

Using Second-order Vectors in a Knowledge-based Method for Acronym Disambiguation

Bridget T. McInnes*

College of Pharmacy
University of Minnesota
Minneapolis, MN 55455

Ted Pedersen

Department of Computer Science
University of Minnesota
Duluth, MN 55812

Ying Liu

College of Pharmacy
University of Minnesota
Minneapolis, MN 55455

Serguei V. Pakhomov

College of Pharmacy
University of Minnesota
Minneapolis, MN 55455

Genevieve B. Melton

Institute for Health Informatics
University of Minnesota
Minneapolis, MN 55455

Abstract

In this paper, we introduce a knowledge-based method to disambiguate biomedical acronyms using second-order co-occurrence vectors. We create these vectors using information about a long-form obtained from the Unified Medical Language System and Medline. We evaluate this method on a dataset of 18 acronyms found in biomedical text. Our method achieves an overall accuracy of 89%. The results show that using second-order features provide a distinct representation of the long-form and potentially enhances automated disambiguation.

1 Introduction

Word Sense Disambiguation (WSD) is the task of automatically identifying the appropriate sense of a word with multiple senses. For example, the word *culture* could refer to *anthropological culture* (e.g., the culture of the Mayan civilization), or a *laboratory culture* (e.g., cell culture).

Acronym disambiguation is the task of automatically identifying the contextually appropriate long-form of an ambiguous acronym. For example, the acronym *MS* could refer to the disease *Multiple Sclerosis*, the drug *Morphine Sulfate*, or the state *Mississippi*, among others. Acronym disambiguation can be viewed as a special case of WSD, although, unlike terms, acronyms tend to be complete phrases or expressions, therefore collocation features are not as easily identified. For example, the feature *rate* when disambiguating the term *interest*, as in

interest rate, may not be available. Acronyms also tend to be noun phrases, therefore syntactic features do not provide relevant information for the purposes of disambiguation.

Identifying the correct long-form of an acronym is important not only for the retrieval of information but the understanding of the information by the recipient. In general English, Park and Byrd (2001) note that acronym disambiguation is not widely studied because acronyms are not as prevalent in literature and newspaper articles as they are in specific domains such as government, law, and biomedicine.

In the biomedical sublanguage domain, acronym disambiguation is an extensively studied problem. Pakhomov (2002) note acronyms in biomedical literature tend to be used much more frequently than in news media or general English literature, and tend to be highly ambiguous. For example, the Unified Medical Language System (UMLS), which includes one of the largest terminology resources in the biomedical domain, contains 11 possible long-forms of the acronym *MS* in addition to the four examples used above. Liu et al. (2001) show that 33% of acronyms are ambiguous in the UMLS. In a subsequent study, Liu et al. (2002a) found that 80% of all acronyms found in Medline, a large repository of abstracts from biomedical journals, are ambiguous. Wren and Garner (2002) found that there exist 174,000 unique acronyms in the Medline abstracts in which 36% of them are ambiguous. The authors also estimated that the number of unique acronyms is increasing at a rate of 11,000 per year.

Supervised and semi-supervised methods have been used successfully for acronym disambiguation

*Contact author : bthomson@umn.edu.

but are limited in scope due to the need for sufficient training data. Liu et al. (2004) state that an acronym could have approximately 16 possible long-forms in Medline but could not obtain a sufficient number of instances for each of the acronym-long-form pairs for their experiments. Stevenson et al. (2009) cite a similar problem indicating that acronym disambiguation methods that do not require training data, regardless if it is created manually or automatically, are needed.

In this paper, we introduce a novel knowledge-based method to disambiguate acronyms using second-order co-occurrence vectors. This method does not rely on training data, and therefore, is not limited to disambiguating only commonly occurring possible long-forms. These vectors are created using the first-order features obtained from the UMLS about the acronym's long-forms and second-order features obtained from Medline. We show that using second-order features provide a distinct representation of the long-form for the purposes of disambiguation and obtains a significantly higher disambiguation accuracy than using first order features.

2 Unified Medical Language System

The Unified Medical Language System (UMLS) is a data warehouse that stores a number of distinct biomedical and clinical resources. One such resource, used in this work, is the Metathesaurus. The Metathesaurus contains biomedical and clinical concepts from over 100 disparate terminology sources that have been semi-automatically integrated into a single resource containing a wide range of biomedical and clinical information. For example, it contains the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT), which is a comprehensive clinical terminology created for the electronic exchange of clinical health information, the Foundational Model of Anatomy (FMA), which is an ontology of anatomical concepts created specifically for biomedical and clinical research, and MEDLINEPLUS, which is a terminology source containing health related concepts created specifically for consumers of health services.

The concepts in these sources can overlap. For example, the concept *Autonomic nerve* exists in both SNOMED CT and FMA. The Metathesaurus assigns

the synonymous concepts from the various sources a Concept Unique Identifiers (CUIs). Thus both the *Autonomic nerve* concepts in SNOMED CT and FMA are assigned the same CUI (C0206250). This allows multiple sources in the Metathesaurus to be treated as a single resource.

Some sources in the Metathesaurus contain additional information about the concept such as a concept's synonyms, its definition and its related concepts. There are two main types of relations in the Metathesaurus that we use: the parent/child and broader/narrower relations. A parent/child relation is a hierarchical relation between two concepts that has been explicitly defined in one of the sources. For example, the concept *Splanchnic nerve* has an *is-a* relation with the concept *Autonomic nerve* in FMA. This relation is carried forward to the CUI level creating a parent/child relations between the CUIs C0037991 (Splanchnic nerve) and C0206250 (Autonomic nerve) in the Metathesaurus. A broader/narrower relation is a hierarchical relation that does not explicitly come from a source but is created by the UMLS editors. We use the entire UMLS including the RB/RN and PAR/CHD relations in this work.

3 Medline

Medline (*Medical Literature Analysis and Retrieval System Online*) is a bibliographic database containing over 18.5 million citations to journal articles in the biomedical domain which is maintained by the National Library of Medicine (NLM). The 2010 Medline Baseline, used in this study, encompasses approximately 5,200 journals starting from 1948 and is 73 Gigabytes; containing 2,612,767 unique unigrams and 55,286,187 unique bigrams. The majority of the publications are scholarly journals but a small number of newspapers, and magazines are included.

4 Acronym Disambiguation

Existing acronym disambiguation methods can be classified into two categories: form-based and context-based methods. Form-based methods, such as the methods proposed by Taghva and Gilbreth (1999), Pustejovsky et al. (2001), Schwartz and Hearst (2003) and Nadeau and Turney (2005), disambiguate the acronym by comparing its letters di-

rectly to the initial letters in the possible long-forms and, therefore, would have difficulties in distinguishing between acronyms with similar long-forms (e.g., RA referring to Refractory anemia or Rheumatoid arthritis).

In contrast, context-based methods disambiguate between acronyms based on the context in which the acronym is used with the assumption that the context surrounding the acronym would be different for each of the possible long-forms. In the remainder of this section, we discuss these types of methods in more detail.

4.1 Context-based Acronym Disambiguation Methods

Liu et al. (2001) and Liu et al. (2002b) introduce a semi-supervised method in which training and test data are automatically created by extracting abstracts from Medline that contain the acronym's long-forms. The authors use collocations and a bag-of-words approach to train a Naive Bayes algorithm and report an accuracy of 97%. This method begins to treat acronym disambiguation as more of a WSD problem by looking at the context in which the acronym exists to determine its long-form, rather than the long-form itself. In a subsequent study, Liu et al. (2004) explore using additional features and machine learning algorithms and report an accuracy of 99% using the Naive Bayes.

Joshi (2006) expands on Liu, et al's work. They evaluate additional machine learning algorithms using unigrams, bigrams and trigrams as features. They found that given their feature set, SVMs obtain the highest accuracy (97%).

Stevenson et al. (2009) re-recreate this dataset using the method described in Liu et al. (2001) to automatically create training data for their method which uses a mixture of linguistics features (e.g., collocations, unigrams, bigrams and trigrams) in combination with the biomedical features CUIs and Medical Subject Headings, which are terms manually assigned to Medline abstracts for indexing purposes. The authors evaluate the Naive Bayes, SVM and Vector Space Model (VSM) described by Agirre and Martinez (2004), and report that VSM obtained the highest accuracy (99%).

Pakhomov (2002) also developed a semi-supervised method in which training data was

automatically created by first identifying the long-form found in the text of clinical reports, replacing the long-form with the acronym to use as training data. A maximum entropy model trained and tested on a corpus of 10,000 clinical notes achieved an accuracy of 89%. In a subsequent study, Pakhomov et al. (2005) evaluate obtaining training data from three sources: Medline, clinical records and the world wide web finding using a combination of instances from clinical records and the web obtained the highest accuracy.

Joshi et al. (2006) compare using the Naive Bayes, Decision trees and SVM on ambiguous acronyms found in clinical reports. The authors use the part-of-speech, the unigrams and the bigrams of the context surrounding the acronym as features. They evaluate their method on 7,738 manually disambiguated instances of 15 ambiguous acronyms obtaining an accuracy of over 90% for each acronym.

5 Word Sense Disambiguation

Many knowledge-based WSD methods have been developed to disambiguate terms which are closely related to the work presented in this paper. Lesk (1986) proposes a definition overlap method in which the appropriate sense of an ambiguous term was determined based on the overlap between its definition in a machine readable dictionary (MRD). Ide and Véronis (1998) note that this work provided a basis for most future MRD disambiguation methods; including the one presented in this paper.

Banerjee and Pedersen (2002) use the Lesk's overlap method to determine the relatedness between two concepts (synsets) in WordNet. They extend the method to not only include the definition (gloss) of the two synsets in the overlap but also the glosses of related synsets.

Wilks et al. (1990) expand upon Lesk's method by calculating the number of times the words in the definition co-occur with the ambiguous words. In their method, a vector is created using the co-occurrence information for the ambiguous word and each of its possible senses. The similarity is then calculated between the ambiguous word's vector and each of the sense vectors. The sense whose vector is most similar is assigned to the ambiguous word.

		FEATURES											
		metabolites	glucose	fructose	phosphoric acid	esters	changed	effect	glycolyte	enzymes	combined	decreases	intensity
Extended Definition for Fructose Diphosphate	disphosphoric	0	0	0	0	0	0	0	0	0	0	0	0
	acid	0	0	0	.3	0	.2	0	0	0	0	.1	0
	esters	0	0	0	0	.5	0	0	0	0	0	0	0
	fructose	0	.1	0	0	0	0	0	0	0	0	0	0
	diphosphate	0	0	0	0	0	0	0	0	0	0	0	0
	isomer	0	0	0	0	0	0	0	0	0	0	0	0
	prevalent	0	0	0	0	0	0	0	0	0	0	0	0
2nd order vector for Fructose Diphosphate		0	.1	0	.3	.5	.2	0	0	0	0	.1	0

Figure 1: 2nd Order Vector for Fructose Diphosphate (FDP)

Patwardhan and Pedersen (2006) introduce a vector measure to determine the relatedness between pairs of concepts. In this measure, a second order co-occurrence vector is created for each concept using the words in each of the concepts definition and calculating the cosine between the two vectors. This method has been used in the task of WSD by calculating the relatedness between each possible sense of the ambiguous word and its surrounding context. The context whose sum is the most similar is assigned to the ambiguous word.

Second-order co-occurrence vectors were first introduced by Schütze (1992) for the task of word sense *discrimination* and later extended by Purandare and Pedersen (2004). As noted by Pedersen (2010), disambiguation requires a sense-inventory in which the long-forms are known ahead of time, where as in discrimination this information is not known a priori.

6 Method

In our method, a second-order co-occurrence vector is created for each possible long-form of the

acronym, and the acronym itself. The appropriate long-form of the acronym is then determined by computing a cosine between the vector representing the ambiguous acronym and each of the vectors representing the long-forms. The long-form whose vector has the smallest angle between it and the acronym vector is chosen as the most likely long-form of the acronym.

To create a second-order vector for a long-form, we first obtain a textual description of the long-form in the UMLS, which we refer to as the *extended definition*. Each long-form, from our evaluation set, was mapped to a concept in the UMLS, therefore, we use the long-form's definition plus the definition of its parent/children and narrow/broader relations and the terms in the long-form.

We include the definition of the related concepts because not all concepts in the UMLS have a definition. In our evaluation dataset, not a single acronym has a definition for each possible long-form. On average, each extended definition contains approximately 453 words. A short example of the extended definition for the acronym FDP when referring to

fructose diphosphate is: “Diphosphoric acid esters of fructose. The fructose diphosphate isomer is most prevalent. fructose diphosphate.”

After the extended definition is obtained, we create the second-order vector by first creating a word by word co-occurrence matrix in which the rows represent the content words in the long-forms, extended definition, and the columns represent words that co-occur in Medline abstracts with the words in the definition. Each cell in this matrix contains the Log Likelihood Ratio (Dunning (1993)) of the word found in the row and the word in the column. Second, each word in the long-forms, extended definition is replaced by its corresponding vector, as given in the co-occurrence matrix. The centroid of these vectors constitutes the second order co-occurrence vector used to represent the long-form.

For example, given the *example corpus* containing two instances: 1) The metabolites, glucose fructose and their phosphoric acid esters are changed due to the effect of glycolytic enzymes, and 2) The phosphoric acid combined with metabolites decreases the intensity. Figure 1 shows how the second-order co-occurrence vector is created for the long-form *fructose diphosphate* using the extended definition and features from our given corpus above.

The second-order co-occurrence vector for the ambiguous acronym is created in a similar fashion, only rather than using words in the extended definition, we use the words surrounding the acronym in the instance.

Vector methods are subject to noise introduced by features that do not distinguish between the different long-forms of the acronym. To reduce this type of noise, we select the features to use in the second order co-occurrence vectors based on the following criteria: 1) second order feature cannot be a stop-word, and 2) second order feature must occur at least twice in the feature extraction dataset and not occur more than 150 times. We also experiment with the location of the second-order feature with respect to the first-order feature by varying the window size of zero, four, six and ten words to the right and the left of the first-order feature. The experiments in this paper were conducted using CuiTools v0.15.¹

Our method is different from other context-based

acronym disambiguation methods discussed in the related work because it does not require annotated training data for each acronym that needs to be disambiguated. Our method differs from the method proposed by Wilks et al. (1990) in two fundamental aspects: 1) using the *extended definition* of the possible long-forms of an acronym, and 2) using second-order vectors to represent the instance containing the acronym and each of the acronym’s possible long-forms.

7 Data

7.1 Acronym Dataset

We evaluated our method on the “Abbrev” dataset² made available by Stevenson et al. (2009). The acronyms and long-forms in the data were initially presented by Liu et al. (2001). Stevenson et al. (2009) automatically re-created this dataset by identifying the acronyms and long-forms in Medline abstracts and replacing the long-form in the abstract with its acronym. Each abstract contains approximately 216 words. The dataset consists of three subsets containing 100 instances, 200 instances and 300 instances of the ambiguous acronym referred to as Abbrev.100, Abbrev.200, Abbrev.300, respectively. The acronyms long-forms were manually mapped to concepts in the UMLS by Stevenson, et al.

A sufficient number of instances were not found for each of the 21 ambiguous acronyms by Stevenson et al. (2009). For example, “ASP” only contained 71 instances and therefore not included in any of the subsets. “ANA” and “FDP” only contained just over 100 instances and therefore, are only included in the Abbrev.100 subset. “ACE”, “ASP” and “CSF” were also excluded because several of the acronyms’ long-forms did not occur frequently enough in Medline to create a balanced dataset.

We evaluate our method on the same subsets that Stevenson et al. (2009) used to evaluate their supervised method. The average number of long-forms per acronym is 2.6 and the average majority sense across all subsets is 70%.

7.2 Feature Extraction Dataset

We use abstracts from Medline, containing ambiguous acronym or long-form, to create the second-

¹<http://cuitools.sourceforge.net>

²<http://nlp.shef.ac.uk/BioWSD/downloads/corpora>

order co-occurrence vectors for our method as described in Section 6. Table 1 shows the number of Medline abstracts extracted for the acronyms.

Acronyms	# Abstracts	Acronym	# Abstracts
ANA	3,267	APC	11,192
BPD	3,260	BSA	10,500
CAT	44,703	CML	8,777
CMV	13,733	DIP	2,912
EMG	16,779	FDP	1,677
LAM	1,572	MAC	6,528
MCP	2,826	PCA	11,044
PCP	5,996	PEG	10,416
PVC	2,780	RSV	5,091

Table 1: Feature Extraction Data for Acronyms

8 Results

Table 2 compares the majority sense baseline and the first-order baseline with the results obtained using our method on the Acronym Datasets (Abbrev.100, Abbrev.200 and Abbrev.300) using a window size of zero, four, six and ten. Differences between the means of disambiguation accuracy produced by various approaches were tested for statistical significance using the pair-wise Student’s t-tests with the significance threshold set to 0.01.

	Window Size	Abbrev		
		100	200	300
Maj. Sense Baseline		0.70	0.70	0.70
1-order Baseline		0.57	0.61	0.61
Our Method	0	0.83	0.83	0.81
	4	0.86	0.87	0.86
	6	0.88	0.90	0.89
	10	0.88	0.90	0.89

Table 2: Overall Disambiguation Results

The majority sense baseline is often used to evaluate supervised learning algorithms and indicates the accuracy that would be achieved by assigning the most frequent sense (long-form) to every instance. The results in Table 2 demonstrate that our method is significantly more accurate than the majority sense baseline ($p \leq 0.01$).

We compare the results using second-order vectors to first-order vectors. Table 2 shows that accuracy of the second-order results is significantly higher than the first-order results ($p \leq 0.01$).

The results in Table 2 also show that, as the window size grows from zero to six, the accuracy of the

system increases and plateaus at a window size of ten. There is no statistically significant difference between using a window size of six and ten but there is a significant difference between a window size of zero and six, as well as four and six ($p \leq 0.01$).

Acronym	# Long forms	Abbrev 100	Abbrev 200	Abbrev 300
ANA	3	0.84		
APC	3	0.88	0.87	0.87
BPD	3	0.96	0.95	0.95
BSA	2	0.95	0.93	0.92
CAT	2	0.88	0.87	0.87
CML	2	0.81	0.84	0.83
CMV	2	0.98	0.98	0.98
DIP	2	0.98	0.98	
EMG	2	0.88	0.89	0.88
FDP	4	0.65		
LAM	2	0.86	0.87	0.88
MAC	4	0.94	0.95	0.95
MCP	4	0.73	0.67	0.68
PCA	4	0.78	0.79	0.79
PCP	2	0.97	0.96	0.96
PEG	2	0.89	0.89	0.88
PVC	2	0.95	0.95	
RSV	2	0.97	0.98	0.98

Table 3: Individual Results using a Window Size of 6.

9 Error Analysis

Table 3 shows the results obtained by our method for the individual acronyms using a window size of six, and the number of possible long-forms per acronym. Of the 18 acronyms, three obtain an accuracy below 80 percent: FDP, MCP and PCA.

FDP has four possible long-forms: Fructose Diphosphate (E1), Formycin Diphosphate (E2), Fibrinogen Degradation Product (E3) and Flexor Digitorum Profundus (E4). The confusion matrix in Table 4 shows that the method was unable to distinguish between the two long-forms, E1 and E2, which are both diphosphates, nor E2 and E3.

Long-Form	E1	E2	E3	E4
E1: Fructose Diphosphate				
E2: Formycin Diphosphate	5	2	11	19
E3: Fibrinogen Degradation Product			4	
E4: Flexor Digitorum Profundus				59

Table 4: FDP Confusion Matrix

MCP also has four possible long-forms: Multicatalytic Protease (E1), Metoclopramide (E2), Monocyte Chemoattractant Protein (E3) and Membrane

Cofactor Protein (E4). The confusion matrix in Table 5 shows that the method was not able to distinguish between E3 and E4, which are both proteins, and E1, which is a protease (an enzyme that breaks down a protein).

Long-Form	E1	E2	E3	E4
E1: Multicatalytic Protease	1	5	6	1
E2: Metoclopramide		15		
E3: Monocyte Chemoattractant Protein	1	3	44	11
E4: Membrane Cofactor Protein				13

Table 5: MCP Confusion Matrix

PCA has four possible long-forms: Passive Cutaneous Anaphylaxis (E1), Patient Controlled Analgesia (E2), Principal Component Analysis (E3), and Posterior Cerebral Artery (E4). The confusion matrix in Table 6 shows that the method was not able to distinguish between E2 and E3. Analyzing the extended definitions of the concepts showed that E2 includes the definition to the concept Pain Management. The words in this definition overlap with many of the words used in E3s extended definition.

Long-Form	E1	E2	E3	E4
E1:Passive Cutaneous Anaphylaxis	18		6	1
E2:Patient Controlled Analgesia		5	15	
E3:Principal Component Analysis			48	
E4:Posterior Cerebral Artery				7

Table 6: PCA Confusion Matrix

10 Comparison with Previous Work

Of the previously developed methods, Liu et al. (2004) and Stevenson et al. (2009) evaluated their semi-supervised methods on the same dataset as we used for the current study. A direct comparison can not be made between our method and Liu et al. (2004) because we do not have an exact duplication of the dataset that they use. Their results are comparable to Stevenson et al. (2009) with both reporting results in the high 90s. Our results are directly comparable to Stevenson et al. (2009) who report an overall accuracy of 98%, 98% and 99% on the Abbrev.100, Abbrev.200 and Abbrev.300 datasets respectively. This is approximately 10 percentage points higher than our results.

The advantage of the methods proposed by Stevenson et al. (2009) and Liu et al. (2004) is that

they are semi-supervised which have been shown to obtain higher accuracies than methods that do not use statistical machine learning algorithms. The disadvantage is that sufficient training data are required for each possible acronym-long-form pair. Liu et al. (2004) state that an acronym could have approximately 16 possible long-forms in Medline but a sufficient number of instances for each of the acronym-long-form pairs were not found in Medline and, therefore, evaluated their method on 15 out of the original 34 acronyms. Stevenson et al. (2009) cite a similar problem in re-creating this dataset. This shows the limitation to these methods is that a sufficient number of training examples can not be obtained for each acronym that needs to be disambiguated. The method proposed in the paper does not have this limitation and can be used to disambiguate any acronym in Medline.

11 Discussion

In this paper, we presented a novel method to disambiguate acronyms in biomedical text using second-order features extracted from the UMLS and Medline. The results show that using second-order features provide a distinct representation of the long-form that is useful for disambiguation.

We believe that this is because biomedical text contains technical terminology that has a rich source of co-occurrence information associated with them due to their compositionality. Using second-order information works reasonably well because when the terms in the extended definition are broken up into their individual words, information is not being lost. For example, the term Patient Controlled Analgesia can be understood by taking the union of the meanings of the three terms and coming up with an appropriate definition of the term (patient has control over their analgesia).

We evaluated various window sizes to extract the second-order co-occurrence information from, and found using locally occurring words obtains a higher accuracy. This is consistent with the finding reported by Choueka and Lusignan (1985) who conducted an experiment to determine what size window is needed for humans to determine the appropriate sense of an ambiguous word.

The amount of data used to extract the second-

order features for each ambiguous acronym varied depending on its occurrence in Medline. Table 1 in Section 7.2 shows the number of abstracts in Medline used for each acronym. We compared the accuracy obtained by our method using a window size of six on the Abbrev.100 dataset with the number of abstracts in the feature extraction data. We found that the accuracy was not correlated with the amount of data used ($r = 0.07$). This confirms that it is not the quantity but the content of the contextual information that determines the accuracy of disambiguation.

We compared using second-order features and first-order features showing that the second-order results obtained a significantly higher accuracy. We believe that this is because the definitions of the possible concepts are too sparse to provide enough information to distinguish between them. This finding coincides to that of Purandare and Pedersen (2004) and Pedersen (2010) who found that with large amounts of data, first-order vectors perform better than second-order vectors, but second-order vectors are a good option when large amounts of data are not available.

The results of the error analysis indicate that for some acronyms using the extended definition does not provide sufficient information to make finer grained distinctions between the long-forms. This result also indicates that, although many long-forms of acronyms can be considered coarse-grained senses, this is not always the case. For example, the analysis of *MCP* showed that two of its possible long-forms are proteins which are difficult to differentiate from given the context.

The results of the error analysis also show that indicative collocation features for acronyms are not easily identified because acronyms tend to be complete phrases. For example, two of the possible long-forms of *DF* are *Fructose Diphosphate* and *Formycin Diphosphate*.

Two main limitations of this work must be mentioned to facilitate the interpretation of the results. The first is the small number of acronyms and the small number of long-forms per acronym in the dataset; however, the acronyms in this dataset are representative of the kinds of acronyms one would expect to see in biomedical text. The second limitation is that the dataset contains only those acronyms whose long-forms were found in Medline abstracts.

The main goal of this paper was to determine if the context found in the long-forms, extended definition was distinct enough to distinguish between them using second-order vectors. For this purpose, we feel that the dataset was sufficient although a more extensive dataset may be needed in the future for improved coverage.

12 Future Work

In the future, we plan to explore three different avenues. The first avenue is to look at obtaining contextual descriptions of the possible long-forms from resources other than the UMLS such as the MetaMapped Medline baseline and WordNet. The second avenue is limiting the features that are used in the instance vectors. The first-order features in the instance vector contain the words from the entire abstract. As previously mentioned, vector methods are subject to noise, therefore, in the future we plan to explore using only those words that are co-located next to the ambiguous acronym. The third avenue is expanding the vector to allow for terms. Currently, we use word vectors, in the future, we plan to extend the method to use terms, as identified by the UMLS, as features rather than single words.

We also plan to test our approach in the clinical domain. We believe that acronym disambiguation may be more difficult in this domain due to the increase amount of long-forms as seen in the datasets used by Joshi et al. (2006) and Pakhomov (2002).

13 Conclusions

Our study constitutes a significant step forward in the area of automatic acronym ambiguity resolution, as it will enable the incorporation of scalable acronym disambiguation into NLP systems used for indexing and retrieval of documents in specialized domains such as medicine. The advantage of our method over previous methods is that it does not require manually annotated training for each acronym to be disambiguated while still obtaining an overall accuracy of 89%.

Acknowledgments

This work was supported by the National Institute of Health, National Library of Medicine Grant #R01LM009623-01.

References

- E. Agirre and D. Martinez. 2004. The Basque Country University system: English and Basque tasks. In *Proceedings of the 3rd ACL workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL)*, pages 44–48.
- S. Banerjee and T. Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using WordNet. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, pages 136–145.
- Y. Choueka and S. Lusignan. 1985. Disambiguation by short contexts. *Computers and the Humanities*, 19(3):147–157.
- T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.
- N. Ide and J. Véronis. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1):2–40.
- M. Joshi, S. Pakhomov, T. Pedersen, and C.G. Chute. 2006. A comparative study of supervised learning as applied to acronym expansion in clinical reports. In *Proceedings of the Annual Symposium of AMIA*, pages 399–403.
- M. Joshi. 2006. Kernel Methods for Word Sense Disambiguation and Abbreviation Expansion. Master's thesis, University of Minnesota.
- M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of the 5th Annual International Conference on Systems Documentation*, pages 24–26.
- H. Liu, YA. Lussier, and C. Friedman. 2001. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method. *Journal of Biomedical Informatics*, 34(4):249–261.
- H. Liu, A.R. Aronson, and C. Friedman. 2002a. A study of abbreviations in MEDLINE abstracts. In *Proceedings of the Annual Symposium of AMIA*, pages 464–468.
- H. Liu, S.B. Johnson, and C. Friedman. 2002b. Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS. *JAMIA*, 9(6):621–636.
- H. Liu, V. Teller, and C. Friedman. 2004. A multi-aspect comparison study of supervised word sense disambiguation. *JAMIA*, 11(4):320–331.
- D. Nadeau and P. Turney. 2005. A supervised learning approach to acronym identification. In *Proceedings of the 18th Canadian Conference on Artificial Intelligence*, pages 319–329.
- S. Pakhomov, T. Pedersen, and C.G. Chute. 2005. Abbreviation and acronym disambiguation in clinical discourse. In *Proceedings of the Annual Symposium of AMIA*, pages 589–593.
- S. Pakhomov. 2002. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 160–167.
- Y. Park and R.J. Byrd. 2001. Hybrid text mining for finding abbreviations and their definitions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 126–133.
- S. Patwardhan and T. Pedersen. 2006. Using WordNet-based context vectors to estimate the semantic relatedness of concepts. In *Proceedings of the EACL 2006 Workshop Making Sense of Sense - Bringing Computational Linguistics and Psycholinguistics Together*, pages 1–8.
- T. Pedersen. 2010. The effect of different context representations on word sense discrimination in biomedical texts. In *Proceedings of the 1st ACM International IHI Symposium*, pages 56–65.
- A. Purandare and T. Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 41–48.
- J. Pustejovsky, J. Castano, B. Cochran, M. Kotecki, M. Morrell, and A. Rumshisky. 2001. Extraction and disambiguation of acronym-meaning pairs in medline. *Unpublished manuscript*.
- H. Schütze. 1992. Dimensions of meaning. In *Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, pages 787–796.
- A.S. Schwartz and M.A. Hearst. 2003. A simple algorithm for identifying abbreviation definitions in biomedical text. In *Proceedings of the Pacific Symposium on Biocomputing (PSB)*, pages 451–462.
- M. Stevenson, Y. Guo, A. Al Amri, and R. Gaizauskas. 2009. Disambiguation of biomedical abbreviations. In *Proceedings of the ACL BioNLP Workshop*, pages 71–79.
- K. Taghva and J. Gilbreth. 1999. Recognizing acronyms and their definitions. *ISRI UNLV*, 1:191–198.
- Y. Wilks, D. Fass, C.M. Guo, J.E. McDonald, T. Plate, and B.M. Slator. 1990. Providing machine tractable dictionary tools. *Machine Translation*, 5(2):99–154.
- J.D. Wren and H.R. Garner. 2002. Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of Information in Medicine*, 41(5):426–434.