Using Lexical Knowledge in Text Classification

Sam Scott and Stan Matwin
{sscott,stan}@csi.uottawa.ca
Department of Computer Science
University of Ottawa
150 Louis Pasteur,
Ottawa, Ontario
K1N 6N5 Canada

Abstract

This paper describes several experiments in text classification using WordNet, a rich source of lexical background knowledge available in the public domain. WordNet is used to map the original words from a text into sets based on synonym and hypernym relationships. This information is used to compute a change of representation from *bag of words* to *hypernym density*. Six binary classification tasks of varying difficulty are defined, and the Ripper system is used to produce discrimination rules for each task using each representation. Experiments show that for some of the more difficult tasks the *hypernym density* representation leads to significantly more accurate and more comprehensible rules.

Key Words

text classification, background knowledge, change of representation, WordNet

Contact Author

Sam Scott
sscott@csi.uottawa.ca
(613) 565-5770

(Submitted to ICML-98, 25/02/98)
(University of Ottawa TR-98-03)

1. Introduction

Text classification is an important application of Machine Learning that can be used in a variety of contexts, including e-mail and news filtering, personal information agents and assistants, information retrieval, and automatic indexing. A number of authors have experimented with automatic text classification systems. In a typical paper, [Lang 95] investigated how text classification could be applied to filtering news to match the interests of a given user. The classification problem was therefore a two-class problem with a very large feature set. In Lang's approach, documents were represented as the set of words that occurred in the document, an approach that is often referred to as the *bag of words* model. No linguistic processing (other than a stop list of most frequent words) was applied to the original text. Two learning techniques, tf-idf weighting[1] and MDL, were applied. Similarly [Joachims 97] and [Koller, Sahami 97] also used the bag of words approach in text classification systems.

In general, the bag of words is the main, if not the only representation used in most text classification experiments and applications. This approach totally neglects the fact that text classification is actually done on written documents. All lexical, syntactic and semantic properties of the text are lost when the bag of words is applied. We believe that the performance of text classification systems could be improved by incorporating different kinds of linguistic knowledge about the text being classified. Our work sets out to test this hypothesis in a simple learning setting.

The general approach explored in this paper has become more feasible in recent years, as the NLP community has produced a number of public domain tools and repositories representing vast amounts of linguistic knowledge. WordNet is one such tool. In this paper, we change the representation of the text to be classified by exploiting some of the linguistic features of the words based on the mappings provided by the WordNet system. We report results from an initial study in which the new representation is compared to the old bag of words model.

The paper proceeds as follows. In section 2 we present the data sets that we work with, the classification tasks defined on this data, and some initial experiments with the Ripper learning system. Section 3 describes the aspects of WordNet on

---

[1] The bag of words was extended here with the frequencies of the words.

which we base our change of representation, and section 4 discusses the change of representation process. Section 5 presents experimental results after the change of representation and discusses accuracy and comprehensibility. Section 6 provides a brief review of related work, and section 7 presents the conclusion and future work.

2. Preliminaries: the Corpora, Classification Tasks, and Learning Algorithm

The classification tasks used in this study are drawn from three different corpora: Reuters-21578, USENET, and the Digital Tradition (DigiTrad). Both Reuters and USENET have been the subject of previous studies in machine learning (see [Koller, Sahami 97] for a study of Reuters and [Weiss et. al. 96] for a study of USENET). In keeping with previous studies, we used topic headings as the basis for the Reuters classification tasks and newsgroup names as the basis for the USENET tasks. The third corpus, DigiTrad is a public domain collection of 6500 folk song lyrics [Greenhaus 96]. To aid searching, the owners of DigiTrad have assigned to each song one or more key words from a fixed list. Some of these key words capture taxonomic information (e.g. "Irish" or "British") while others relate to subject matter (e.g. "murder" or "marriage"). The latter type of key words served as the basis for the classification tasks in this study.

Not all types of text are equally difficult to classify. Reuters consists of articles written purely as a source of factual information. The writing style tends to be direct and to the point, and uses a restricted vocabulary to aid quick comprehension. It has been observed that the topic headings in Reuters tend to consist of words that appear frequently in the text, and this observation has been exploited to help improve classification accuracy [Rodríguez et. al. 97]. DigiTrad and USENET are good examples of the opposite extreme. The texts in DigiTrad make heavy use of metaphoric, rhyming, unusual and archaic language. Often the lyrics do not explicitly state what a song is about. Contributors to USENET often vary in their writing styles, stray from the topic, or use unusual or ungrammatical language. All of these qualities tend to make subject-based classification tasks from DigiTrad and USENET more difficult than those of a comparable size from Reuters.

From the three corpora described above, six binary classification tasks were defined, as shown in table 1. The tasks were chosen to be roughly the same size, and cover cases in which the classes seemed to be semantically related

(REUTER2 and USENET2) as well as those in which the classes seemed unrelated (REUTER1 and USENET1). In all cases the classes were made completely disjoint by removing any overlapping examples.[2]

The machine learning algorithm chosen for this study was Ripper, a rule-based learner developed by William Cohen [Cohen 95]. Ripper was specifically designed to handle the high dimensionality of bag of words text classification by being fast and using set-valued features [Cohen 96]. Table 1 shows that our intuitions about the difficulty of the three corpora for bag of words classification are valid in the case of the Ripper algorithm. Error rates for the Reuters tasks were under 5%, while error rates for the other tasks ranged from approximately 19% to 38%. We believe that with the growing applications of text classification on the Internet, it is likely that the kinds of texts to be automatically classified will share many features with the kinds of texts that are difficult for the bag of words approach.

Task Name	Source	Classes	Size	Balance	Words	Error
REUTER1	Reuters-21578	*livestock / gold*	224	98/126	154	1.75
REUTER2	Reuters-21578	*corn / wheat*	313	130/183	173	3.87
SONG1	DigiTrad	*murder / marriage*	424	200/224	331	30.23
SONG2	DigiTrad	*political / religion*	432	194/238	241	32.64
USENET1	USENET	*soc.history / misc.taxes.moderated*	249	79/170	166	19.92
USENET2	USENET	*bionet.microbiology / bionet.neuroscience*	280	117/163	152	37.86

Table 1: Information on the classification tasks discussed in this paper. "Size" refers to total number of examples in each task. "Balance" shows number of examples in each class. "Words" shows the average length of the documents in each task. "Error" shows the average percentage error rates for each task using Ripper with bag of words *and 10-fold cross-validation.*

It is worth noting that difficult tasks for Ripper are not necessarily difficult for human classifiers. 200 examples from each of the SONG1 and SONG2 tasks were classified by hand (with no special training phase) and the results were compared to the classes indicated by the DigiTrad labels. The error rates were approximately 1% for SONG1 and 4% for SONG2. Clearly the background knowledge a human brings to a classification task can be used to overcome the

[2] USENET articles that were cross-posted or tagged as follow-ups were excluded so that the remaining articles reflected a wide variety of attempts to launch discussions within the given topics. Non-text objects such as uuencoded bitmaps were also removed from the postings.

difficulties posed by the text itself.

3. WordNet

WordNet is a public domain on-line lexical reference system that captures a number of important conceptual relations in addition to the synonymy and antonymy found in most thesauri [Miller 95].[3] These relations can be used to engineer a change of representation in text data by transforming vectors of words into vectors of word meanings. The synonymy relation can be used to map words with similar meanings together, and hypernymy (corresponding to the "is a" relation) can be used to generalize noun and verb meanings to a higher level of abstraction.[4]

The nodes in WordNet's semantic network are referred to as "synsets". Each synset represents a particular semantic meaning shared by a group of words and phrases. Each word or phrase in the synset is a synonym of the others with respect to this particular meaning, and a word or phrase with more than one meaning will appear in more than one synset. The synset nodes therefore embody the concept of synonymy, while other semantic relations are modeled as arcs between the nodes (see figure 1).

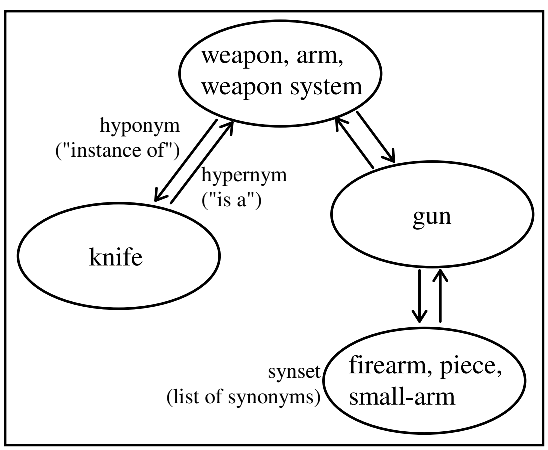

Figure 1: a piece of the WordNet semantic hierarchy (idealized model)

[3] Version 1.5 of WordNet was used in this study.
[4] The hypernymy relation is defined only for nouns and verbs.

Synsets are also organized by part of speech (e.g. nouns and verbs). Figure 2 shows the output of WordNet when a user requests all the hypernyms for the noun senses of the word *gun*. WordNet contains 3 noun senses for *gun*, and each corresponding synset is displayed, followed by all the synsets that appear above it in the hypernym hierarchy.

```
Synonyms/Hypernyms (Ordered by Frequency) of noun gun
3 senses of gun
Sense 1
gun
          ⇒ weapon, arm, weapon system
                    ⇒ weaponry, arms, implements of war, weapons system, munition
                              ⇒ instrumentality, instrumentation
                                        ⇒ artifact, artefact
Sense 2                                           ⇒ object, inanimate object, physical object
artillery, heavy weapon, gun, ordnance                      ⇒ entity
          ⇒ armament
                    ⇒ weaponry, arms, implements of war, weapons system, munition
                              ⇒ instrumentality, instrumentation
Sense 3                                 ⇒ artifact, artefact
gunman, gun                                       ⇒ object, inanimate object, physical object
          ⇒ shot, shooter                                   ⇒ entity
                    ⇒ expert
                              ⇒ person, individual, someone, somebody, mortal, human
                                        ⇒ life form, organism, being, living thing
                                                  ⇒ entity
                                        ⇒ causal agent, cause, causal agency
                                                  ⇒ entity
```

Figure 2: Some output typical of WordNet 1.5. The first line for each "sense" is a synset containing the word "gun". Each line preceded by "⇒" indicates a step upwards through the hypernym hierarchy of generalization.

In order to take advantage of the part of speech organization of the synsets, an automatic part of speech tagger can be used. For this study, the public domain supervised tagger by Eric Brill was chosen [Brill 92].

4. Using WordNet to Accomplish a Change of Representation

The proposed algorithm for computing the change of representation requires three passes through the corpus.

a) During the first pass, the Brill tagger assigns a part of speech tag to each word in the corpus.

b) During the second pass, all nouns and verbs are looked up in WordNet and a global list of all synonym and hypernym synsets is assembled[5]. Infrequently occurring synsets are discarded, and those that remain form the feature set.

c) During the third pass, the density of each synset (defined as the number of occurrences of a synset divided by the number of words in the document) is computed for each example resulting in a set of numerical feature vectors. Every unique synset in the WordNet output counts as one occurrence of that synset (for example in figure 2 there are 15 unique synsets).

(In step *b* a synset is defined as infrequent if its frequency of occurrence over the entire corpus is less than 0.05N, where N is the number of documents in the corpus. This cut off point was found to more than halve the number of features without significantly affecting classification accuracy. An upper cutoff point was also considered, but the reduction in dimensionality was judged too small to justify the risk of excluding possibly relevant features.)

In the new representation, each feature represents a set of either nouns or verbs. These sets overlap. More general features, corresponding to synsets higher up in the hypernym hierarchy, represent supersets of the less general features. Because most words have multiple meanings in WordNet, features at the same level of generalization also overlap but do not form subsets of one another. Ideally, a word sense disambiguation algorithm could be used to resolve some of this latter overlap by mapping each word to its correct meaning.

In the only other published application of WordNet to classification, word sense disambiguation was performed by manual inspection [Rodríguez et. al. 97]. This approach was feasible in the context of that study because of the small number of words involved. In the current study, the words number in the tens of thousands. Manual disambiguation is not an option. Automatic disambiguation is possible using WordNet, but the procedure is complex and may require the use of a parser [Li et. al. 95]. Instead all senses returned by WordNet are judged equally likely to

---

[5] Note that WordNet also maps words to their roots. For instance, the word "guns" would map to the word "gun" and the word "am" maps to the word "be".

be correct, and all of them are included in the feature set. The use of the density measurement is an attempt to capture some measure of relevancy. The learner is aided by the fact that many different but synonymous words will map to common synsets, thus raising the densities of the "correct" synsets. Similarly, different but related synsets are connected by common hypernyms at a higher level of generalization. In other words, a relatively low value for a feature indicates that little evidence was found for the meaningfulness of that synset to the document.

It may seem that treating all hypernyms as equals would lead to the inclusion of excessively general hypernyms, such as the synset *entity* in figure 2. This is not a problem, since the rule-based learner will choose features from the level of generalization most suitable to the task: if a feature is too general to be useful, it will not be included in any classification rules. Furthermore, in some instances these seemingly overly general features do prove to be useful (see section 5.2), so excluding them may actually hurt classification accuracy

The fact that each word potentially maps to dozens of features in the new representation could lead to an explosion of dimensionality far worse than the bag of words approach. Fortunately, the overlapping nature of the synsets along with the part of speech and frequency filtering described above counteracts this tendency and reduces the total number of features in the new representation significantly. A comparison of the dimensionality of the two representations is shown in table 2.

Task	**Bag of Words**	**Hypernym Density**	**Reduction (%)**
REUTER1	4227	2573	39.1
REUTER2	4738	2591	45.3
SONG1	9902	3844	61.2
SONG2	10550	3347	68.3
USENET1	6295	2627	58.2
USENET2	6852	2451	64.2
Average:	7094	2906	56.1

Table 2: Comparison of the number of features in each representation.

Clearly the change of representation process leaves a lot of room for inaccuracies to be introduced to the feature set. Some sources of potential error are: a) the tagger, b) words missing from WordNet, c) the shallowness of WordNet's semantic hierarchy in some knowledge domains, and d) the lack of true word sense disambiguation.

5. Experiments and results

5.1. Accuracy

The new hypernym density representation differs in three important ways from the bag of words: a) all words are discarded except nouns and verbs, b) filtered normalized density vectors replace binary vectors, and c) synonym and hypernym synsets replace words. To show convincingly that improvements in accuracy are derived solely from the use of hypernyms, three normalizing experiments were performed using the following representations:

a) bag of words using only nouns and verbs,
b) filtered normalized density vectors for nouns and verbs, and
c) filtered normalized density vectors for noun and verb synonym synsets (no hypernyms).

The results of these three runs were compared to the bag of words approach (see table 3) and in no case was any statistically significant difference found, leading to the conclusion that any improvements in accuracy derive mainly from the use of hypernyms.

Task	Bag of Words	Bag of Nouns and Verbs	Noun and Verb Dens.	Synonym Synset Dens.
REUTER1	1.75	1.75	1.75	1.75
REUTER2	3.87	4.19	5.48	4.84
SONG1	30.23	27.35	27.67	27.65
SONG2	32.64	28.23	29.56	34.68
USENET1	19.92	18.56	20.45	22.54
USENET2	37.86	37.86	35.00	38.21

Table 3: Comparison of error rates over 10-fold cross-validation for the normalizing experiments. No statistically significant benefit or harm is derived from any of these changes of representation.

For the main experiment, average error rates over 10-fold cross-validation were compared for all six classification tasks using three representations: bag of words, hypernym density, and a combined representation in which both the hypernyms and the original words were used . For each classification task the same 10 folds were used on every run so the results could be tested for significance using a paired-t test. The results are summarized in table 4.

Task	**Bag of Words**	**Hypernym Density**	**Bag of Words + Hypernym Density**
REUTER1	1.75	2.38	2.20
REUTER2	3.87	*6.13*	*7.39*
SONG1	30.23	22.04	20.59
SONG2	32.64	34.45	*37.97*
USENET1	19.92	14.36[6]	13.11
USENET2	37.86	40.00	40.71

Table 4: Comparison of error rates over 10-fold cross-validation for the six data sets in the study. Statistically significant improvements over bag of words are underlined. Significantly worse error rates are in italics.

In the case of the Reuters tasks, no improvements over bag of words were expected and none were observed. On the other hand, a dramatic reduction in the error rate was seen for SONG1 and USENET1. In both cases, the average number of errors was reduced by approximately a third. For the SONG2 and USENET2 data sets, the use of hypernyms produced a comparable error rate to bag of words, except in the case of combining features for SONG2, which lead to a slight performance degradation The discussion of these results is left to section 5.4.

5.2 Comprehensibility

Figure 3 shows a comparison of the rules learned by Ripper on the USENET1 data set. The results for both bag of words and hypernym density are shown for the same fold of data.

[6] Significant only to 94.5% confidence rather than the usual 95%

possession $\leq 2.9 \Rightarrow$ soc.history
default $\Rightarrow$ misc.taxes.moderated

Rule learned using hypernym frequency

for a document D,
("tax" $\notin$ D & "history" $\in$ D) OR
("tax" $\notin$ D & "s" $\in$ D & "any" $\in$ D) OR
("tax" $\notin$ D & "is" $\notin$ D & "and" $\notin$ D & "if" $\notin$ D & "roth" $\notin$ D) OR
("century" $\in$ D) OR
("great" $\in$ D) OR
("survey" $\in$ D) OR
("war" $\in$ D) $\Rightarrow$ soc.history
default $\Rightarrow$ misc.taxes.moderated

Rule learned using bag of words

Figure 3: A comparison of the rules learned by Ripper using hypernym frequency (top) and bag of words (bottom) on a single fold of the USENET1 data. The bottom rule produced twice as many errors on the holdout data.

In the case of hypernyms, Ripper has learned a simple rule saying that if the synset *possession* has a low density, the document probably belongs in the history category. Over the 10 folds of the data, seven folds produced a rule almost identical to the one shown. For the remaining three folds, the *possession* hypernym appeared along with other synsets in slightly different rules. The hyponyms of *possession* include words such as *ownership*, *asset*, and *liability* - the sorts of words used often during discussions about taxes, but rarely during discussions about history. On the other hand, the rules learned on the bag of words data seem less comprehensible: they are more elaborate and less semantically clear. Furthermore, the rules tended to vary widely across the 10 folds, suggesting that they were less robust and more dependent on the specifics of the training data.

A similar situation is depicted in figure 4 for the SONG1 data set. In this case it is interesting to note the use of the hypernym *weapon_arm_weapon_system* which appeared in the rules across all 10 folds. The use of this hypernym is significant because its hyponyms, such as *knife* and *gun* were never used in the bag of words approach. Another interesting feature of the hypernym rule is the presence of *abstraction*, a highly general hypernym that has hundreds of hyponyms beneath it in the hierarchy.

(marry $\geq$ 0.8 & termination $\leq$ 0) OR
(kill $\leq$ 0.45 & weapon_arm_weapon_system $\leq$ 1.09) OR
(abstraction $\leq$ 12.15) $\Rightarrow$ marriage
default $\Rightarrow$ murder

Rule learned using hypernym frequency

for a document D,
("married" $\in$ D) OR
("dt" $\notin$ D & "was" $\notin$ D & "child" $\notin$ D & "came" $\notin$ D) OR
("wife" $\in$ D & "murder" $\notin$ D & "no" $\in$ D & "killed" $\notin$ D) OR
("wed" $\in$ D) OR
("tailor" $\in$ D) OR $\Rightarrow$ marriage
default $\Rightarrow$ murder

Rule learned using bag of words

Figure 4: A comparison of the rules learned by Ripper using hypernym frequency (top) and bag of words (bottom) on a single fold of the SONG1 data. Some hypernym names have been shortened to save space. The bottom rule produced 60% more errors on the holdout data.

A systematic investigation of the comprehensibility of rules produced from the changed representation is beyond the scope of this work. However, we often see evidence of the better comprehensibility of the rules produced from the hypernym density representation. If one is to see the approach presented here as an integrated system, which takes texts as input, automatically engineers the new features using WordNet, and then learns classification rules, then the results shown above represent an example of constructive learning [Michalski 86].

5.3 Discussion

Clearly the change of representation to hypernym density greatly improves classification accuracy in some cases. Why does it sometimes *not* work?

In the case of the Reuters tasks, the lack of benefit is not a particular worry. It is very unlikely that any change of representation could have improved on the accuracy of bag of words for these tasks. But why was performance actually degraded in the case of REUTER2? Visual inspection of the rules learned using bag of words versus hypernym density reveals that they are very similar. With bag of words, Ripper used the words *corn*, *maize*, and *wheat* to do most classification, while with hypernym density the synsets *corn_maize_IndianCorn_ZeaMays* and *wheat* were used extensively. Clearly the learner was attempting to use the same information in both cases, but with slightly different results probably due to small errors introduced during the change of representation. (One source of error was definitely the tagger, which tagged the important word *maize* incorrectly 25% of the time.)

In the case of the SONG2 task, the main problem seems to be that the classes (*political* and *religion*) are more closely semantically related than their class labels suggest. Visual inspection of these songs revealed that many of the political songs contain statements about religion, make references to religious concepts, or frame their messages in religious terminology. This was the source of the higher error rate reported in section 2 when these songs were classified by hand. Inspection of Ripper's output revealed that bag of words rules make heavy use of religious words such as *Jesus, lord,* and *soul*, while the hypernym rules mostly contained highly abstract political synsets such as *social group* and *political unit*. It is possible that subtle differences in religious terminology (for instance between gospel hymns and political parodies of religion) were mapped to common synsets in WordNet.

Finally, in the case of USENET2, the problem is two-fold. The classes are semantically closely related (*microbiology* and *neuroscience*) and the writers tend to use highly technical terms that are not found in WordNet 1.5. Some examples of missing words include *neuroscientist*, *haemocytometer*, *HIV*, *kinase*, *neurobiology*, and *retrovirus*[7]. An attempt was made to add the missing words manually into the WordNet hierarchy, but even then the extended semantic hierarchy was not fine-grained enough to represent the level of expertise necessary to allow meaningful generalizations.

6. Related Work

In general, the idea of using domain knowledge in an inductive system that learns classification rules from example is not new (e.g. see [Bergadano, Giordana 88], [Clark, Matwin 93]). However, little work has been done using available NLP resources for text classification. Recent work by [Rodríguez et. al. 97] is an exception. In their study, WordNet was used in conjunction with a neural network to classify texts from the Reuters corpus. Only the synonymy relation was used, no syntactic information was included, and the words were looked up by hand in

[7] Some of these terms do appear in WordNet 1.6

WordNet. Furthermore, the approach took advantage of the fact noted in section 2, that Reuters topic headings are themselves good indicators for classification.

7. Conclusions and future work.

This paper describes a change of representation for text data that can lead to significant reductions in error rates on certain types of text classification tasks. The new representation uses the lexical and semantic knowledge embodied in WordNet to move from a *bag of words* representation to a representation based on *hypernym density*. A side benefit of this process is that the classification rules induced using hypernym density are often simpler and more comprehensible than rules induced using the bag of words.

Our experience indicates that the hypernym density representation will work well for texts that vary in writing style, use an extended or unusual vocabulary, or are written by multiple authors employing different styles. It is not likely to work well for text that is guaranteed to be written concisely and efficiently, such as the text in Reuters-21578. In particular, hypernym density is more likely to perform well on classification tasks involving narrowly defined and/or semantically distant classes (such as SONG1 and USENET1). In the case of classes which are broadly defined and/or semantically related (such as SONG2 and USENET2) hypernym density does not always outperform bag of words.

Abundant future work stems from this preliminary study. The first task is to experiment with different levels of generalization. The representation used in this study included all hypernyms as features, but recent experiments have suggested that in some cases even better results can be obtained by restricting the number of steps that are taken up the hypernym tree. In particular for the SONG1 task (but not for any others) the error rate dropped from 22.04% using all hypernyms to 17.04% using only the first step in the hierarchy. It is possible that similar improvements will be seen at other levels of generalization. Another area for development is the incorporation of more of the semantic relations from WordNet, such as meronymy ("part of"). This will give the change of representation an even more semantic character. More sophisticated word sense disambiguation (but still short of full syntax analysis) could produce more accurate hypernym density features, and the use of other linguistic resources available in the public domain, such as the Unified Medical Language System® Metathesaurus® [NLM 98], could improve classifier performance in knowledge domains which are semantically close and highly expert.

Acknowledgments

The authors are grateful to William Cohen for making the Ripper system available for this research. The first author acknowledges the support of the National Sciences and Engineering Research Council of Canada and the Interactive Information Group of the National Research Council. The second author acknowledges the support of the Natural Sciences and Engineering Research Council of Canada.

References

[Bergadano, Giordana 88] F. Bergadano and A. Giordana. A knowledge-intensive Approach to Concept Induction. In *Proc. ICML-88*, 305-317, 1988.

[Brill 92] Eric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, ACL, 1992.

[Clark, Matwin 93] P. Clark and S. Matwin. Learning Domain Theories Using Abstract Background Knowledge. *In Proc. ECML-93*, 360-365, 1993.

[Cohen 95] William W. Cohen. Fast Effective Rule Induction. In *Proc. ICML-95*. Lake Tahoe, California, 1995.

[Cohen 96] William W. Cohen. Learning Trees and Rules with Set-valued Features. In *Proc. AAAI-96*, 1996

[Greenhaus 96] Dick Greenhaus. *About the Digital Tradition*. (www.mudcat.org/DigiTrad-blurb.html), 1996.[8]

[Joachims 97] T. Joachims. A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In *Proc. ICML-97*, 143-146, 1997.

[Koller, Sahami 96] D. Koller and M. Sahami. Hierarchically Classifying Documents Using Very Few Words. In *Proc. ICML-97*, 170-176, 1997.

[Lang 95] K. Lang. NewsWeeder: Learning to Filter News. In *Proc. ICML-95*, 331-336, 1995.

[8] The DigiTrad database itself is at www.deltablues.com/folksearch.html

[Li et. al. 95] Xiaobin Li, Stan Szpakowicz and Stan Matwin. A WordNet-based Algorithm for Word Sense Disambiguation. In *Proc. IJCAI-95*, Montréal, Canada, 1995.

[Michalski 86] R. Michalski. Understanding the nature of Learning: Issues and Research Directions. In R. Michalski, J. Carbonell and T. Mitchell, eds., *Machine Learning, vol. II*, Morgan Kaufmann 1986..

[Miller 95] G. Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 1995.

[NLM 98] National Library of Medicine. *Unified Medical Language System Overview*. www.nlm.nih.gov/research/umls/UMLSDOC.HTML, February, 1998.

[Rodríguez et. al. 97] Manuel de Buenaga Rodríguez, José María Gómez-Hidalgo and Belén Díaz-Agudo. Using WordNet to Complement Training Information in Text Categorization. In *Proc. RANLP-97.* Stanford, March 25-27, 1997

[Weiss et. al. 96] Scott A. Weiss, Simon Kasif, and Eric Brill. Text Classification in USENET Newsgroups: A Progress Report. In *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access*, Bulgaria, September 11-13, 1996.