An Unsupervised Vector Approach to Biomedical Term Disambiguation: Integrating UMLS and Medline

Bridget T. McInnes
Computer Science Department
University of Minnesota Twin Cities
Minneapolis, MN 55155, USA
bthomson@cs.umn.edu

Abstract

This paper introduces an unsupervised vector approach to disambiguate words in biomedical text that can be applied to all-word disambiguation. We explore using contextual information from the Unified Medical Language System (UMLS) to describe the possible senses of a word. We experiment with automatically creating individualized stoplists to help reduce the noise in our dataset. We compare our results to SenseClusters and Humphrey et al. (2006) using the NLM-WSD dataset and with SenseClusters using conflated data from the 2005 Medline Baseline.

1 Introduction

Some words have multiple senses. For example, the word *cold* could refer to a viral infection or the temperature. As humans, we find it easy to determine the appropriate sense (concept) given the context in which the word is used. For a computer, though, this is a difficult problem which negatively impacts the accuracy of biomedical applications such as medical coding and indexing. The goal of our research is to explore using information from biomedical knowledge sources such as the Unified Medical Language System (UMLS) and Medline to help distinguish between different possible concepts of a word.

In the UMLS, concepts associated with words and terms are enumerated via Concept Unique Identifiers (CUIs). For example, two possible senses of *cold* are “C0009264: Cold Temperature” and “C0009443: Common Cold” in the UMLS release 2008AA. The UMLS is also encoded with different semantic and syntactic structures. Some such information includes related concepts and semantic types. A semantic type (ST) is a broad subject categorization assigned to a CUI. For example, the ST of “C0009264: Cold Temperature” is “Idea or Concept” while the ST for “C0009443: Common Cold” is “Disease or Syndrome”. Currently, there exists approximately 1.5 million CUIs and 135 STs in the UMLS. Medline is an online database that contains 11 million references biomedical articles.

In this paper, we introduce an unsupervised vector approach to disambiguate words in biomedical text using contextual information from the UMLS and Medline. We compare our approach to Humphrey et al. (2006) and SenseClusters. The ability to make disambiguation decisions for words that have the same ST differentiates SenseClusters and our approach from Humphrey et al.’s (2006). For example, the word *weight* in the UMLS has two possible CUIs, “C0005912: Body Weight” and “C0699807: Weight”, each having the ST “Quantitative Concept”. Humphrey et al.’s (2006) approach relies on the concepts having different STs therefore is unable to disambiguate between these two concepts.

Currently, most word sense disambiguation approaches focus on lexical sample disambiguation which only attempts to disambiguate a predefined set of words. This type of disambiguation is not practical for large scale systems. All-words disambiguation approaches disambiguate all ambiguous words in a running text making them practical for large scale systems. Unlike SenseClusters, Humphrey, et al. (2006) and our approach can be

Proceedings of the ACL-08: HLT Student Research Workshop (Companion Volume), pages 49–54,
Columbus, June 2008. ©2008 Association for Computational Linguistics

used to perform all-words disambiguation.

In the following sections, we first discuss related work. We then discuss our approach, experiments and results. Lastly, we discuss our conclusions and future work.

2 Related Work

There has been previous work on word sense disambiguation in the biomedical domain. Leroy and Rindflesch (2005) introduce a supervised approach that uses the UMLS STs and their semantic relations of the words surrounding the target word as features into a Naive Bayes classifier. Joshi et al. (2005) introduce a supervised approach that uses unigrams and bigrams surrounding the target word as features into a Support Vector Machine. A unigram is a single content word that occurs in a window of context around the target word. A bigram is an ordered pair of content words that occur in a window of context around the target word. McInnes et al. (2007) introduce a supervised approach that uses CUIs of the words surrounding the target word as features into a Naive Bayes classifier.

Humphrey et al. (2006) introduce an unsupervised vector approach using Journal Descriptor (JD) Indexing (JDI) which is a ranking algorithm that assigns JDs to journal titles in MEDLINE. The authors apply the JDI algorithm to STs with the assumption that each possible concept has a distinct ST. In this approach, an ST vector is created for each ST by extracting associated words from the UMLS. A target word vector is created using the words surrounding the target word. The JDI algorithm is used to obtain a score for each word-JD and ST-JD pair using the target word and ST vectors. These pairs are used to create a word-ST table using the cosine coefficient between the scores. The cosine scores for the STs of each word surrounding the target word are averaged and the concept associated with the ST that has the highest average is assigned to the target word.

3 Vector Approaches

Patwardhan and Pedersen (2006) introduce a vector measure to determine the relatedness between pairs of concepts. In this measure, a co-occurrence matrix of all words in a given corpus is created containing how often they occur in the same window of context with each other. A gloss vector is then created for each concept containing the word vector for each word in the concepts definition (or gloss). The cosine between the two gloss vectors is computed to determine the concepts relatedness.

SenseClusters [1] is an unsupervised knowledge-lean word sense disambiguation package The package uses clustering algorithms to group similar instances of target words and label them with the appropriate sense. The clustering algorithms include Agglomerative, Graph partitional-based, Partitional biased agglomerative and Direct k-way clustering. The clustering can be done in either vector space where the vectors are clustered directly or similarity space where vectors are clustered by finding the pair-wise similarities among the contexts. The feature options available are first and second-order co-occurrence, unigram and bigram vectors. First-order vectors are highly frequent words, unigrams or bigrams that co-occur in the same window of context as the target word. Second-order vectors are highly frequent words that occur with the words in their respective first order vector.

We compare our approach to SenseClusters v0.95 using direct k-way clustering with the I2 clustering criterion function and cluster in vector space. We experiment with first-order unigrams and second-order bigrams with a Log Likelihood Ratio greater than 3.84 and the exact and gap cluster stopping parameters (Purandare and Pedersen, 2004; Kulkarni and Pedersen, 2005).

4 Our Approach

Our approach has three stages: i) we create a the feature vector for the target word ($instance\ vector$) and each of its possible concepts ($concept\ vectors$) using SenseClusters, ii) we calculate the cosine between the instance vector and each of the concept vectors, and iii) we assign the concept whose concept vector is the closest to the instance vector to the target word.

To create the the instance vector, we use the words that occur in the same abstract as the target word as features. To create the concept vector, we explore four different context descriptions of a possible concept to use as features. Since each possible concept

[1] http://senseclusters.sourceforge.net/

has a corresponding CUI in the UMLS, we explore using: i) the words in the concept's CUI definition, ii) the words in the definition of the concept's ST definition, iii) the words in both the CUI and ST definitions, and iv) the words in the CUI definition unless one does not exist then the words in its ST definition.

We explore using the same feature vector parameters as in the SenseCluster experiments: i) first-order unigrams, and ii) second-order bigram. We also explore using a more judicious approach to determine which words to include in the feature vectors. One of the problems with an unsupervised vector approach is its susceptibility to noise. A word frequently seen in a majority of instances may not be useful in distinguishing between different concepts. To alleviate this problem, we create an individualized stoplist for each target word using the inverse document frequency (IDF). We calculate the IDF score for each word surrounding the target word by taking the log of the number of documents in the training data divided by the number of documents the term has occurred in the dataset. We then extract those words that obtain an IDF score under the threshold of one and add them to our basic stoplist to be used when determining the appropriate sense for that specific target word.

5 Data

5.1 Training Data

We use the abstracts from the 2005 Medline Baseline as training data. The data contains 14,792,864 citations from the 2005 Medline repository. The baseline contains 2,043,918 unique tokens and 295,585 unique concepts.

5.2 NLM-WSD Test Dataset

We use the National Library of Medicine's Word Sense Disambiguation (NLM-WSD) dataset developed by (Weeber et al., 2001) as our test set. This dataset contains 100 instances of 50 ambiguous words from 1998 MEDLINE abstracts. Each instance of a target word was manually disambiguated by 11 human evaluators who assigned the word a CUI or "None" if none of the CUIs described the concept. (Humphrey et al., 2006) evaluate their approach using a subset of 13 out of the 50 words whose majority sense is less than 65% and whose possible concepts do not have the same ST. Instances tagged as "None" were removed from the dataset. We evaluate our approach using these same words and instances.

5.3 Conflate Test Dataset

To test our algorithm on a larger biomedical dataset, we are creating our own dataset by conflating two or more unambiguous words from the 2005 Medline Baseline. We determine which words to conflate based on the following criteria: i) the words have a single concept in the UMLS, ii) the words occur approximately the same number of times in the corpus, and iii) the words do not co-occur together.

We create our dataset using $name\text{-}conflate$ [2] to extract instances containing the conflate words from the 2005 Medline Baseline. Table 4 shows our current set of conflated words with their corresponding number of test (test) and training (train) instances. We refer to the conflated words as their pseudowords throughout the paper.

6 Experimental Results

In this section, we report the results of our experiments. First, we compare the results of using the IDF stoplist over a basic stoplist. Second, we compare the results of using the different context descriptions. Third, we compare our approach to SenseClusters and Humphrey et al. (2006) using the NLM-WSD dataset. Lastly, we compare our approach to SenseClusters using the conflated dataset.

In the following tables, CUI refers to the CUI definition of the possible concept as context, ST refers to using the ST definition of the possible concept as context, CUI+ST refers to using both definitions as context, and CUI→ST refers to using the CUI definition unless if one doesn't exist then using ST definition. Maj. refers to the "majority sense" baseline which is accuracy that would be achieved by assigning every instance of the target word with the most frequent sense as assigned by the human evaluators.

6.1 Stoplist Results

Table 2 shows the overall accuracy of our approach using the basic stoplist and the IDF stoplist on the

[2] http://www.d.umn.edu/ tpederse/namedata.html

target word	Unigram				Bigram			
	CUI	ST	CUI+ST	CUI→ST	CUI	ST	CUI+ST	CUI→ST
adjustment	44.57	31.61	46.74	44.57	**47.83**	38.04	27.17	**47.83**
blood pressure	39.39	34.34	41.41	38.38	43.43	27.27	**47.47**	38.38
degree	3.13	**70.31**	**70.31**	**70.31**	3.13	48.44	48.44	48.44
evaluation	50.51	50.51	53.54	51.52	50.51	**54.55**	52.53	51.52
growth	**63.64**	51.52	42.42	**63.64**	**63.64**	51.52	48.48	**63.64**
immunosuppression	50.51	46.46	50.51	50.51	43.43	**57.58**	48.48	43.43
mosaic	0	33.33	27.08	**37.50**	0	28.13	22.92	22.92
nutrition	28.41	34.09	35.23	25.00	38.64	**39.77**	36.36	37.50
radiation	57.73	44.78	58.76	57.73	**60.82**	28.36	**60.82**	**60.82**
repair	74.63	25.00	41.79	37.31	**76.12**	54.69	44.78	41.79
scale	32.81	48.00	42.19	51.56	0	18.00	95.31	**96.88**
sensitivity	6.00	**50.56**	48.00	48.00	8.00	44.94	18.00	18.00
white	48.31	38.61	46.07	**49.44**	44.94	38.16	43.82	**49.44**
average	38.43	43.01	46.46	**48.11**	36.96	40.73	45.74	47.74

Table 1: Accuracy of Our Approach using Different Context Descriptions

NLM-WSD dataset using each of the different context descriptions described above. The results show an approximately a 2% higher accuracy over using the basic stoplist. The exception is when using the CUI context description; the accuracy decreased by approximately 2% when using the unigram feature set and approximately 1% when using the bigram feature set.

context	Basic stoplist		IDF stoplist	
	unigram	bigram	unigram	bigram
CUI	**41.02**	**37.68**	38.43	36.96
ST	42.74	37.14	**43.01**	**40.73**
CUI+ST	44.13	42.71	**46.46**	**45.74**
CUI→ST	46.61	45.58	**48.11**	**47.74**

Table 2: Accuracy of IDF stoplist on the NLM-WSD dataset

6.1.1 Context Results

Table 1 shows the results of our approach using the CUI and ST definitions as context for the possible concepts on the NLM-WSD dataset and Table 4 shows similar results using the conflate dataset.

On the NLM-WSD dataset, the results show a large difference in accuracy between the contexts on a word by word basis making it difficult to determine which of the context description performs the best. The unigram results show that CUI→ST and CUI+ST obtain the highest accuracy for five words, and CUI and ST obtain the highest accuracy for one word. The bigram results show that CUI→ST and CUI obtains the highest accuracy for two words, ST obtains the highest accuracy for four words, and CUI+ST obtains the highest accuracy for one word. The overall results show that using unigrams with the context description CUI→ST obtains the highest overall accuracy.

On the conflated dataset, the pseudowords a_a, a_o, d_d and e_e have a corresponding CUI definition for each of their possible concepts therefore the accuracy for CUI and CUI→ would be the same for these datasets and is not reported. The pseudowords a_a_i, x_p_p and d_a_m_e do not have a CUI definitions for each of their possible concepts. The results show that CUI obtained the highest accuracy for six out of the seven datasets and CUI→ST obtained the highest accuracy for one. These experiments were run using the unigram feature.

6.2 NLM-WSD Results

Table 3 shows the accuracy of the results obtained by our unsupervised vector approach using the CUI→ST context description, SenseClusters, and the results reported by Humphrey et al. (2006).

As seen with the context description results, there exists a large difference in accuracy on a word by word basis between the approaches. The results show that Humphrey et al. (2006) report a higher overall accuracy compared to SenseClusters and our approach. Although, Humphrey et al. (2006) performed better for 5 out of the 13 words where as SenseClusters performed better for 9. The unigram feature set with gap cluster stopping returned the highest overall accuracy for SenseClusters. The number of clusters for all of the gap cluster stopping experiments were two except for *growth* which returned one. For our approach, the unigram feature set returned the highest overall accuracy.

target word	senses	Maj.	Humphrey	SenseClusters				Our Approach	
			et al. 2006	exact cluster stopping		gap cluster stopping		CUI→ST	
				unigram	bigram	unigram	bigram	unigram	bigram
adjustment	3	66.67	**76.67**	49.46	38.71	55.91	45.16	44.57	47.83
blood pressure	3	**54.00**	41.79	40.00	46.00	51.00	**54.00**	38.38	38.38
degree	2	96.92	**97.73**	53.85	55.38	53.85	55.38	70.31	48.44
evaluation	2	50.00	59.70	**66.00**	50.00	**66.00**	50.00	51.52	51.52
growth	2	63.00	**70.15**	66.00	52.00	66.00	63.00	63.64	63.64
immunosuppression	2	59.00	74.63	67.00	**80.00**	67.00	**80.00**	50.51	43.43
mosaic	2	53.61	67.69	**72.22**	58.57	61.86	50.52	37.50	22.92
nutrition	2	**50.56**	35.48	40.45	47.19	44.94	41.57	25.00	37.50
radiation	2	62.24	**78.79**	69.39	56.12	69.39	56.12	57.73	60.82
repair	2	76.47	86.36	**86.76**	73.53	86.76	73.53	37.31	41.79
scale	2	**100.0**	60.47	**100.0**	**100.0**	**100.0**	**100.0**	51.56	96.88
sensitivity	2	**96.08**	82.86	41.18	41.18	52.94	54.90	48.00	18.00
white	2	54.44	55.00	**80.00**	53.33	**80.00**	53.33	49.44	49.44
average		67.92	**68.26**	64.02	57.85	65.82	59.81	48.11	47.74

Table 3: Accuracy of Approaches using the NLM-WSD Dataset

target word	pseudo-	test	train	Maj.	Sense	Our Approach			
	word				Clusters	CUI	ST	CUI+ST	CUI→ST
actin-antigens	a_a	33193	298723	63.44	**91.30**	53.95	44.81	54.17	
angiotensin II-olgomycin	a_o	5256	47294	93.97	**56.76**	16.62	20.68	17.73	
dehydrogenase-diastolic	d_d	22606	203441	58.57	**95.85**	45.78	43.94	45.70	
endogenous-extracellular matrix	e_e	19820	178364	79.92	71.21	**74.34**	65.37	73.37	
allogenic-arginine-ischemic	a_a_i	22915	206224	57.16	**69.03**	47.68	24.60	33.77	32.07
X chromosome-peptide-plasmid	x_p_p	46102	414904	74.61	**66.21**	20.04	31.60	42.89	42.98
diacetate-apamin-meatus-enterocyte	d_a_m_e	1358	12212	25.95	**74.23**	28.87	24.08	26.07	22.68

Table 4: Accuracy of Approaches using the Conflate Dataset

6.3 Conflate Results

Table 4 shows the accuracy of the results obtained by our approach and SenseClusters. The results show that SenseClusters returns a higher accuracy than our approach except for the e_e dataset.

7 Discussion

We report the results for four experiments in this paper: i) the results of using the IDF stoplist over a basic stoplist, ii) the results of our approach using different context descriptions of the possible concepts of a target word, iii) the results of our approach compared to SenseClusters and Humphrey et al. (2006) using the NLM-WSD dataset, and iv) the results of our approach compared to SenseClusters using the conflated dataset.

The results of using an individualized IDF stoplist for each target word show an improvement over using the basic stoplist. The results of our approach using different context descriptions show that for the NLM-WSD dataset the large differences in accuracy makes it unclear which of the context descriptions performed the best. On the conflated dataset, adding the ST definition to the context description improved the accuracy of only one pseudoword. When comparing our approach to Humphrey et al. (2006) and SenseClusters, our approach did not return a higher accuracy.

When analyzing the data, we found that there does not exist a CUI definition for a large number of possible concepts. Table 5 shows the number of words in the CUI and ST definitions for each concept in the NLM-WSD dataset. Only four target words have a CUI definition for each possible concept. We also found the concept definitions vary widely in length. The CUI definitions in the UMLS come from a variety of sources and there may exist more than one definition per source. Unlike CUI definitions, there does exist an ST definition for each possible concept. The ST definitions come from the same source and are approximately the same length but they are a broad categorization. We believe this makes them too coarse grained to provide descriptive enough information about their associated concepts.

This can also be seen when analyzing the conflate datasets. The conflate dataset d_a_m_e is missing two definition which is a contributing factor to its low accuracy for CUI. Adding the ST definition

target word	CUI Definition			ST Definition		
	c1	c2	c3	c1	c2	c3
adjustment	41	9	48	31	19	10
blood pressure	26	18	0	20	31	22
degree	0	0		15	23	
evaluation	54	0		33	17	
growth	91	91		20	19	
immunosuppression	130	41		30	20	
mosaic	0	38	0	10	10	23
nutrition	152	152	0	10	31	30
radiation	71	207		14	30	
repair	0	51		30	20	
scale	0	10	144	47	23	8
sensitivity	0	0	0	25	50	22
white	0	60		15	28	

Table 5: Number of words in CUI and ST Definitions of Possible the Concepts in the NLM-WSD Dataset

though did not provide enough distinctive information to distinguish between the possible concepts.

8 Conclusions and Future Work

This paper introduces an unsupervised vector approach to disambiguate words in biomedical text using contextual information from the UMLS. Our approach makes disambiguation decisions for words that have the same ST unlike Humphrey et al. (2006). We believe that our approach shows promise and leads us to our goal of exploring the use of biomedical knowledge sources.

In the future, we would also like to increase the size of our conflated dataset and possibly create a biomedical all-words disambiguation test set to test our approach. Unlike SenseClusters, our approach can be used to perform all-words disambiguation. For example, given the sentence: *His weight has fluctuated during the past month.* We first create a instance vector containing $fluctuated$, $past$ and $months$ for the word $weight$ and a concept vector for each of its possible concepts, “C0005912: Body Weight” and “C0699807: Quantitative Concept” using their context descriptions. We then calculate the cosine between the instance vector and each of the two concept vectors. The concept whose vector has the smallest cosine score is assigned to $weight$. We then repeat this process for $fluctuated$, $past$ and $months$.

We also plan to explore using different contextual information to improve the accuracy of our approach. We are currently exploring using co-occurrence and relational information about the possible CUIs in the UMLS. Our IDF stoplist experiments show promise, we are planning to explore other measures to determine which words to include in the stoplist as well as a way to automatically determine the threshold.

Acknowledgments

The author thanks Ted Pedersen, John Carlis and Siddharth Patwardhan for their comments.

Our experiments were conducted using CuiTools v0.15, which is freely available from http://cuitools.sourceforge.net.

References

S.M. Humphrey, W.J. Rogers, H. Kilicoglu, D. Demner-Fushman, and T.C. Rindflesch. 2006. Word sense disambiguation by selecting the best semantic type based on journal descriptor indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technolology*, 57(1):96–113.

M. Joshi, T. Pedersen, and R. Maclin. 2005. A comparative study of support vectors machines applied to the supervised word sense disambiguation problem in the medical domain. In *Proceedings of 2nd Indian International Conference on AI*, pages 3449–3468, Dec.

A. Kulkarni and T. Pedersen. 2005. SenseClusters: unsupervised clustering and labeling of similar contexts. In *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, pages 105–108, June.

G. Leroy and T.C. Rindflesch. 2005. Effects of information and machine learning algorithms on word sense disambiguation with small datasets. *International Journal of Medical Info.*, 74(7-8):573–85.

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 533–37, Chicago, IL, Nov.

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*, volume 1501, pages 1–8, Trento, Italy, April.

A. Purandare and T. Pedersen. 2004. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the Conference on CoNLL*, pages 41–48.

M. Weeber, J.G. Mork, and A.R. Aronson. 2001. Developing a test collection for biomedical word sense disambiguation. In *Proceedings of the American Medical Informatics Association Symposium*, pages 746–750.