Disambiguation of Biomedical Abbreviations

Mark Stevenson[1], **Yikun Guo**[2], **Abdulaziz Al Amri**[3] and **Robert Gaizauskas**[4]
Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello
Sheffield, S1 4DP
United Kingdom
[1,2,4]{initial.surname}@dcs.shef.ac.uk, [3]abdulazizmail@gmail.com

Abstract

Abbreviations are common in biomedical documents and many are ambiguous in the sense that they have several potential expansions. Identifying the correct expansion is necessary for language understanding and important for applications such as document retrieval. Identifying the correct expansion can be viewed as a Word Sense Disambiguation (WSD) problem. A WSD system that uses a variety of knowledge sources, including two types of information specific to the biomedical domain, is also described. This system was tested on a corpus of ambiguous abbreviations, created by automatically identifying the correct expansion in Medline abstracts, and found to identify the correct expansion with up to 99% accuracy.

1 Introduction

Many abbreviations are ambiguous in the sense that they have more than one possible expansion. For example, expansions for "NLP" include "Neuro-linguistic Programming" as well as "Natural Language Processing". Ambiguous abbreviations form a challenge to language understanding since identification of the correct expansion is often important. The query "NLP", for example, returns pages which refer to "Neuro-linguistic programming" for most web search engines, pages which are of limited value to those interested in Natural Language Processing. In some cases this problem could be obviated by altering the query terms, for example including "Natural", "Language" and "Processing". However, this will not help when the abbreviation's expansion does not occur within the document. Fred and Cheng (1999) point out that this is often the case in biomedical documents, in this domain ubiquitous abbreviations (such as DNA and mRNA) often appear without an expansion.

It has been reported that misinterpretation of abbreviations in biomedical documents has lead to medical practitioners making fatal errors (Fred and Cheng, 1999). However, identifying the correct expansion is not a straightforward task since an abbreviation may have several possible expansions. Chang et al. (2002) reported that abbreviations in biomedical journal articles consisting of six characters or less have an average of 4.61 possible meanings and Pustejovsky et al. (2002) mention that the simple abbreviation "AC" is associated with at least 10 strings in different biomedical documents including "atrioventricular connection", "anterior colporrhaphy procedure", "auditory cortex" and "atypical carcinoid".

The problem of identifying the correct expansion of an ambiguous abbreviation can be viewed as a Word Sense Disambiguation (WSD) task where the various expansions are the "senses" of the abbreviation. In this paper we approach the problem in this way by applying a WSD system which has previously been applied to biomedical text (Stevenson et al., 2008). The WSD system uses a variety of information sources, including those traditionally applied to the WSD problem in addition to two knowledge sources that are specific to the biomedical domain.

Evaluation of systems for disambiguating ambiguous abbreviations has been hindered by the fact

Proceedings of the Workshop on BioNLP, pages 71–79,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

that there is no freely available benchmark corpus against which approaches can be compared. We describe a process whereby such a corpus can be created by automatically mining abstracts from Medline. This corpus is being made publicly available to encourage comparative research in this area. Our abbreviation disambiguation system was evaluated against this corpus and found to identify the correct abbreviation with up to 99% accuracy.

The remainder of this paper is organised as follows. The next section describes relevant previous work on disambiguation of abbreviations. Section 3 describes a supervised learning WSD system tailored specifically to the biomedical domain. Section 4 describes the automatic creation of a corpus of ambiguous abbreviations designed specifically for the training and evaluation of abbreviation disambiguation systems. Section 5 describes the evaluation of our system on this corpus. Our conclusions are presented in Section 6.

2 Previous Work

Gaudan et al. (2005) distinguish two types of abbreviations: global and local. *Global abbreviations* are those found in documents without the expansion explicitly stated, while *local abbreviations* are defined in the same document in which the abbreviation occurs. Our work is concerned with the problem of disambiguating global abbreviations. Gaudan et al. (2005) point out that global abbreviations are often ambiguous.

Various researchers have explored the problem of disambiguating global abbreviations in biomedical documents. Liu et al. (2001)(2002) used several domain-specific knowledge sources to identify terms which are semantically related to each possible expansion but which have only one sense themselves. Instances of these terms were identified in a corpus of biomedical journal abstracts and used as training data. Their learning algorithm uses a variety of features including all words in the abstract and collocations of the ambiguous abbreviation. They report an accuracy of 97% on a small set of abbreviations. Liu et al. (2004) present a fully supervised approach. They compared a variety of supervised machine learning algorithms and found that the best performance over a set of 15 ambiguous abbreviations, 98.6%, was obtained using Naive Bayes. Gaudan et al. (2005) use a Support Vector Machine trained on a bag-of-words model and report an accuracy of 98.5%. Yu et al. (2006) experimented with two supervised learning algorithms: Naive Bayes and Support Vector Machines. They extracted a corpus containing examples of 60 abbreviations from a set of biomedical journal articles which was split so that abstracts in which the abbreviations were defined were used as training data and those in which no definition is found as test data. Abbreviations in the test portion were manually disambiguated. They report 79% coverage and 80% precision using a Naive Bayes classifier. Pakhomov (2002) applied a maximum entropy model to identify the meanings of ambiguous abbreviations in 10,000 rheumatology notes with around 89% accuracy. Joshi et al. (2006) disambiguated abbreviations in clinical notes using three supervised learning algorithms (Naive Bayes, decision trees and Support Vector Machines). They used a range of features and found that the best performance was obtained when these were combined. Unfortunately direct comparison of these methods is made difficult by the fact that various researchers have evaluated their approaches on different data sets.

A variety of approaches have also been proposed for the problem of disambiguating local abbreviations in biomedical documents. This task is equivalent to identifying the abbreviation's expansion in the document. The problem is relatively straightforward for abbreviations which are created by selecting the first character from each word in the expansion, such as "angiotensin converting enzyme (ACE)", but is more difficult when this convention is not followed, for example "acetylchlinesterase (ACE)", "antisocial personality (ASP)" and "catalase (CAT)". Okazaki et al. (2008) recently proposed an approach to this problem based on discriminative alignment that has been shown to perform well. However, the most common solutions are based on heuristic approaches, for example Adar (2004) and Zhou et al. (2006). Pustejovsky et al. (2002) used hand-built regular expressions. Schwartz and Hearst (2003) describe an approach which starts by identifying the set of candidate expansions in the same sentence as an abbreviation. The most likely one is identified by searching for the shortest candidate which contains all the characters in the abbreviation in the correct order.

3 Abbreviation Disambiguation System

Our abbreviation disambiguation system is based on a state-of-the-art WSD system that has been adapted to the biomedical domain by augmenting it with additional knowledge sources. The system on which our approach is based (Agirre and Martínez, 2004) participated in the Senseval-3 challenge (Mihalcea et al., 2004) with a performance close to the best system for the lexical sample tasks in two languages while the version adapted to the biomedical domain has achieved the best recorded results (Stevenson et al., 2008) on a standard test set consisting of ambiguous terms (Weeber et al., 2001).

This system is based on a supervised learning approach with features derived from text around the ambiguous word that are domain independent. We refer to these as *general* features. This feature set has been adapted for the disambiguation of biomedical text by adding further linguistic features and two different types of domain-specific features: CUIs (as used by McInnes et al. (2007)) and Medical Subject Heading (MeSH) terms. This set of features is more diverse than have been explored by previous approaches to abbreviation disambiguation.

3.1 Features

Our feature set contains a number of parameters (e.g. thresholds for unigram and CUI frequencies). These parameters were set to the same values that were used when the system was applied to general biomedical terms (Stevenson et al., 2008) since these were found to perform well. We also use the entire abstract as the context of the ambiguous term for relevant features rather than just the sentence containing the term. Effects of altering these variables are consistent with previous results (Liu et al., 2004; Joshi et al., 2005; McInnes et al., 2007) and are not reported here.

General features: The system uses a wide range of domain-independent features that are commonly employed for WSD.

- Local collocations: A total of 41 features which extensively describe the context of the ambiguous word and fall into two main types: (1) bigrams and trigrams containing the ambiguous word constructed from lemmas, word forms or PoS tags and (2) preceding/following lemma/word-form of the content words (adjective, adverb, noun and verb) in the same sentence as the ambiguous abbreviation. For example, consider the sentence below with the target abreviation *BSA*.

  > "Lean BSA was obtained from height and lean body weight ..."

  The features would include the following: left-content-word-lemma "*lean BSA*", right-function-word-lemma "*BSA be*", left-POS "JJ NNP", right-POS "NNP VBD", left-content-word-form "*Lean BSA*", right-function-word-form "*BSA was*", etc.

- Salient bigrams: Salient bigrams within the abstract with high log-likelihood scores, as described by Pedersen (2001).

- Unigrams: Lemmas of all content words in the abstract and words within a ±4-word window around the target word, excluding those in a list of stopwords. In addition, the lemmas of any unigrams appearing at least twice in the entire corpus and which are found in the abstract are also included as features.

Concept Unique Identifiers (CUIs): We follow the approach presented by McInnes et al. (2007) to generate features based on UMLS Concept Unique Identifiers (CUIs). The MetaMap program (Aronson, 2001) identifies all words and terms in a text which could be mapped onto a UMLS CUI. MetaMap does not disambiguate the senses of the concepts, instead it enumerates likely candidate concepts. For example, MetaMap will segment the phrase "Lean BSA was obtained from height and lean body weight ..." into four chunks: "Lean BSA", "obtained", "from height" and "lean body weight". The first chunk will be mapped onto three CUIs: "C1261466: BSA (Body surface area)", "C1511233: BSA (NCI Board of Scientific Advisors)" and "C0036774: BSA (Serum Albumin, Bovine)". The chunk "lean body weight" is mapped onto two concepts: "C0005910: Body Weight" and "C1305866: Body Weight (Weighing patient)"[1]. CUIs occurring more than twice in an abstract are included as features. CUIs have been used for various disambiguation tasks in the biomedical domain, including disambiguation of ambiguous general terms (McInnes et al., 2007) and gene symbol disambiguation (Xu et al., 2007), but not, to our knowledge, for abbreviation disambiguation.

Medical Subject Headings (MeSH): The final feature is also specific to the biomedical domain. Medical Subject Headings (MeSH) (Nelson et al., 2002) is a controlled vocabulary for indexing biomedical and health-related information and documents. MeSH terms are manually assigned to abstracts by human indexers. The latest version of MeSH (2009) contains over 25,000 terms organised into an 11 level hierarchy.

The MeSH terms assigned to the abstract in which each ambiguous word occurs are used as features. For example, the abstract containing our example phrase has been assigned 16 terms including "Body Surface Area", "Body Weight", "Humans" and "Organ Size" . MeSH terms have previously been used for abbreviation disambiguation by Yu et al. (2006).

3.2 Learning Algorithms

We compared three machine leaning algorithms which have previously been shown to be effective for WSD tasks.

The **Vector Space Model (VSM)** is a memory-based learning algorithm which was used by Agirre and Martínez (2004). Each occurrence of an ambiguous word is represented as a binary vector in which each position indicates the occurrence/absence of a feature. A single centroid vector is generated for each sense during training. These centroids are compared with the vectors that represent new examples using the cosine metric to compute similarity. The sense assigned to a new example is that of the closest centroid.

The **Naive Bayes (NB)** classifier is based on a probabilistic model which assumes conditional independence of features given the target classification. It calculates the posterior probability that an instance belongs to a particular class given the prior probabilities of the class and the conditional probability of each feature given the target class.

Support Vector Machines (SVM) have been widely used in classification tasks. SVMs map feature vectors onto a high dimensional space and construct a classifier by searching for the hyperplane that gives the greatest separation between the classes.

We used our own implementation of the Vector Space Model and Weka implementations (Witten and Frank, 2005) of the other two algorithms.

4 Evaluation Corpus

The most common method for generating corpora to train and test WSD systems is to manually annotate instances of ambiguous terms found in text with the appropriate meaning. However, this process is both time-consuming and difficult (Artstein and Poesio, 2008). An alternative to manual tagging is to find a way of automatically creating sense tagged corpora. For the translation of ambiguous English words Ng et al. (2003) made use of the fact that the various senses are often translated differently. For example when "bank" is used in the 'financial institution' sense it is translated to French as "banque" and "bord" when it is used to mean 'edge of river'. However, a disadvantage of this approach is that it relies on the existence of parallel text which may not be available. In the biomedical domain Liu et al. (2001)(2002) created a corpus using unambiguous related terms (see Section 2) although they found that it was not always possible to identify suitable related terms.

4.1 Corpus Creation

Liu et al. (2001) also made use of the fact that when abbreviations are introduced they are often accompanied by their expansion, for example "BSA (bovine serum albumin)". This phenomenon was exploited to automatically generate a corpus of abbreviations and associated definitions by replacing the abbreviation and expansion with the abbreviation alone. For example, the sentence *"The adsorption behavior of bovine serum albumin (BSA) on a Sepharose based hydrophobic interaction support has been studied."* becomes *"The adsorption behav-*

[1] The first of these, C0005910, refers to the weight of a patient as a property of that individual while the second, C1305866, refers to the process of weighing a patient as part of a diagnostic procedure.

```
“BSA” AND “body surface area” NOT “bovine serum albumin”
“BSA” AND “bovine serum albumin” NOT “body surface area”
```

Figure 1: Example queries for abbreviation “BSA”

ior of BSA on a Sepharose based hydrophobic interaction support has been studied.”

We used this approach to create a corpus of sense tagged abbreviations in biomedical documents using a set of 21 three letter abbreviations used in previous research on abbreviation disambiguation (Liu et al., 2001; Liu et al., 2002; Liu et al., 2004). Possible expansions for the majority of these abbreviations were listed in these papers. For the few remaining ones possible expansions were taken from the Medstract database (Pustejovsky et al., 2002). We searched for instances of these abbreviations in Medline, a database containing more than 18 million abstracts from publications in biomedicine and the life sciences. For each abbreviation we queried Medline, using the Entrez interface, to identify documents containing one of its meanings. For example the abbreviation “BSA” has two possible expansions: “body surface area” and “bovine serum albumin”. Medline is searched to identify documents that contain each possible expansion of the abbreviation using the queries shown in Figure 1. Each query matches documents containing the abbreviation and relevant expansion and no mentions of the other possible expansion(s).

The retrieved documents are then processed to remove the expansions of each abbreviation. The Schwartz and Hearst (2003) algorithm for identifying abbreviations and the relevant expansion (see Section 2) is then run over each of the retrieved abstracts to identify the correct expansion. The expansion is removed from the document and stored separately, effectively creating a sense tagged corpus. For convenience the abstracts are converted into a format similar to the one used for the NLM-WSD corpus (Weeber et al., 2001).

The resulting corpus consists of 55,655 documents. For each abbreviation Table 1 shows the number of abstracts retrieved from Medline (in the column labeled “Abstracts”) and the number of expansions (“Count” column). The column labelled “Rare” lists the number of expansions that account for fewer than 1% of the occurrences of an abbreviation and “Frequent” lists the percentage of occurances represented by the most frequent expansion. It can be seen that there is a wide variation between the number of abstracts retrieved for each abbreviation. CSF occurs in 14,871 abstracts and ASP in just 71. There is also a wide variation between the frequency of the most common expansion with over 99% of the occurrences of “CSF” representing one expansion (“cerebrospinal fluid”) while for “ASP” two of the five possible expansions (“antisocial personality” and “aspartate”) each account for almost 34% of the documents. In addition, several abbreviations have expansions which occur only rarely. For example, two of the expansions of “APC” (“atrial pressure complexes” and “aphidicholin”) each have only a single document and account for just 0.03% of the instances of that abbreviation.

4.2 Corpus Reduction

Given the diversity of the abbreviations which were downloaded from Medline, both in terms of number of documents and distribution of senses, subsets of this corpus that are more suitable for WSD experiments were created. Corpora containing 100, 200 and 300 randomly selected examples of each abbreviation were generated and these are referred to as Corpus.100, Corpus.200 and Corpus.300 respectively.

Some of the 21 abbreviations were not suitable for inclusion in these corpora. Abbreviations were not included in the relevant corpus if an insufficient number of examples were retrieved from Medline. For example, only 71 abstracts containing “ASP” were retrieved and it is is not included in any of the three corpora. Similarly, “ANA” and “FDP” are not included in Corpus.200 or Corpus.300 and “DIP” not included in Corpus.300. In addition, rare senses, those which represent fewer than 1% of the occurrences of an abbreviation in all retrieved abstracts, were discarded. Finally, two abbreviations (“ACE” and “CSF”) have only one sense that is not “Rare”

		Expansions		
	Abstracts	Count	Rare	Frequent
ACE	3105	3	2	98.7
ANA	100	3	0	58.0
APC	3146	5	2	39.4
ASP	71	5	0	33.8
BPD	1841	3	0	46.7
BSA	5373	2	0	86.4
CAT	4636	3	1	55.2
CML	2234	4	2	91.7
CMV	7665	2	0	96.7
CSF	14871	3	2	99.1
DIP	209	2	0	75.1
EMG	2052	2	0	88.4
FDP	130	4	0	78.5
LAM	325	4	1	48.3
MAC	955	5	1	64.3
MCP	815	5	1	50.2
PCA	2442	5	1	68.9
PCP	1642	2	0	57.8
PEG	607	2	0	94.1
PVC	234	2	2	78.2
RSV	3202	2	0	76.7
Average	2650	3.2	0.6	70.8

Table 1: Properties of abbreviations corpus retrieved from Medline

(see Table 1) and these were also excluded from the reduced corpora.

Consequently, Corpus.100 contains 18 abbreviations (“ACE”, “ASP” and “CSF” are excluded), Corpus.200 contains 16 (“ANA” and “FDP” are also excluded) and Corpus.300 contains 14 (“DIP” and “PVC” also excluded). Where an abbreviation is included in more than one corpus, all the examples in the smaller corpus are included in the larger one(s). For example, the 100 examples of “APC” in Corpus.100 are also included in Corpus.200 and Corpus.300.

5 Experiments

Various combinations of learning algorithms and features were applied to the three reduced corpora described in Section 4.2. Performance of the WSD system is measured in terms of the proportion of abbreviation instances for which the correct expansion is identified. 10-fold cross validation was used for all experiments and all quoted results refer to the average performance across the 10 folds. Results are shown in Table 2. The baseline figures, based on selecting the most frequent expansion for each abbreviation, are shown for each corpus. Note that these figures vary slightly across the three corpora because of the different abbreviations each contains (see Section 4.2).

A first observation is that performance of the WSD system is consistently better than the baseline for the relevant corpus and, with a few exceptions, above 90%. As might be expected, performance improves as additional training examples are added. However, even when the number of examples is relatively low, just 100, performance of the best configuration (VSM learning algorithm with all three types of feature) is 97.4%.

The best result, 99% (300 training examples, VSM learning algorithm with all feature types), exceeds reported performance of previous abbreviation disambiguation systems (see Section 2). Although these results are not directly comparable, since these studies used different evaluation corpora, the set of ambiguous abbreviations used in this study and methodology for corpus creation are similar to those used by Liu et al. (2001)(2002)(2004).

The best performance for each learning algorithm is obtained when all three types of features are combined. The difference between performance obtained using all three feature types and using only the MeSH or CUI features is statistically significant (Wilcoxon Signed Ranks test, $p < 0.01$) although the difference between this and performance using just the linguistic features is not.

The VSM learning algorithm generally performs better than either the SVM or Naive Bayes learning algorithms. The difference between performance of VSM and the other algorithms is statistically significant for Corpus.100 but not for the other two, suggesting that this learning algorithm is better able to cope with small number of training examples than Naive Bayes and Support Vector Machines. Strong performance of the VSM algorithm is consistent with previous work which has shown that this algorithm performs well on the disambiguation of ambiguous terms in both biomedical and general text (Agirre and Martínez, 2004; Stevenson et al., 2008).

Algorithm	Features						
	Linguistic	CUI	MeSH	Linguistic +CUI	Linguistic +MeSH	CUI+ MeSH	Linguistic+ MeSH+CUI
	Corpus.100 (Baseline = 69.0%)						
SVM	0.934	0.900	0.949	0.947	0.946	0.938	0.954
NB	0.940	0.917	0.949	0.951	0.947	0.944	0.958
VSM	0.968	0.937	0.888	0.970	0.971	0.939	**0.974**
	Corpus.200 (Baseline = 69.1%)						
SVM	0.957	0.911	0.964	0.964	0.964	0.947	0.965
NB	0.966	0.926	0.962	0.969	0.971	0.955	0.972
VSM	0.979	0.930	0.894	0.982	0.981	0.947	**0.984**
	Corpus.300 (Baseline = 68.7%)						
SVM	0.966	0.914	0.970	0.968	0.974	0.954	0.975
NB	0.971	0.933	0.960	0.971	0.976	0.960	0.978
VSM	0.981	0.938	0.894	0.987	0.985	0.957	**0.990**

Table 2: Performance of WSD system using various combinations of learning algorithms and features.

Performance of our system on this task is higher than would be expected for most WSD tasks suggesting that the problem of abbreviation disambiguation is simpler than the disambiguation of general terms. The most probable reason for this is that the various expansions of abbreviations in our corpus are more distinct and better defined than senses for general terms. For example, the three possible expansions for "ANA" in our corpus are a professional body ("American Nurses Association"), a type of medical test ("antinuclear") and a neurotransmitter ("Anandamide"). It is likely that these diverse meanings will tend to occur in very different contexts and in documents with different topics. On the other hand it is widely accepted that distinctions between possible meanings of words in natural language are often vague (Kilgarriff, 1993). It is likely that clearer distinctions between possible expansions of abbreviations make the task of identifying the correct one more straightforward than identifying meanings of ambiguous words. In addition, the creation of annotated data for WSD is often hampered by the difficulty in obtaining sufficient agreement between annotators (Artstein and Poesio, 2008; Weeber et al., 2001) and this problem does not apply to our automatically-generated corpus.

Results in Table 2 indicate that CUIs are useful features in the disambiguation of abbreviations. This is in contrast with previous experiments on ambiguous terms in biomedical documents (Stevenson et al., 2008) in which it was found that the best performance as obtained using only linguistic and MeSH features. It is likely that the clear distinction between expansions of abbreviations is the reason behind this difference. CUIs are assigned automatically by the MetaMap program (Aronson, 2001). However, this assignment is very noisy. It is likely that the various expansions of abbreviations are distinct enough for this noise to be tolerated by the learning algorithms while it causes problems when the meanings are closer together, such as in the case of ambiguous terms.

5.1 Performance of Individual Abbreviations

Table 3 shows the performance of the best WSD system (VSM learning algorithm with all features) for each abbreviation in the three subsets of our corpus. Our system performs well for all abbreviations. Accuracy is no lower than 92% for any abbreviation using Corpus.100 and no lower than 97% for Corpus.300, demonstrating that the approach is robust. In fact, the approach still performs well for abbreviations with low baseline scores, such as "APC", "BPD" and "LAM".

It is interesting to note that the abbreviations with the lowest performance tend to have expansions that are closely related. For example, the two expansions of "EMG" are 'electromyography' and 'electromyo-

	Corpus		
	100	200	300
ANA	0.980	-	-
APC	0.980	1.000	1.000
BPD	1.000	1.000	1.000
BSA	0.970	0.970	0.982
CAT	0.990	0.990	1.000
CML	0.960	0.963	0.978
CMV	0.970	0.970	0.970
DIP	1.000	1.000	-
EMG	0.920	0.960	0.980
FDP	0.970	-	-
LAM	0.960	0.980	0.980
MAC	0.970	0.990	0.989
MCP	0.980	0.978	1.000
PCA	0.960	0.987	0.992
PCP	0.990	1.000	1.000
PEG	0.980	0.982	1.000
PVC	0.990	1.000	-
RSV	0.960	0.972	0.978
Overall	0.974	0.984	0.990

Table 3: Performance of WSD system over individual abbreviations in three reduced corpora

gram' while for "LAM" one expansion ('Lymphangioleiomyomatosis') is a rare lung disease and the other ('Lipoarabinomannan') a molecule associated with another lung disease (tuberculosis). On the other hand, abbreviations that are more accurately disambiguated tend to have expansions with more distinct meanings. For example, "BPD" can be an acronym for 'borderline personality disorder' (a psychiatric diagnosis), 'bronchopulmonary dysplasia' (a lung disease) or 'biparietal diameter' (diameter of a foetus' head in an ultrasound) and the expansions of "DIP" are 'desquamative interstitial pneumonia' (a lung disease) and 'distal interphalangeal joints' (types of joints in the human hand and foot).

6 Conclusions

This paper has presented an approach to the disambiguation of ambiguous abbreviations in biomedical documents. We treat this problem as a form of WSD and apply a system that combines a wider range of features than have been previously applied, including those which are commonly used within WSD systems in addition to information from two domain-specific knowledge sources. The approach is evaluated using a corpus of abbreviations automatically mined from Medline and found to identify the correct expansion with accuracy of up to 99%. This figure is higher than previously reported results for abbreviation disambiguation systems, although direct comparison is difficult due to the use of different data sets. It was also found that best performance could be obtained using a simple machine learning algorithm and a diverse range of knowledge sources. Performance of our system is higher than is normally achieved by WSD systems when applied to general terms and we suggest that the reason for this is that the various expansions of abbreviations are better defined and more distinct than the senses of ambiguous words.

This study has been limited to the disambiguation of abbreviations consisting of exactly three letters. Possibilities for future work include experimenting with abbreviations of various lengths.

Data

The corpus described in Section 4 has been made freely available for research and may be obtained from `http://nlp.shef.ac.uk/BioWSD/downloads/abbreviationdata/`.

Acknowledgments

We are grateful to the anonymous reviewers of this paper for their valuable feedback.

References

E. Adar. 2004. SaRAD: A simple and robust abbreviation dictionary. *Bioinformatics*, 20(4):527–533.

E. Agirre and D. Martínez. 2004. The Basque Country University system: English and Basque tasks. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 44–48, Barcelona, Spain, July.

A. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the American Medical Informatics Association (AMIA)*, pages 17–21.

R. Artstein and M. Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

J. Chang, H. Schütze, and R. Altman. 2002. Creating an Online Dictionary of Abbreviations from MEDLINE. *The Journal of the American Medical Informatics Association*, 9(6):612–620.

H. Fred and T. Cheng. 1999. Acronymesis: the exploding misuse of acronyms. *Texas Heart Institute Journal*, 30:255–257.

S. Gaudan, H. Kirsch, and D. Rebholz-Schuhmann. 2005. Resolving abbreviations to their senses in Medline. *Bioinformatics*, 21(18):3658–3664.

M. Joshi, T. Pedersen, and R. Maclin. 2005. A Comparative Study of Support Vector Machines Applied to the Word Sense Disambiguation Problem for the Medical Domain. In *Proceedings of the Second Indian Conference on Artificial Intelligence (IICAI-05)*, pages 3449–3468, Pune, India.

M. Joshi, S. Pakhomov, T. Pedersen, and C. Chute. 2006. A comparative study of supervised learning as applied to acronym expansion in clinical reports. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 399–403, Washington, DC.

A. Kilgarriff. 1993. Dictionary word sense distinctions: An enquiry into their nature. *Computers and the Humanities*, 26:356–387.

H. Liu, Y. Lussier, and C. Friedman. 2001. Disambiguating ambiguous biomedical terms in biomedical narrative text: An unsupervised method. *Journal of Biomedical Informatics*, 34:249–261.

H. Liu, S. Johnson, and C. Friedman. 2002. Automatic Resolution of Ambiguous Terms Based on Machine Learning and Conceptual Relations in the UMLS. *Journal of the American Medical Informatics Association*, 9(6):621–636.

H. Liu, V. Teller, and C. Friedman. 2004. A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation. *Journal of the American Medical Informatics Association*, 11(4):320–331.

B. McInnes, T. Pedersen, and J. Carlis. 2007. Using UMLS Concept Unique Identifiers (CUIs) for Word Sense Disambiguation in the Biomedical Domain. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 533–537, Chicago, IL.

R. Mihalcea, T. Chklovski, and A. Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain.

S. Nelson, T. Powell, and B. Humphreys. 2002. The Unified Medical Language System (UMLS) Project. In Allen Kent and Carolyn M. Hall, editors, *Encyclopedia of Library and Information Science*. Marcel Dekker, Inc.

H. Ng, B. Wang, and S. Chan. 2003. Exploiting Parallel Texts for Word Sense Disambiguation: an Empirical Study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, pages 455–462, Sapporo, Japan.

N. Okazaki, S. Ananiadou, and J. Tsujii. 2008. A discriminative alignment model for abbreviation recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 657–664, Manchester, UK.

S. Pakhomov. 2002. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Philadelphia, PA.

T. Pedersen. 2001. A Decision Tree of Bigrams is an Accurate Predictor of Word Sense. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01)*, pages 79–86, Pittsburgh, PA.

J. Pustejovsky, J. Castano, R. Saur, A. Rumshisky, J. Zhang, and W. Luo. 2002. Medstract: Creating Large-scale Information Servers for Biomedical Libraries. In *ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain*.

A. Schwartz and M. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of the Pacific Symposium on Biocomputing*, Kauai.

M. Stevenson, Y. Guo, R. Gaizauskas, and D. Martinez. 2008. Disambiguation of biomedical text using diverse sources of information. *BMC Bioinformatics*, 9(Suppl 11):S7.

M. Weeber, J. Mork, and A. Aronson. 2001. Developing a Test Collection for Biomedical Word Sense Disambiguation. In *Proceedings of AMAI Symposium*, pages 746–50, Washington, DC.

I. Witten and E. Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.

H. Xu, J. Fan, G. Hripcsak, E. Mendonça, Markatou M., and Friedman C. 2007. Gene symbol disambiguation using knowledge-based profiles. *Bioinformatics*, 23(8):1015–22.

H. Yu, W. Kim, V. Hatzivassiloglou, and J. Wilbur. 2006. A large scale, corpus-based approach for automatically disambigutaing biomedical abbreviations. *ACM Transactions on Information Systems*, 24(3):380–404.

W. Zhou, I. Vetle, and N. Smalheiser. 2006. ADAM: another database of abbreviations in MEDLINE. *Bioinformatics*, 22(22):2813–2818.