Using UMLS Concept Unique Identifiers (CUIs) for Word Sense Disambiguation in the Biomedical Domain

Bridget T. McInnes, MS[1], Ted Pedersen, PhD[2], and John Carlis, PhD[1]

[1] **University of Minnesota, Minneapolis, MN, USA**
[2] **University of Minnesota, Duluth, MN, USA**

Abstract

This paper explores the use of Concept Unique Identifiers (CUIs) as assigned by MetaMap as features for a supervised learning approach to word sense disambiguation of biomedical text. We compare the use of CUIs that occur in abstracts containing an instance of the target word with using the CUIs that occur in sentences containing an instance of the target word. We also experiment with frequency cutoffs for determining which CUIs should be included as features. We find that a Naive Bayesian classifier where the features represent CUIs that occur two or more times in abstracts containing the target word attains accuracy 9% greater than Leroy and Rindflesch's approach, which includes features based on semantic types assigned by MetaMap. Our results are comparable to those of Joshi, et. al. and Liu, et. al., who use feature sets that do not contain biomedical information.

Introduction

Some words have multiple senses, for example, the word $culture$ could mean an anthropological culture, such as the culture of a Mayan civilization, or a laboratory culture, such as a cell culture. The different senses of a word are often obtained from a sense inventory such as a dictionary or other resource. The Unified Medical Language System (UMLS) is one such sense inventory for the biomedical and clinical domain. In the UMLS, senses (or concepts) associated with words and terms are enumerated via Concept Unique Identifiers (CUIs). For example, the two senses of $culture$ are "C0010453: Anthropological Culture" and "C0430400: Laboratory Culture" in the UMLS release 2007AB.

The UMLS is more than just a dictionary of different word senses but also a framework encoded with different semantic and syntactic structures. Some such information includes related concepts, semantic types and semantic relations. A semantic type is a broad subject categorization assigned to a CUI. A semantic relation is the relationship between two semantic types. For example, the semantic type of "C0010453: Anthropological Culture" is "Idea or Concept" while the semantic type for "C0430400: Laboratory Culture" is "Laboratory Procedure". The semantic relations between "Idea or Concept" and "Laboratory Procedure" with the semantic type "Mental Process" are "result_of" and "assesses_effect_of", respectively.

MetaMap[1] maps terms in biomedical text to senses (i.e. concepts) in the UMLS by identifying the CUIs of the content words in the text. MetaMap can be thought of as an all-words disambiguation system, while our approach is focused on particular target words. MetaMap assigns a CUI (sense) to every word or term that it can in a running text using rules and patterns. Our approach is based on supervised learning, where we collect some number of manually disambiguated examples of a given word, and learn a model from that data that only assigns senses to that target word. Thus, MetaMap is a broad coverage tool while our approach is more fine-grained and specific to a few words.

There has been previous work on supervised learning approaches to word sense disambiguation in the biomedical domain. However, some of that work has simply used features that are known to work in general English, and do not take advantage of biomedical information (e.g.,[2],[3]).

On the other hand, Leroy and Rindflesch[4] include the use of biomedical information generated by MetaMap. They analyze combinations of the following features: i) whether the target word is a head word, ii) the part-of-speech (POS) of the target word, iii) the semantic relations between the words surrounding the target word and between surrounding words and the target word itself, and iv) the semantic types of the words or terms surrounding the target word. Leroy and Rindflesch's best reported feature set contains whether the target word is a head word, the POS of the target word and the semantic types of the words in the same sentence as the target word. They use the Naive Bayes algorithm from the WEKA[5] data-mining package and report their results using 10-fold cross validation.

Leroy and Rindflesch's use of features generated from MetaMap lead us the question of *whether CUIs generated by MetaMap would be an improvement over semantic types*?

Joshi, et. al.[3] employ features that have been used in supervised learning of word sense disambiguation for general English, and apply them to the biomedical domain. Their approach utilizes features based on the unigrams and bigrams of the words in the same window of context as the target word. A unigram is a content word that frequently occurs in a window of context around the target word. A bigram is an ordered pair of content words that frequently occur in a window of context around the target word. Joshi et. al. report highly accurate results, especially when their features are unigrams where the window is the same sentence as the target word and unigrams where the window of context is the same abstract as the target word. They compare the Naive Bayesian classifier and Support Vector Machine from the WEKA data-mining package and report their results using 10-fold cross validation.

Joshi, et. al.'s use of unigrams led us to the question *whether the biomedical specific feature CUIs would be an improvement over the more general feature unigrams?*

Joshi, et. al. also compare using unigrams in the same sentence as the target word versus the same abstract. This lead us to the question of *whether increasing the size of the context window in which surrounding CUIs are found improve the results, as seen with unigrams*?

Liu, et. al.[2] utilize combinations of the following features: i) surrounding words, ii) orientation, iii) distance, iv) collocations and v) unigrams. Orientation is whether the surrounding word is to the left or the right of the target word. Distance is how far the surrounding word is from the target word and collocation is a unit of words that represent a single idea, for example, "White House". Their best reported feature set contains all word within a window size of three, their orientation, and the three nearest two word collocations. They compare the Naive Bayes, a modified Decision List and a combination Naive Bayes/exemplar-based algorithm. They report their results using 10-fold cross validation and record the best per word accuracy over all feature sets and algorithms.

We address the above questions by evaluating the following feature sets: the CUIs of the words in the same sentence as the target word and the CUIs of the words in the same abstract as the target word; each with frequency cutoff of one and two. We compare our approach to the previous approaches described above.

Methods

Dataset : We use the National Library of Medicine's Word Sense Disambiguation (NLM-WSD) dataset[6]. This data contains 100 randomly selected instances of 50 frequent and highly 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.

Joshi, et. al. evaluated their approach using 28 out of the 50 target words in the dataset; referred to here as Joshi subset. Leroy and Rindflesch evaluated their approach using 15 out of the 28 words used by Joshi, et. al.; referred to as Leroy subset. Liu, et. al. evaluated their approach using 22 out of the 28 words; referred to as Liu subset. There are nine words that all three authors use to evaluated their approach; referred to as Common subset. There are 22 words that were not used by any of the authors; referred to as the Excluded subset. These words were not used because a large majority of their instances have the same concept. We report our results for all 50 words.

WSD Approach : We explore the use of CUIs as features in a supervised learning approach to word sense disambiguation. For a CUI to serve as a feature, it must occur in the appropriate window of context around the target word more than a specified number of times. Our windows of context include the sentence in which a given target word occurs, or the entire abstract; this would automatically include the sentence that contains the target word as well.

The CUIs of the terms in the same window of context as the target word were obtained from MetaMap and are encoded in the NLM-WSD dataset. The CUIs of the target words were manually assigned.

The frequency cutoffs are implemented by counting the number of times in which a CUI occurs surrounding the target word in the window of contexts. A cutoff of one indicates that we include only those CUIs that occur two or more times surrounding the target word, and a cutoff of two indicates that the CUI should occur three or more times surrounding the target word to be a feature.

So overall, we have four different feature sets: i) the CUIs in the same sentence as the target word with a frequency cutoff of one, ii) the CUIs in the same sentence as the target word with a frequency cutoff of two, iii) the CUIs in the same abstract as the target word with a frequency cutoff of one, iv) the CUIs in the same abstract as the target word with a frequency cutoff of two.

We use the Naive Bayes algorithm from the WEKA data-mining package as our learning algorithm and re-

Table 1: **Accuracy of Approaches Based on 10-fold Cross Validation**

		Our approaches				Previous approaches			
target word	baseline	s-01-cui	s-2-cui	a-1-cui	a-2-cui	s-4-Joshi	a-4-Joshi	s-0-Leroy	s-0-Liu
adjustment	62.0	74.0	68.0	70.0	67.0	70.0	71.0	57.0	
blood pressure	54.0	57.0	56.0	46.0	45.0	62.0	53.0	46.0	
evaluation	50.0	58.0	59.0	73.0	73.0	62.0	69.0	57.0	
immunosuppression	59.0	74.0	74.0	75.0	74.0	72.0	80.0	63.0	
radiation	61.0	78.0	72.0	81.0	81.0	69.0	82.0	72.0	
sensitivity	49.0	81.0	81.0	92.0	92.0	76.0	88.0	70.0	
cold	86.0	85.0	85.0	89.0	89.0	88.0	90.0		90.9
depression	85.0	79.0	76.0	81.0	82.0	87.0	86.0		88.8
discharge	74.0	91.0	90.0	96.0	96.0	82.0	95.0		90.8
extraction	82.0	84.0	83.0	86.0	85.0	84.0	84.0		89.7
fat	71.0	73.0	74.0	77.0	76.0	80.0	84.0		85.9
implantation	81.0	91.0	91.0	92.0	93.0	86.0	94.0		90.0
japanese	73.0	81.0	79.0	76.0	77.0	81.0	77.0		79.8
lead	71.0	92.0	92.0	90.0	91.0	83.0	89.0		91.0
mole	83.0	89.0	87.0	87.0	88.0	98.0	95.0		91.1
pathology	85.0	82.0	79.0	84.0	83.0	88.0	85.0		88.2
reduction	89.0	92.0	92.0	92.0	93.0	93.0	91.0		91.0
sex	80.0	85.0	83.0	87.0	88.0	85.0	88.0		89.9
ultrasound	84.0	87.0	87.0	85.0	87.0	85.0	92.0		87.8
degree	63.0	73.0	73.0	79.0	80.0	92.0	89.0	68.0	98.0
growth	63.0	62.0	60.0	69.0	66.0	63.0	71.0	62.0	72.2
man	58.0	84.0	85.0	80.0	81.0	92.0	89.0	80.0	91.0
mosaic	52.0	73.0	71.0	75.0	75.0	77.0	87.0	66.0	87.8
nutrition	45.0	46.0	43.0	49.0	48.0	63.0	52.0	48.0	58.1
repair	52.0	84.0	81.0	93.0	92.0	72.0	87.0	81.0	76.1
scale	65.0	80.0	78.0	83.0	82.0	80.0	81.0	84.0	90.9
weight	47.0	68.0	69.0	79.0	80.0	80.0	83.0	68.0	78.0
white	49.0	74.0	73.0	74.0	74.0	72.0	79.0	62.0	75.6
Joshi subset	66.8	77.7	76.4	80.0	79.9	79.3	82.5		
Leroy subset	55.2	71.0	69.5	74.5	74.0	73.4	77.4	65.6	
Liu subset	69.9	79.7	78.6	81.9	82.0	82.3	84.9		85.5
Common subset	54.8	71.5	70.3	75.6	75.3	76.7	79.7	68.7	80.8

port the accuracy of our approach using 10-fold cross validation. In 10-fold cross validation, the features come from the entire set of instances in the NLM-WSD dataset. To test our algorithm, the instances are then divided into ten blocks where each block contains an equal number of instances. Then nine blocks are used as a training data and the remaining block is used as test data. The classifier is built using the nine blocks as training data and tested using the remaining block. This is repeated ten times such that each block has been used as test data exactly once with the other nine as training data. The accuracy reported is the average over all ten runs.

Results

Table 1 and Table 2 show the accuracy (%) of using the CUIs of the words surrounding the target word as features into a Naive Bayes algorithm. The label "s-x-cui" refers to the feature set containing the CUIs in the sentences that occur more than x times with the target word. The label "a-x-cui" refers to the feature set containing the CUIs in the abstracts that occur more than x times with the target word. The tables also show the "majority sense" baseline. This is the accuracy that would be achieved by assigning every instance of the target word with the most frequent sense as assigned by the human evaluators.

Table 1 shows the results of previous supervised WSD approaches introduced by Joshi, et. al. (s-4-Joshi and a-4-Joshi), Leroy and Rindflesch (s-0-Leroy), and Liu, et. al. (Liu) for their respective subsets. The labels "s-4-Joshi" and "a-4-Joshi" refers to Joshi, et. al.'s feature set containing the unigrams that occur in the window of context around the target word (sentence and abstract respectively) in more than four training examples. The label "s-0-Leroy" refers to Leroy and Rindflesch's feature set containing semantic types in the sentences that contain the target word; they do not employ a cutoff. The label "Liu" for Liu, et. al.'s feature set refers to the best reported results over all feature combinations and algorithms used by the authors.

Table 2 shows the results of the words that were not evaluated by the previous authors (Excluded subset) and our overall results for the entire NLM-WSD dataset.

In this section, we first compare our feature sets (s-x-cui and a-x-cui) to the baseline and then to each other to determine which window of context and frequency cutoff performed best. Second, we compare the results of our best performing feature set to the best performing feature set of each of the three previous approaches. The $p\text{-}values$ reported are calculated using the one-sided pairwise t-test.

Impact of Different Feature Sets : In this section, we first compare our four different features sets to the majority sense baseline and then analyze the impact the different window of contexts and frequency cutoffs used in our approach. We report the overall accuracy of the feature sets using the entire NLM-WSD dataset; the results of which can be seen in Table 2.

We show that our four feature sets, s-1-cui, s-2-cui, a-1-cui and a-2-cui, increased the accuracy over the baseline by 5.82% ($p \leq .00007$), 4.24% ($p \leq .0037$), 7.56% ($p \leq .00002$) and 7.48% ($p \leq .00002$) respectively.

Table 2: **Accuracy of Our Approaches Based on 10-fold Cross Validation**

target word	baseline	s-1-cui	s-2-cui	a-1-cui	a-2-cui
association	100	100	100	97	97
condition	90	90	86	89	89
culture	89	86	85	94	95
determination	79	76	74	81	80
energy	99	98	98	99	99
failure	71	63	58	73	73
fit	82	88	86	87	86
fluid	100	95	95	99	99
frequency	94	90	89	94	94
ganglion	93	94	93	94	95
glucose	91	83	84	90	90
inhibition	98	98	98	98	98
pressure	96	88	88	93	92
resistance	97	96	95	96	97
secretion	99	93	92	99	99
strains	92	91	92	92	93
support	90	91	90	91	90
surgery	98	94	93	94	93
transient	99	99	99	98	98
transport	93	94	95	93	93
variation	80	83	84	91	90
Excluded subset	92.2	90.4	89.6	92.7	92.6
NLM-WSD dataset	78.0	83.3	82.2	85.6	85.5

When comparing the window of context in which the CUIs are extracted, the results show that increasing the context window from sentences to abstracts improves the accuracy. a-1-cui returns a 2.28% ($p \leq .0006$) increase over s-1-cui, and a-2-cui returns a 3.24% ($p \leq .00000$) increase over s-2-cui.

We show that using a frequency cutoff of one returns a significantly higher overall accuracy than a frequency cutoff of two using sentences as the window of context. This is not the case though when using abstracts as the window of context. s-1-cui results show an increase in overall accuracy of 1.04% ($p \leq .00004$) over s-2-cui while a-1-cui results show a non-significant increase in overall accuracy of 0.08% ($p \leq .2871$) over a-2-cui.

Overall, the results show that a-1-cui, which incorporates a frequency cutoff of one and includes the entire abstract as the window of context, returns the highest overall accuracy.

Comparison with Previous Approaches : Table 1 shows a comparison between our WSD approach and the previous supervised approaches by Joshi, et. al. (s-4-Joshi and a-4-Joshi), Liu, et. al. (Liu), and Leroy and Rindflesch (s-0-Leroy). We make a direct comparison by calculating the overall average of the words from the NLM-WSD dataset that were used by the respective authors for their evaluation.

We compare Joshi, et. al.'s feature sets, s-4-Joshi and a-4-Joshi, with our feature sets, s-1-cui and a-1-cui, respectively. The results using the Joshi subset show that s-1-cui performs equivalently to s-4-Joshi but a-4-Joshi performs slightly better than a-1-cui. The overall accuracy of s-1-cui shows a non-significant decrease of 1.68% ($p \leq .135$) compared to s-4-Joshi whereas the overall accuracy of a-1-cui shows a significant decrease of 2.36% ($p \leq .003$) compared to a-4-Joshi.

We compare Liu, et. al.'s best performing feature set, Liu, with a-1-cui. The results reported by Liu, et. al. are the best per-word accuracies over all algorithms and feature sets examined. The results using the Liu subset show that a-1-cui returns a significant decrease in overall accuracy of 3.62% ($p \leq .01$) compared to Liu.

We compare Leroy and Rindflesch's best performing feature set, s-0-Leroy, with our s-1-cui and a-1-cui feature sets. The results using the Leroy subset show that s-1-cui significantly improves the accuracy of s-0-Leroy by 5.47% ($p \leq .001$) and a-1-cui significantly improves the accuracy by 8.93% ($p \leq .00005$).

Discussion

We pose three questions in this paper: i) whether CUIs are a more effective feature for word sense disambiguation than semantic types ii) whether the biomedical specific feature CUIs are more effective than unigrams in biomedical text, and iii) whether disambiguation accuracy with CUIs improves as the window of context is increased.

We found that CUIs result in more accurate disambiguation than semantic types, and are comparable with unigrams. Thus our answers to i) and ii) are ”yes” and ”no”. We believe CUI's performed better than semantic types because they give a more specific description of the surrounding context.

We hypothesized that CUIs would also perform better than unigrams, however, we found that this was not the case. We believe that this might reflect the fact that the CUIs we used as features were assigned by MetaMap, which is a broad-coverage rule-based tool, and may at times lack sufficient contextual information to assign CUIs correctly. Thus, there is a certain amount of noise in the CUI features, and when used as features they may have mislead the supervised learning algorithm. While unigrams are potentially ambiguous, the supervised learning algorithm may detect patterns among individually ambiguous unigrams that essentially disambiguate each other, and result in a high level of performance. We find it very encouraging though that CUIs assigned by a rule based approach result in a high level of disambiguation accuracy, which suggests that even minor improvements to MetaMap could significantly enhance the power of CUIs as features in approaches such as these.

We also found that a wider window of context results in higher disambiguation accuracy, so our answer to iii) is ”yes”. Joshi, et. al. also reported that using the entire abstract as a source of features improved upon just using the sentence in which the target word occurs, so our findings are in agreement. However, it should be noted that the text from MedLine abstracts is very focused and carefully crafted, so that this finding may well be different for more general biomedical texts. This remains an important issue for future work.

Conclusion

This papers introduces a supervised approach to word sense disambiguation of biomedical text that is based on using CUIs as assigned by MetaMap as features. We compare our approach to previous work by Leroy and Rindflesch, which uses semantic types as assigned by MetaMap as features, and to previous work by Joshi, et. al. and Liu, et. al. that relies on features that are also employed when disambiguating general English text.

Since MetaMap assigns CUIs (and other information) to all terms in running text, we believe that it can serve a very useful role as a generator of features for other systems such as ours. In effect, MetaMap makes a first pass through the data, identifies the terms in the data, and assigns senses (CUIs) to those terms with reasonable accuracy. We have shown that these CUIs can be used as input to a supervised word sense disambiguation systems that is focused on obtaining very high accuracy for a smaller number of words. Over the longer term, we would like our results to feed back into and enhance MetaMap, thus resulting in an iterative approach that combines rule based on supervised learning.

Finally, we remain interested in exploring whether biomedical features such as CUIs and semantic types can result in higher disambiguation accuracy than features that are also used in general English text. We believe that the UMLS and similar resources present a great opportunity for obtaining features that are specific to the biomedical domain and will ultimately enhance the performance of word sense disambiguation and other natural language processing tasks.

In conclusion, the results of this paper make three points. First, incorporating more surrounding context improves results. Second, using the CUIs extracted from MetaMap perform better than semantic types also extracted from MetaMap. And third, using CUIs are comparable to the more general English unigram feature.

Acknowledgments

The authors wish to thank Serguei Pakhomov and Mahesh Joshi for their valuable comments and suggestions.

Bridget T. McInnes’ participation was supported by a Graduate Assistance in Areas of National Need (GAANN) Fellowship to the University of Minnesota.

The experiments in this paper were conducted with CuiTools version 0.05, which is freely available from http://cuitools.sourceforge.net.

Address for Correspondence

Bridget McInnes, University of Minnesota, Department of Computer Science and Engineering, 200 Union Street SE 55455 Minneapolis (MN), USA
bthomson@cs.umn.edu

References

1. Aronson A. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: Proceedings of the American Medical Informatics Association (AMIA) Symposium; 2001. p. 17–21.
2. Liu H, Teller V, Friedman C. A multi-aspect comparison study of supervised word sense disambiguation. Journal of the American Medical Informatics Association. 2004;11(4):320–331.
3. Joshi M, Pedersen T, Maclin R. A Comparative Study of Support Vectors Machines Applied to the Supervised Word Sense Disambiguation Problem in the Medical Domain. In: Proceedings of Second Indian International Conference on Artificial Intelligence; 2005. p. 3449–3468.
4. Leroy G, Rindflesch TC. Effects of information and machine learning algorithms on word sense disambiguation with small datasets. International Journal of Medical Informatics. 2005;74(7-8):573–585.
5. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann; 1999.
6. Weeber M, Mork J, Aronson A. Developing a test collection for biomedical word sense disambiguation. In: Proceedings of the American Medical Informatics Association (AMIA) Symposium; 2001. p. 746–750.