International Journal of Medical Informatics (2005) **xxx**, xxx—xxx

ELSEVIER

International Journal of
Medical
Informatics

www.intl.elsevierhealth.com/journals/ijmi

Effects of information and machine learning algorithms on word sense disambiguation with small datasets

Gondy Leroy[a,*], Thomas C. Rindflesch[b]

[a] *School of Information Science, Claremont Graduate University, 130 E. Ninth Street, Claremont, CA 91711, USA*
[b] *National Library of Medicine, Bethesda, MD, USA*

Received 16 November 2004; received in revised form 17 February 2005; accepted 17 March 2005

KEYWORDS
Word sense disambiguation;
Machine learning;
Naïve Bayes;
Decision tree;
Neural network;
UMLS

Summary Current approaches to word sense disambiguation use (and often combine) various machine learning techniques. Most refer to characteristics of the ambiguity and its surrounding words and are based on thousands of examples. Unfortunately, developing large training sets is burdensome, and in response to this challenge, we investigate the use of symbolic knowledge for small datasets. A naïve Bayes classifier was trained for 15 words with 100 examples for each. Unified Medical Language System (UMLS) semantic types assigned to concepts found in the sentence and relationships between these semantic types form the knowledge base. The most frequent sense of a word served as the baseline. The effect of increasingly accurate symbolic knowledge was evaluated in nine experimental conditions. Performance was measured by accuracy based on 10-fold cross-validation. The best condition used only the semantic types of the words in the sentence. Accuracy was then on average 10% higher than the baseline; however, it varied from 8% deterioration to 29% improvement. To investigate this large variance, we performed several follow-up evaluations, testing additional algorithms (decision tree and neural network), and gold standards (per expert), but the results did not significantly differ. However, we noted a trend that the best disambiguation was found for words that were the least troublesome to the human evaluators. We conclude that neither algorithm nor individual human behavior cause these large differences, but that the structure of the UMLS Metathesaurus (used to represent senses of ambiguous words) contributes to inaccuracies in the gold standard, leading to varied performance of word sense disambiguation techniques.

1. Introduction

Although many words we use in conversation and writing are ambiguous, we usually do not experi-

[*] Corresponding author. Tel.: +1 909 607 3270.
E-mail address: gondy.leroy@cgu.edu (G. Leroy).

1386-5056/$ — see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2005.03.013

ence problems with interpreting these words in context. People seem to take the context of a conversation effortlessly into account and assign the correct meanings to individual words. Such disambiguation, however, is not easily accomplished with automated methods. Since this is a problem for machine translation, information retrieval, thematic analysis, spelling correction, or any type of speech and text processing, researchers have devoted considerable effort to word sense disambiguation (WSD).

WSD techniques choose the correct sense for a word from a predefined set of available senses. Most existing techniques use the surrounding words and specific features of these to learn the correct sense of the ambiguous word. They are usually supervised machine learning algorithms based on large annotated datasets where the correct sense is indicated for each instance. Ide and Véronis [1] provide an overview of WSD from the early years (1950s) to the late 1990s.

We evaluated the effect of different types of symbolic information for terms in medical text by mapping sentences to the Unified Medical Language System (UMLS). We used small datasets to evaluate how much this knowledge base can contribute when few examples are available. For our first set of tests, we used a naïve Bayes classifier. We continued our study with the best condition by comparing with a neural network (feedforward/backpropagation) and a decision tree algorithm. Accuracy was similar for all three, but the variance between different words was very large. We then tried to discover why the variance was so high. We believe that it may be the different meanings available in the UMLS (a compilation of vocabularies not intended as a WSD resource) which led to the confusion in compiling the gold standard used for learning. Using individual expert's gold standards or specific gold standard characteristics could not explain the variance.

2. Word sense disambiguation

There exist many techniques that are used for word sense disambiguation. Which one is chosen depends on the final goal, the available information per word, and the number of available examples. In some cases, it is sufficient to distinguish between different meanings of words, without having to label the words. For example, a label may be unnecessary when clustering documents together that have similar topics. Schütze [2] labels this task as ''word sense discrimination''. He distinguishes this from ''sense labeling'' where each sense receives the correct label. This distinction often — but not always — coincides with unsupervised (discrimination) versus supervised (labeling) machine learning techniques.

2.1. Approaches to word sense disambiguation

2.1.1. Unsupervised learning techniques

Unsupervised learning algorithms learn patterns solely from input parameters without trying to match to pre-specified categories. In the case of word sense disambiguation, they learn to group words based on the information in the feature sets. But there is no label specified in advance for the group nor is the number of possible groups specified. Assigning a specific meaning can still be achieved by finding the common theme in the established clusters and mapping these to established meanings for the word in a dictionary or other knowledge source. This mapping can be done by a human or automatically based on similarity metrics.

Clustering techniques are especially useful for this type of disambiguation. For example, Pedersen and Bruce [3] tested three unsupervised learning algorithms: Wards and McQuitty's clustering and the EM algorithm. They mapped these clusters to dictionary senses so that there was maximal agreement.

2.1.2. Supervised learning techniques

Supervised learning is used more often for WSD. These techniques rely on outcome feedback provided to an algorithm so that it can take corrective action during its learning or training phase. The possible outcomes are known in advance and algorithms need to learn to combine a particular input with such an output. In the case of word sense disambiguation, the input usually consists of features of the ambiguous word and surrounding text. The output is the correct sense for the word. During the learning phase, supervised techniques learn to associate these feature sets with one particular sense of a limited list of provided senses. This happens by providing the techniques with feedback on its decision for every example. The supervised learning techniques rely on a training set comprised of example ambiguous words and their correct sense. Decision trees, such as ID3 or C4.5, artificial neural networks (ANN), such as the feedforward/backpropagation ANN, and probabilistic-based methods, such as naïve Bayes, are commonly used.

Mooney [4] tested seven such supervised learning methods with the word *line*. His work demonstrates the importance of a large dataset. The input information consisted of surrounding words, a bag-of-words approach. He tested naïve Bayes, a perceptron, and a decision tree, among others. Naïve Bayes was a top performer for both accuracy and the amount of training time required. With 1200 examples, the accuracy was more than 70%. It was less than 60% accurate when trained on only 300 examples. In addition to different algorithms, the amount, relevance, and precision of the information affect performance. Hoste et al. [5] manipulated the feature information and algorithm settings in an extensive set of studies. They argued that algorithms with default settings do not provide a sufficient base for comparison. They demonstrated that different settings and different types of information lead to large variances in accuracy. Pedersen [6] evaluated the use of bigrams (sequences of two words) for WSD with a decision tree and naïve Bayes classifier. He tested different bigrams that occur close to the ambiguous words (within approximately 50 words to the left or right of the ambiguous word) as possible disambiguation features. The decision tree with the most accurate disambiguation was based on bigrams selected with a power divergence statistic (a goodness-of-fit measure).

Although individual algorithms perform well when the datasets provide sufficient examples, combining several together improves accuracy. Florian et al. [7] worked with the Senseval-2 dataset (www.itri.brighton.ac.uk/events/senseval/). They started with Bayes-based methods and used an enriched bag-of-words technique that included a weighted bag-of-lemmas and local *n*-gram context with specific syntactic relations. Their approaches were among the top performers for English (approximately 65% accuracy) and the best for Spanish, Swedish, and Basque. In later studies [8], they combined different types of classifiers, such as vector-based methods (e.g., naïve Bayes), variance-based methods (e.g., Maximum Variance Correction), and Brill's transformation-based learning. They also evaluated different feature spaces such as words, lemmas, and part-of-speech tags in different settings such as traditional bag-of-word approaches, but also local bigram, trigram, and other syntactic relationships. They combined their classifiers with five different voting schemes and found that combinations always outperformed individual classifiers.

2.1.3. Additional disambiguation techniques

In addition to the classic learning algorithms, there are several approaches that rely on heuristics, rules, statistics, or a combination of these. Many researchers develop rules to assign words to a specific sense based on the semantic similarity between a word and sense. The rules are often a combination of statistics and insights by the researchers.

For example, Mihalcea and Moldovan [9] based their approach on semantic density between words and focus on verb—noun pairs. Their distance measure is based on an evaluation of common words between two sets of words. They use WordNet as their knowledge source and calculate statistics for the most probable senses using the Internet as a corpus. Hoste et al. [5] describe a memory-based learning approach where the algorithm keeps all training examples in memory. A classification decision is made based on similarity between new input and stored examples. MetaMap, provided by the National Library of Medicine, uses rules to map between words in the text and UMLS Metathesaurus [10] and provides a score to indicate the fit of each mapping.

2.2. Information sources

Word sense disambiguation techniques, both supervised and unsupervised, sometimes use only the presence or absence of words surrounding the ambiguous word as input information. This is called a bag-of-words approach. In this case, information about the co-occurrence of the ambiguous word with others is used to determine its correct sense. However, quite often, an external source of information is used to provide more advanced features, such as part-of-speech of the ambiguous word itself or surrounding words.

A popular information source for general text is WordNet, a general-English lexical resource [11]. It is frequently used for both its semantic and syntactic information to disambiguate words in general texts. For example, Inkpen and Hirst [12] used WordNet to disambiguate near-synonyms in dictionary entries. Their supervised learning techniques (C4.5, a decision tree algorithm) were based on the overlap of words in the dictionary description and the WordNet glosses, synsets, antonyms, and polysemy information. They achieved 83% accuracy. Santamaría et al. [13] associated Web directories from the Open Directory Project with WordNet synsets with 86% precision. Multiple directories could be assigned to multiple senses. They used vector representations for the surrounding words. Co-occurrence-based comparisons were used to select senses that were closely related to directories. They tested their approach on the Senseval-2 dataset and evaluated whether directories were correctly assigned to words. Magnini et al. [14]

also used WordNet for the Senseval-2 dataset but extended it by adding domain names such as *Medicine* or *Architecture* to every synset. They assigned these to a subset of words in the text based on frequencies of the domains and a few additional rules. Then, with a vector-based approach using a window of 100 surrounding words, they achieved 75% precision in their best conditions.

The general medical domain has recently been the focus of WSD research. Here, the UMLS [15] is a readily available resource to provide syntactic and semantic information. For example, Liu et al. [16] focused on ambiguous abbreviations. They first created a gold standard automatically for the ambiguous abbreviations. For each abbreviation, they retrieved related concepts from the UMLS Metathesaurus. If all related concepts were associated with a particular sense of the ambiguous abbreviation, this sense was accepted as the correct one. Otherwise, the sense with the most associations with these related concepts was deemed correct. A naïve Bayes classifier using stemmed words learned from the gold standard. Ruch et al. [17,18] used another UMLS component, the Semantic Network, to improve WSD. They evaluated a Hidden Markov Model augmented by Semantic Network information to improve spelling correction in medical text with WSD. However, the word sense disambiguation module had no effect on their overall accuracy for spelling corrections.

In biomedicine, WSD has been applied to specific categories of words such as DNA, RNA, and proteins. Hatzivassiloglou and Duboué [19] used three supervised learning techniques, C4.5 decision trees, naïve Bayes, and inductive learning. They tested different features with an automatically created gold standard to distinguish between genes, proteins, and mRNA. Their best technique, naïve Bayes, achieved 84% accuracy. Ginter et al. [20] developed their own algorithms based on feature vectors and frequency of word overlap for a similar task. They compared this with other classic algorithms and achieved the highest accuracy with their own algorithms, 2—5% higher than the best naïve Bayes classifier. They used more than 200,000 documents. Liu et al. [21] evaluated different feature sets and classifiers in an extensive study to disambiguate biomedical abbreviations with automatically created gold standards. They trained their classifiers per abbreviation and achieved high accuracy (over 90%) especially when there were thousands of examples from which to learn.

Word sense disambiguation research and the need for it are not limited to the domains described above. For example, recognizing individuals who use different aliases or different versions of their name, or distinguishing between individuals who have the same name are useful for law enforcement and for intelligence agencies. Han et al. [22] addressed this problem in a similar context, namely that of scientific citations. They used a Bayesian approach and also support vector machines (SVM) with two datasets: a list of Web pages with author publication lists for several researchers with the same name and the DBLP Web site (http://dblp.uni-trier.de/). They tested several combinations of input data comprised of co-author information, words in the paper title, and words in the journal title. Both approaches achieve high accuracy, more than 90%, and it was especially information about co-authors that proved useful in this disambiguation task.

3. Research question

Many learning algorithms for WSD have been tested for both generic and domain specific topics. One common aspect of this research is the use of large datasets for learning (training). Each dataset consists of hundreds of examples vetted by domain specialists who indicate the correct meaning for all targeted words and so construct the gold standard. For example, Mooney [4] used 300, 600, and 1200 examples for training and showed that performance increased with more examples. Hoste et al. [5] argues that increasing the size of the gold standard, e.g., by a factor of 1000, has more effect on performance than individual algorithm biases. Alas, compiling such gold standards is time-consuming and difficult. Some researchers have built gold standards automatically [16,19,21] to sidestep the difficulty of finding experts to create them. These standards are an excellent approach to comparing different algorithms. However, because they are systematically built, they deviate from the standard human experts would establish. This is illustrated by Hatzivassiloglou and Duboué [19], who asked human experts to assign labels to the same terms as in the artificial gold standard (the disambiguating terms were deleted). The pair-wise agreement of the experts was 78%. The question remains whether the artificial standards are more or less correct and suitable than the human created ones.

The notion driving this project is the use of smaller gold standards for machine learning approaches to WSD. In particular, we investigate whether the explicit use of human knowledge allows algorithms to perform as well with a small gold standard as with a large one. The hypothesis is that by supplying algorithms with additional, external knowledge, comparable to the knowledge of the experts who compiled the gold standards, fewer examples will be needed for learning. In this way, our approach, if successful, may augment existing approaches.

4. Word sense disambiguation study with naïve Bayes classifier

4.1. Dataset

This study was performed with a dataset provided by the National Library of Medicine (available from http://wsd.nlm.nih.gov/), in which eleven human evaluators disambiguated words occurring in MEDLINE abstracts [23]. The dataset contains 50 English terms, such as *cold* or *growth*, which are commonly ambiguous. Each ambiguous term is mapped to multiple UMLS concepts. For each, 100 instances were disambiguated by indicating the correct sense with a UMLS concept or the option ''None'' if no UMLS concept described the correct sense.

Each instance is provided with its original MEDLINE abstract, and linguistic and symbolic knowledge is made available for all terms in the entire abstract. MetaMap [10] (available at http://mmtx.nlm.nih.gov/) was used to provide the linguistic information, e.g., part-of-speech (POS), and to map all terms to UMLS concepts and semantic types. All these mappings are provided in the online dataset. We limited our input data for the WSD classifiers to only those mappings (described below) that can be made based on the words occurring in the same sentence as the ambiguous word.

4.2. External knowledge source

We chose the UMLS [15] Semantic Network as our external knowledge source and tested which portions of this network help disambiguate words automatically. In considering a sentence containing an ambiguity, we use the symbolic representation of that sentence in the UMLS Semantic Network [24] and do not use the actual words surrounding the ambiguous term.

Our goal was to train a machine learning technique that can disambiguate the words by choosing the correct mapping. Each mapped concept is also connected to semantic types in the UMLS Semantic Network. We used these to represent the different meanings of ambiguous terms. For example, based on the UMLS, there are three senses and their related semantic types for *blood pressure*. One extra sense (none of the above) was added to be used when none of the previous meanings was correct. The resulting UMLS concepts and semantic types are: Blood Pressure (Organism Function), Blood Pressure Determination (Diagnostic Procedure), Arterial Pressure (Laboratory or Test Result), and none of the above.

4.3. WSD classifier

For our initial study, we chose a naïve Bayes classifier since it was a top performer in several other WSD studies. A naïve Bayes classifier is based on Bayes' probability rules; it takes all presented information into account and is called naïve because it assumes independence between all the features presented to it. We used the Weka software packet to train and test the classifier with 10-fold cross-validation [25]. Follow-up studies included algorithms that represent the different paradigms of decision trees (C4.5) and neural networks (FF/BP) (see below).

4.4. Study design

We report here on experimental conditions in which different combinations of UMLS Semantic Network symbolic knowledge are used. A subset of this work has been presented at Medinfo [26]. The first two conditions (not reported here) used a minimal set of linguistic information about the ambiguous word itself. All other knowledge, added in subsequent experimental conditions, is based exclusively on the sentence in which the ambiguous word appears. The intuition is that more complete symbolic information about the ambiguous word, its context, and how the word interacts with this context will lead to better disambiguation. Fig. 1 illustrates the relationship between the available symbolic knowledge and the experimental conditions.

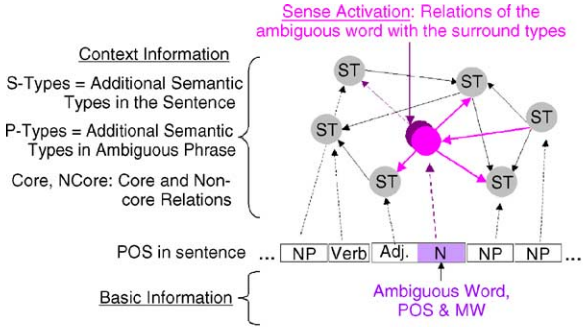
Sense Activation: Relations of the ambiguous word with the surround types
Context Information
S-Types = Additional Semantic Types in the Sentence
P-Types = Additional Semantic Types in Ambiguous Phrase
Core, NCore: Core and Non-core Relations
ST
ST
ST
ST
ST
ST
POS in sentence
... NP Verb Adj. N NP NP ...
Basic Information
Ambiguous Word, POS & MW

Fig. 1 Symbolic knowledge used (ST: semantic types).

- *Basic semantic context* (*condition 1*): the first conditions provide information about the word status, the part-of-speech, and the semantic context. The word's status in the phrase, single word or head of the phrase, is denoted as *main word* (MW). We also use the ambiguous word's part-of-speech. Information about the semantic types of the words occurring in the same phrase with the ambiguous one (PTypes) or elsewhere in the sentence (STypes) is also included. For each of these types we specify how many occurrences there are in the phrase or sentence.
- *Semantic context relations* (*conditions 2a and 3a*): the following conditions add details to the surrounding context of the ambiguous word by adding core (Core) and non-core (NCore) relations. These are Semantic Network relations between the unambiguous semantic types found in the sentence. The UMLS Semantic Network has 54 relations that can exist between 135 semantic types. We considered seven relations to be core relations because they closely link concepts in a hierarchical fashion: is a, conceptual part of, consists of, contains, ingredient of, part of, and process of. We counted the number of such relations (both core and non-core) that exist between any two semantic types found in the context.
- *Normalized semantic context relations* (*conditions 2b and 3b*): to take the granularity of the UMLS Semantic Network into account, we normalized the context relation information (previous conditions) by dividing the number of relations between the pairs of semantic types by the total possible relationships for the individual semantic types. If many relations exist for a semantic type, e.g., 15, but only a few are found based on the pairs of semantic types found in the sentence, e.g., 5, a relative number (5/15) will be more representative when comparing this with a semantic type for which all possible relations are found in the sentence (5/5).
- *Sense activation* (*conditions 4a and 5a*): we evaluated how each ambiguous sense fits into its surrounding context. For this, we added the semantic relations that each ambiguous type can have with its surrounding types (sense activation) as a feature to be used by the classifier. The rationale was that the correct sense would have more interaction with its surroundings.
- *Normalized sense activation* (*conditions 4b and 5b*): comparable to the normalization of the context, we normalized the sense activation by dividing the number of relations found by the number of possible relationship for the particular sense of the ambiguous word.

4.5. Study results

We selected 15 words from the NLM dataset for which the most frequent sense in each case was correct in less than 65% of the instances. This *majority sense performance*, also called *lexical default* [5], served as the baseline for our study. We choose 65% because others have found that high majority sense results in a very skewed dataset that provides insufficient examples to automatically learn from [3]. As mentioned above, additional information is added in each condition. For easy reference, we have numbered the conditions, e.g., the baseline is (0). Table 1 provides an overview of the accuracy for each word. The bottom two rows in the table provide the results for pair-wise *t*-test between the experimental conditions and the baseline (baseline comparison) and between consecutive experimental conditions (incremental comparison, e.g., 0 versus 1, 1 versus 2).

4.5.1. Basic semantic context

In the condition (1), we evaluated the combination of linguistic information with semantic context information. When the semantic types of all unambiguous words in the entire sentence (1) were available for learning, average accuracy was at its peak (66%). This condition was significantly more accurate than the baseline. For some words, disambiguation accuracy increased by 20—30% compared to the baseline.

4.5.2. Semantic context relations

In the following conditions, we added the semantic relations between the unambiguous semantic types that form the context. In conditions (2a), the non-core relations are added, while in (3a) both core and non-core relations are added. Including information about non-core relations (2a) has a significant adverse affect on accuracy. The core relation information had a small beneficial effect for some words, but the effect was not significant. Performance was not better than the baseline and drastically decreased compared to condition (1) with only semantic types.

We then tested whether a normalized representation of these relations in the Semantic Network that was either more detailed or more accurate would improve the results. Conditions (2b) and (3b) are similar to the previous two, but the simple counts for relations were replaced with numbers that take the granularity of the Semantic Network into account. We tested simple division, percentages, and logarithms of the division. The logarithm-based set resulted in the best performance and

Table 1 Accuracy (%) of the naïve Bayes classifier for word sense disambiguation

Word	Information provided to classifier context relations									
	Base		Added semantic context relations				Added sense activation			
	Majority sense (0)	MW, POS, PTypes, STypes (1)	Basic		Normalized		Basic		Normalized	
			MW, POS, PTypes, STypes, NCore (2a)	MW, POS, PTypes, STypes, NCore, Core (3a)	MW, POS, PTypes, STypes, NCore (2b)	MW, POS, PTypes, STypes, NCore, Core (3b)	MW, POS, PTypes, STypes, NCSA (4a)	MW, POS, PTypes, STypes, NCSA, CSA (5a)	MW, POS, PTypes, STypes, NCSA (4b)	MW, POS, PTypes, STypes, NCSA, CSA (5b)
Adjustment	62	57	50	51	34	32	48	50	44	44
Blood pressure	54	46	56	54	37	37	48	48	48	46
Degree	63	68	60	59	58	58	67	70	70	69
Evaluation	50	57	53	55	54	57	53	54	56	56
Growth	63	62	50	50	65	64	56	60	58	58
Immunosuppression	59	63	61	64	65	63	67	65	64	65
Man	58	80	62	66	68	68	70	70	72	72
Mosaic	52	66	42	42	50	50	52	56	48	55
Nutrition	45	48	37	39	39	37	38	40	48	47
Radiation	61	72	54	54	58	59	63	62	63	64
Repair	52	81	68	62	56	57	70	69	70	69
Scale	65	84	72	71	68	67	71	72	75	75
Sensitivity	48	70	65	66	72	70	70	70	73	71
Weight	47	68	54	53	56	56	62	59	62	63
White	49	62	48	50	54	55	59	59	61	64
Average	55	66	55	56	56	55	60	60	61	61
Baseline comparison—*t*-test, α: 0.05, *p*-value		(0 vs. 1) <0.005						(0 vs. 5a) <0.05		(0 vs. 5b) <0.05
Other comparison—*t*-test, α: 0.05, *p*-value			(1 vs. 2a) <0.001		(1 vs. 2b) <0.001	(1 vs. 3b) <0.001	(1 vs. 4a) <0.001		(1 vs. 4b) <0.05	(1 vs. 5b) <0.05

is reported here. However, thus normalizing the scores did still not improve accuracy.

4.5.3. Sense activation

Since performance was lowered so much by adding the relations between semantic types (conditions 2—3), we decided not to pursue them further, but rather to add information about sense activation (3—4) directly to condition 1 (context). Sense activation consists of the relations that the different ambiguous types can have with the unambiguous context. Sense activation based on non-core relations (3a) had a significant adverse effect on accuracy when compared to condition (1) and was not significantly better than the baseline. Adding additional core sense activation did not improve the overall accuracy, but seems to have made results somewhat more consistent. The average accuracy is significantly better than the baseline and not significantly worse than condition (1).

5. Follow-up studies

We assumed that more symbolic information would be better, but this was not the case. The best condition was found when the semantic types were added without detailed information about mappings to relations between types in the UMLS Semantic Network (condition 1). However, there was large variability in the results. Several words responded well to the experimental conditions, while others did not. For example, *repair* had almost 30% increased accuracy in condition (1) compared to the baseline, but the accuracy for *blood pressure* was actually lower in condition (1) than in the baseline. We performed three sets of follow-up studies to try to explain these results.

5.1. Machine learning algorithms

First, we tested machine learning algorithms from different paradigms using the best condition found with the naïve Bayes classifier (condition 1). The purpose of this study was to investigate whether other algorithms would perform better. If they showed different performance characteristics for the 15 words, it would be worthwhile to evaluate different settings for the different conditions.

We chose a decision tree algorithm and a neural network (feedforward/backpropagation) because they are two different paradigms and others have found excellent results with them. We performed the test for the same 15 ambiguous words.

As with the naïve Bayes algorithms, we used Weka to test the other algorithms. J48 is Weka's implementation of C4.5, a decision tree algorithm that can handle continuous values. We used the basic settings (unpruned tree) [25]. The neural network is a feedforward/backpropagation network It was trained over 500 generations, had 3 layers (input, hidden, and output layer), and a learning rate of 0.3.

The resulting performance in this condition (1) is very similar for all three approaches as can be seen in Fig. 2. The average accuracy was 66% with the naïve Bayes classifier, 65% with the decision tree, and 66% with the neural network. There is little variance in performance for the different words.

5.2. Gold standard characteristics

The three algorithms showed very similar performance for all words. A potential explanation for the variability in the results may be related to the gold standard used for learning. We looked at several gold standard characteristics. If these can be associated with performance, future WSD accuracy could be predicted based on such characteristics.

We tested six metrics that describe the data set and correlated these with the accuracy of the WSD algorithm using the Pearson product moment correlation. These metrics all describe somehow the number of examples available for a sense or the diversity of the input.

- Number of choices = the number of possible meanings for the word.
- Smallest category size = the number of instances for the least frequent sense.
- Number of PTypes = the average number of different semantic types in the same phrase as the ambiguous word.
- Number of STypes = the average number of different semantic types in the sentence but not in the same phrase as the ambiguous word.
- Total number of types = the average number of different semantic types in sentence (sum of previous two).

Table 2 provides an overview of the results. None of the correlations were significant. Superficial metrics do not explain the variability in accuracy.

Our gold standard was developed by multiple experts and may display variability and inconsistencies because of the consensus that needed to be reached. If this is the case, we expect that gold standards based on individual expert evalua-

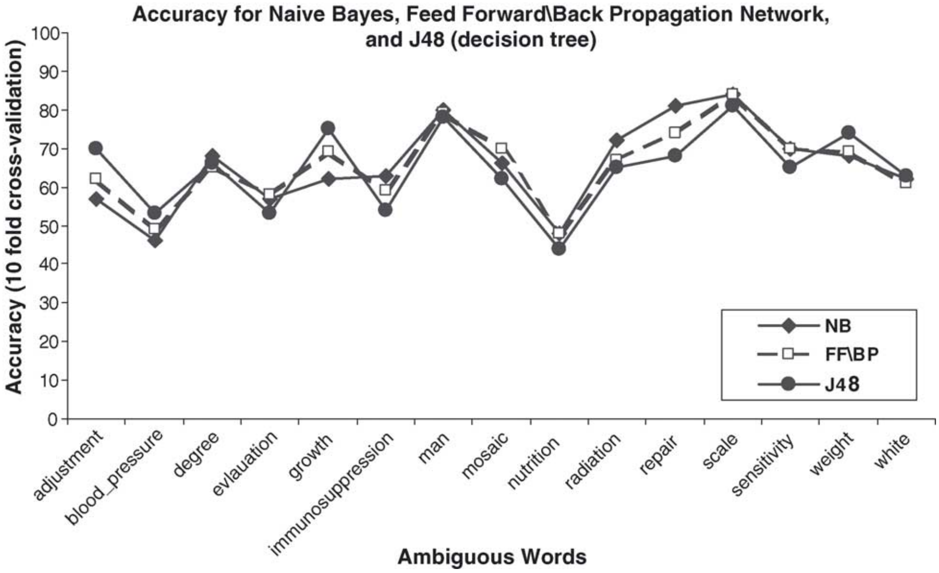

Fig. 2 Accuracy of naïve Bayes (NB), a decision tree (J48), and neural network (FF/BP).

Table 2 Pearson product moment correlation

Pearson correlation	Output information		Input variables		
	No. of choices	Small category size	No. of PTypes	No. of STypes	No. of types
NB	−0.10	−0.01	0.07	−0.29	−0.20
FF/BP	0.04	0.03	0.13	−0.25	−0.12
J48	0.00	−0.01	0.04	0.03	0.06

tion would be more consistent and lead to better results. To test this, we determined whether our original classifier (naïve Bayes) behaved differently for each individual expert's gold standard.

Table 3 provides an overview of accuracy for the combined gold standard (used in all of the above) and for each expert separately for naïve Bayes. Two experts did not evaluate all ambiguities and for one there was no individual data at all. Two gold standards led to significantly worse results (GS2 and GS11). Although some seemed to result in better performance, the results were not significant, due to relatively high variance in the data.

These results show clearly that the average accuracy is not better for individual experts than for the combined, consensus gold standard.

5.3. Troublesome instances

Finally, we sought to define why some instances were more troublesome to the classifiers than others. We evaluated whether there was a relation between the baseline performance for each word, the ambiguity in the instances for each word, and the actual accuracy. Fig. 3 shows our expectations for accuracy determined by baseline accuracy (part A) and example ambiguity (part B).

When the baseline accuracy is low (part A), one would expect improvement to be easier to achieve because there are more examples to learn from per sense (the baseline is the maximum percent correct from one sense) and because there is more room for improvement. Such relations have been reported in [5]. For example, if the baseline ambiguity were high (e.g., 95/100), it would be hard for any algorithm to learn to correctly classify the additional five cases. These five cases would very likely be di-

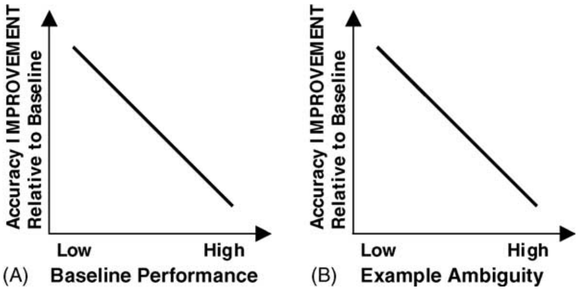
Accuracy I MPROVEMENT Relative to Baseline
Low
High
(A) Baseline Performance
Accuracy I MPROVEMENT Relative to Baseline
Low
High
(B) Example Ambiguity

Fig. 3 Expected improvement in accuracy.

Table 3 Accuracy (%) of naïve Bayes per gold standard (GS) (no separate data for GS8)

Word	MW, POS, PTypes, STypes, naïve Bayes classifier										
	Combined	GS1	GS2	GS3	GS4	GS5	GS6	GS7	GS9	GS10	GS11
Adjustment	57	51	57	56	56	50	63	55	72	57	40
Blood pressure	46	70	51		48	57	42	79	100	56	44
Degree	68	68	68		68	68	71	63	72	68	63
Evaluation	57	72	60		62	75	58	41	64	56	57
Growth	62	66	59		70	59	60	66	64	73	56
Immunosuppression	63	68	47		62	59	66	73	60	71	
Man	80	74	80		78	79	79	81	80	82	68
Mosaic	66	47	57		44	53	65	47	68	68	46
Nutrition	48	56	43		45	67	62	44	78	53	48
Radiation	72	71	70	52	59	61	72	63	62	60	
Repair	81	78	69		76	79	74	78	74	78	
Scale	84	73	75		66	61	70	80	73	98	
Sensitivity	70	69	57		68	69	67	58	69	74	
Weight	68	69	69		66	57	62	66	68	67	59
White	62	62	63		66	66	68	58	60	60	64
Average	66	66	62	54	62	64	65	63	71	68	55
Pair-wise t-test, α: 0.05, with ''combined'', p-value			<0.05								<0.05

vided over the test and training sets, leaving less than five cases to learn from.

For clear, unambiguous examples, the ambiguity is low (part B) and one would expect better learning and so better performance. We define example ambiguity as the number of choices multiplied by the disagreement between experts. For instances with high example ambiguity, one expects lower performance. For example, if there are many closely related, possible senses, it will be difficult to learn the difference between them. The NLM dataset contains information about the evaluation of all 100 instances of each word by the 11 experts. In some cases, the experts did not agree on the correct sense of a word and only chose one sense after extensive discussion. Those requiring discussion are reported as *unresolved counts*. We labeled words with many senses and unresolved counts as words with high example ambiguity (numbers were multiplied).

To visualize these ideas, we ordered the 15 ambiguous words based on their baseline score (Fig. 4)

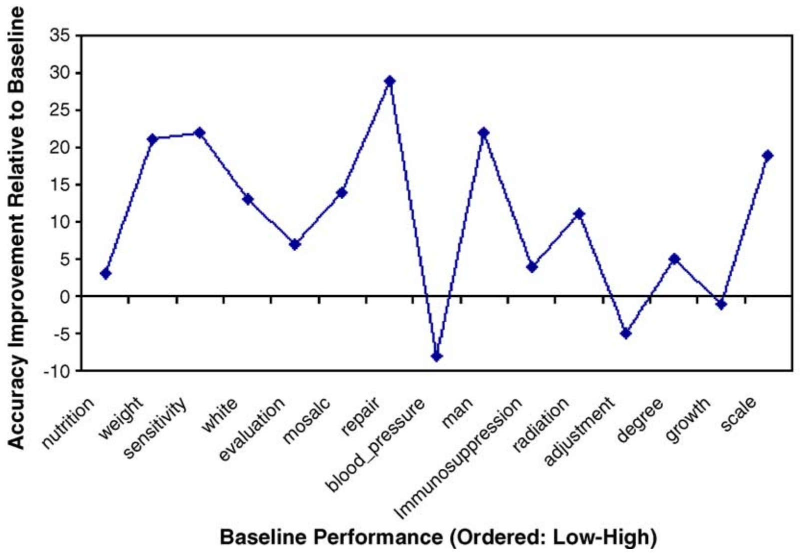

Fig. 4 Actual improvement in accuracy (baseline—ordered).

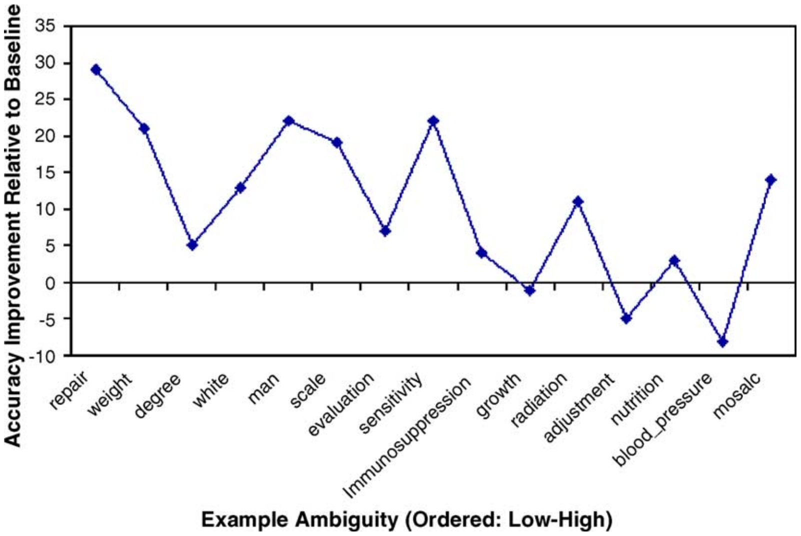

Fig. 5 Actual improvement in accuracy (ambiguity—ordered).

as well as based on the example ambiguity score (Fig. 5). We measured the percentage improvement as the improvement in accuracy for the best experimental condition (condition 1) compared with the baseline.

Fig. 4 shows the actual performance improvement for the words ordered by their baseline performance. This figure should be compared against our expected results in Fig. 3A. There is no improvement with a lower baseline (no significant correlation). However, actual performance seems to decrease when the example ambiguity is higher (Fig. 5).This result looks similar to our expected results in Fig. 3B. Although this is a small test set, a trend can be seen for words with lower example ambiguity (left side) to have higher performance scores. Words with higher example ambiguity (right side) tend to have lower performance scores. We tested the correlation with the Pearson coefficient (one-tailed, since the direction is known) and found a strong trend ($r = -0.379$, $p = 0.8$). If we exclude the last word (mosaic), the correlation is significant ($r = -0.725$, $p < 0.01$).

6. Discussion

We started our work by testing several different conditions for their ability to provide information for automated WSD with a naïve Bayes classifier. Although we expected that more information and more correct information would improve the accuracy, this was not the case. We found that the best accuracy was achieved with only information about the non-ambiguous semantic types assigned to words in the sentences. Adding more information had a negative effect.

The WSD worked extremely well for some words but not for others. To exclude the potential explanation that this variability was due to the particular algorithm we had chosen, we tested two additional algorithms, a decision tree and a neural network. The results were comparable. We then proceeded to look more closely at the gold standard and tested the original algorithm for gold standards based on each expert's opinion. The original gold standard represented the combined evaluation of 11 experts. Again, the results were comparable. Finally, we looked closely at the individual words and associated accuracy. For some words, experts disagreed on their correct meaning. Although the classifiers did not receive any such information as input, they showed the same trend in the resulting accuracy.

We found that effectiveness of the classifier is at least partially dependent on the representation of the senses of the ambiguous words, which in turn is based on the choices available in the underlying dictionary. The NLM test collection uses UMLS Metathesaurus concepts to represent ambiguous senses. The Metathesaurus was not designed as a dictionary, but rather is a compilation of terminologies used for differing purposes. The meaning of terms in these terminologies does not necessarily reflect meaning as encountered in academic text such as the MEDLINE citations on which the test collection is based.

- For example, *blood pressure* is not ambiguous in normal English usage. One standard dictionary (Random House, College Edition) has only a single meaning for this term: ''The pressure of the blood against the inner walls of the blood vessels''. Dorland's Illustrated Medical Dictionary gives a similar definition (along with an explanation of the functional underpinnings of the phenomenon). However, as noted above, the senses allowed for *blood pressure* in the NLM test collection are represented by three Metathesaurus concepts and corresponding semantic types: M1—-''Blood Pressure'' (Organism Function); M2—''Blood Pressure Determination'' (Diagnostic Procedure); M3—''Arterial Pressure'' (Laboratory or Test Result). The Metathesaurus thus represents the senses of *blood pressure* as the phenomenon itself (M1), the procedure for determining the value of the phenomenon (M2), and the result of the determination (M3). The availability of these three ''senses'' invites distinctions where none exists. Often a particular sense is assigned to an ambiguity instance that reflects the meaning of the phrase containing that instance. For example, the phrases *ambulatory blood pressure* and *clinical blood pressure* denote the result of blood pressure measurement, although the ambiguity itself refers only to the phenomenon. In both instances, the judges assigned M3 (Laboratory or Test Result) as the sense of *blood pressure* in these phrases. Similarly the phrase *blood pressure monitoring* was assigned M2 (Diagnostic Procedure).

- For *adjustment*, two of the three senses available are not distinguishable in a principled way. Both M1 and M3 refer to the psychological state of being ''well adjusted''. Although M1 has synonym ''Individual Adjustment'' (with semantic type Individual Behavior) and M3 has synonym ''Psychological Adjustment'' (with semantic type Mental Process), the definitions for both are almost identical.
- A similar situation is seen with *growth*. The Metathesaurus concepts available for disambiguation encourage a distinction between the growth of an entire organism and the development of other entities, such as cells and body parts: the M1 sense of *growth* has semantic type Organism Function. Normal English usage does not make this distinction.

Although we have not examined all ambiguities regarding representation of meaning in the Metathesaurus, the three terms that scored lower than the baseline (*adjustment*, *blood pressure*, and *growth*) have senses represented infelicitously in the Metathesaurus. When these ''senses'' are applied frequently, they have a negative effect on the accuracy of the classifier. Although the M2 sense of *blood pressure* was assigned to only 2 of the 100 ambiguous instances, the M3 sense was used to disambiguate 44 instances. For *adjustment*, M1 was assigned to 18 ambiguous instances and M3 to 13. The M1 sense of *growth* appears 37 times in the 100 instances.

This phenomenon observed in the NLM WSD test collection is related to high sense granularity in WordNet, which also interferes with effective WSD. Magnini et al. [14] demonstrate that collapsing multiple senses to a set of senses belonging to a particular domain can address this problem. Although that exact solution is not relevant here, it would be profitable to investigate collapsing spurious senses in both the NLM training and testing data before applying the naïve Bayes classifier.

7. Conclusion

The purpose of this study was to discover if symbolic knowledge can be used by machine learning algorithms so that it can be added to the common, bag-of-words approaches and so facilitate learning on small datasets. We used a naïve Bayes classifier to disambiguate medical terms and the UMLS for its symbolic knowledge. Only information from the sentence in which the ambiguous word appeared was used.

We tested different experimental conditions and compared them with the majority sense baseline. In each condition, more (or more precise) information was provided to a naïve Bayes classifier. However, it was not the condition with the most information that resulted in the best performance. Two types of information helped accuracy: information about the word being the main word or not [26] and UMLS semantic types associated with unambiguous words in the sentence. When evaluating the potential causes for the high variability between the performances of different words, we discovered an unexpected trend related to example ambiguity. Words that were troublesome to the human evaluators were generally also harder to disambiguate automatically. This was unexpected because we did not provide the algorithms with any information that was related to this difficulty (such as the unresolved counts). We performed additional tests so that we could exclude the possibility that the algorithm itself caused this variability, or that it was mainly due to a gold standard based on consensus between 11 people. Instead, it may be due to the different meanings available in the UMLS, which led to the confusion of the experts compiling the gold standard. To avoid such confusion, subsets from the UMLS appropriate to the domain may result in better results.

Acknowledgements

We thank the NLM experts who evaluated the medical terms, Halil Kilicoglu and Jim Mork for their assistance in making the data accessible, and Olivier Bodenreider for his suggestions. The first author was supported by an appointment to the National Library of Medicine Research Participation Program administered by the Oak Ridge Institute for Science and Education through an inter-agency agreement between the U.S. Department of Energy and the National Library of Medicine.

References

[1] N. Ide, J. Véronis, Word sense disambiguation: the state of the art, Comput. Linguist. 24 (1998) 1—41.

[2] H. Schütze, Automatic word sense discrimination, Comput. Linguist. 24 (1998) 97—123.

[3] T. Pedersen, R. Bruce, Distinguishing word senses in untagged text, in: Presented at Second Conference on Empirical Methods in Natural Language Processing, Providence, RI, 1997.

[4] R.J. Mooney, Comparative experiments on disambiguating word senses: an illustration of the role of bias in machine learning, in: Presented at Conference on Empirical Methods in Natural Language Processing, 1996.

[5] V. Hoste, I. Hendrickx, W. Daelemans, A. Van Den Bosch, Parameter optimization of machine-learning of word sense disambiguation, Nat. Lang. Eng. 8 (2002) 311—325.

[6] T. Pedersen, A decision tree of bigrams is an accurate predictor of word senses, in: Presented at Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics, 2001.

[7] R. Florian, S. Cucerzan, C. Schafer, D. Yarowsky, Combining classifiers for word sense disambiguation, Nat. Lang. Eng. 1 (2002) 1—14.

[8] R. Florian, S. Cucerzan, C. Schafer, D. Yarowsky, Combining classifiers for word sense disambiguation, Nat. Lang. Eng. 8 (2002) 327—341.

[9] R. Mihalcea, D.I. Moldovan, Word sense disambiguation based on semantic density, in: Presented at Coling-ACL'98 Wirjdshop on the Usage of WordNet in Natural Language Processing Systems, Montreal, Canada, 1998.

[10] A.R. Aronson, Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program, in: Presented at AMIA Symposium, 2001.

[11] G.A. Miller, R. Beckwidth, C. Fellbaum, D. Gross, K. Miller, Introduction to WordNet: An On-line Lexical Database, 1998.

[12] D.Z. Inkpen, G. Hirst, Automatic sense disambiguation of the near-synonyms in a dictionary entry, in: Presented at 4th Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2003), Mexico City, 2003.

[13] C. Santamaría, J. Gonzalo, F. Verdejo, Automatic association of Web directories with word senses, Comput. Linguist. 29 (2003) 485—502.

[14] B. Magnini, C. Strapparava, G. Pezzula, A. Gliozzo, The role of domain information in word sense disambiguation, Nat. Lang. Eng. 8 (2002) 359—373.

[15] B. Humphreys, D. Lindberg, H. Schoolman, G. Barnett, The unified medical language system: an informatics research collaboration, J. Am. Med. Inform. Assoc. 5 (1998) 1—11.

[16] H. Liu, S. Johnson, C. Friedman, Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS, J. Am. Med. Inform. Assoc. 9 (2002) 621—636.

[17] P. Ruch, R. Baud, A. Geissbühler, Using lexical disambiguation and named-entity recognition to improve spelling correction in the electronic patient record, Artif. Intell. Med. 29 (2003) 169—184.

[18] P. Ruch, J. Wagner, J. Bouillon, R. Baud, A.M. Rassinoux, J. Scherrer, MEDTAG: tag-like semantics for medical document indexing, in: Presented at AMIA Symposium, 1999.

[19] V. Hatzivassiloglou, P.A. Duboué, Disambiguating proteins, genes, and RNA in text: a machine learning approach, Bioinformatics 1 (2001) 1—10.

[20] F. Ginter, J. Boberg, J. Järvinen, T. Salakoski, New techniques for disambiguation in natural language and their application to biological text, J. Mach. Learn. Res. 5 (2004) 605—621.

[21] H. Liu, Y.A. Lussier, C. Friedman, Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method, J. Biomed. Inform. 34 (2001) 249—261.

[22] H. Han, L. Giles, H. Zha, C. Li, K. Tsioutsiouliklis, Two supervised learning approaches for name disambiguation in author citations, in: Presented at 4th ACM/IEEE-CS Joint Conference on Digital Libraries, Tucson, Arizona, 2004.

[23] M. Weeber, J. Mork, A. Aronson, Developing a test collection for biomedical word sense disambiguation, in: Presented at AMIA Symposium, 2001.

[24] A. McCray, Representing biomedical knowledge in the UMLS Semantic Network, in: N.C. Broering (Ed.), High-Performance Medical Libraries: Advances in Information Management for the Virtual Era, Meckler Publishing, Westport, CT, 1993, pp. 45—55.

[25] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java, Morgan Kaufmann, San Francisco, 2000.

[26] G. Leroy, T.C. Rindflesch, Using symbolic knowledge in the UMLS to disambiguate words in small datasets with a naïve Bayes classifier, in: Presented at MedInfo, San Francisco, 2004.