82.20%

Table 4. Result of experiment using bag-of-word features for Set A

Abbreviation	Accuracy
INF	76.58%
PN	79.11%
BD	90.48%
PA	78.50%
NR	85.61%
RA	83.84%
Average	82.35%

resource, it extracts the data using long forms of the abbreviation disambiguated;

(4) One sense per discourse hypothesis is used to extract labeled data, as in our method, however, in the testing phase the hypothesis is not used; on the contrary, our method applies the hypothesis to correct the output produced by SVM;

(5) The accuracy for the abbreviations in Set B is 79.17%, when remove rare senses the accuracy is 83.93%, our algorithm attains 87.47% accuracy for the same abbreviations and long forms without any removing rare senses.

Serguei Pakhomov in (Pakhomov, 2002) provided a Maximum Entropy based supervised classifying method to disambiguate abbreviations in clinical notes. The abbreviations disambiguated by him and their long forms are as in table 1 (Set A). The training data and testing data are extracted automatically from the clinical notes through searching the long forms in the clinical notes, but one sense per discourse hypothesis is not considered. The obtained accuracy is 89.14%, which is higher than ours for the same abbreviations and long forms, however, because the clinical notes have different literary characteristic from the Medline abstracts, it is difficult to compare the performances between methods proposed in (Pakhomov, 2002) and in this paper.

5. CONCLUSIONS AND FUTURE WORK

We developed an algorithm to disambiguate abbreviations in Medline abstracts based on SVM and one sense per discourse hypothesis. The one sense per discourse hypothesis is used not only to automatically extract training and testing data, but also to unify the outputs of the SVM classifier by a majority voting method. The results obtained in our experiments have showed that

Table 5. Result of majority voting for Set B

Abbreviation	Accuracy
ACE	81.38%
APC	92.29%
ASP	83.48%
BSA	90.77%
CSF	87.47%
EMG	76.04%
IBD	83.98%
MAS	88.08%
RSV	94.86%
VCR	96.28%
Average	87.47%

Table 6. Result of majority voting for Set A

Abbreviation	Accuracy
INF	79.66%
PN	82.93%
BD	92.37%
PA	78.75%
NR	86.59%
RA	85.56%
Average	84.31%

using one sense per discourse hypothesis in the phase of testing can improve the accuracy about 2%.

We have utilized the hypothesis that all occurrences of an abbreviation within an abstract have the same long form. However, this hypothesis has not been verified. In the future we will check whether one long form per abstract for an abbreviation is tenable. We will also use Maximum Entropy based classifying algorithm to disambiguate abbreviations in the Medline to verify the efficiency of the algorithm in the Medline. The accuracies obtained by us and in (Liu et al., 2002b) are based on local context, we will try to use global context to improve performance of abbreviation disambiguation in the future.

6. REFERENCES

[1] Gale W.A., Church K.W., and Yarowsky D. 1992. One Sense Per Discourse. Proceedings of the ARPA Workshop on Speech and Natural Language Processing. Pages 233-237.

[2] Liu H., Aronson R.A. and Friedman C. 2002a. A study of Abbreviations in MEDLINE Abstracts. American Medical Informatics Association Symposium, San Antonio, TX, November, Pages 9-13.

[3] Yu H., Hripcsak G. and Friedman C. 2002. Mapping Abbreviations to Full Forms in Biomedical Articles. Journal of the American Medical Informatics Association, Vol. 9, No. 3, Pages 262-272.

[4] Cortes C. and Vapnik V. 1995. Support-vector networks. Machine Learning, 20: 273-297, November.

[5] Liu H., Johnson S. B. and Friedman C. 2002b. Automatic Resolution of Ambiguous Terms Based on Machine Learning and Conceptual Relations in the UMLS. Journal of the American Medical Informatics Association, Vol. 9, No. 6, Pages 621-636.

[6] Pakhomov S. 2002. Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July, pages 160-167.

[7] Joachims T. 2001. Learning to classify text using support vector machines. Kluwer Academic Publishers.

[8] Cabezas C., Resnik P. and Stevens J. 2001. Supervised Sense Tagging using Support Vector Machines. Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2), Toulouse, France, July, Pages 5-6.

[9] Medline.2003. http://www.nlm.nih.gov.

[10] UMLS Knowledge Sources, 2000 Edition. 2000. US Dept of Health and Human Services, National Institutes of Health, National Library of Medicine.

[11] Chang J., Scheütze H. and Altman R. 2002. Creating an Online Dictionary of Abbreviations from MEDLINE. Journal of American Medical Informatics Association. Vol. 9, No. 6, Pages 612-620.