Text Classification Using WordNet Hypernyms

Sam Scott
Computer Science Dept.
University of Ottawa
Ottawa, ON K1N 6N5 (Canada)
`sscott@csi.uottawa.ca`

Stan Matwin
Computer Science Dept.
University of Ottawa
Ottawa, ON K1N 6N5 (Canada)
`stan@csi.uottawa.ca`

Abstract

This paper describes experiments in Machine Learning for text classification using a new representation of text based on WordNet hypernyms. Six binary classification tasks of varying difficulty are defined, and the Ripper system is used to produce discrimination rules for each task using the new *hypernym density* representation. Rules are also produced with the commonly used *bag-of-words* representation, incorporating no knowledge from WordNet. Experiments show that for some of the more difficult tasks the *hypernym density* representation leads to significantly more accurate and more comprehensible rules.

1. Introduction

The task of *Supervised Machine Learning* can be stated as follows: given a set of classification labels C, and set of training examples E, each of which has been assigned one of the class labels from C, the system must use E to form a hypothesis that can be used to predict the class labels of previously unseen examples of the same type [Mitchell 97]. In machine learning systems that classify text, E is a set of labeled documents from a corpus such as Reuters-21578. The labels can signify topic headings, writing styles, or judgements as to the documents' relevance. Text classification systems are used in a variety of contexts, including e-mail and news filtering, personal information agents and assistants, information retrieval, and automatic indexing.

Before a set of documents can be presented to a machine learning system, each document must be transformed into a feature vector. Typically, each element of a feature vector represents a word from the corpus. The feature values may be binary, indicating presence or absence of the word in the document, or they may be integers or real numbers indicating some measure of frequency of the word's appearance in the text. This text representation, referred to as the *bag-of-words*, is used in most typical approaches to text classification (for recent work see [Lang 95], [Joachims 97], and [Koller & Sahami 97]). In these approaches, no linguistic processing (other than a stop list of most frequent words) is applied to the original text.

This paper explores the hypothesis that incorporating linguistic knowledge into text representation can lead to improvements in classification accuracy. Specifically, we use part of speech information from the Brill tagger [Brill 92] and the synonymy and hypernymy relations from WordNet [Miller 90] to change the representation of the text from bag-of-words to *hypernym density*. We report results from an ongoing study in which the hypernym density representation at different *heights of generalization* is compared to the old bag-of-words model. We focus on using the new representation of text with a particular machine learning algorithm (*Ripper*) that was designed with the high dimensionality of text classification tasks in mind. The issue of whether our results will generalize to other machine learning systems is left as future work.

The only published study comparable to this one is [Rodríguez et al. 97]. Their study used WordNet to enhance neural network learning algorithms for significant improvements in classification accuracy on the Reuters-21578 corpus. However, their approach only made use of synonymy and involved a manual word sense disambiguation step, whereas our approach uses synonymy and hypernymy and is completely automatic. Furthermore, their approach took advantage of the fact that the Reuters topic headings are themselves good indicators for classification, whereas our approach makes no such assumptions. Finally their approach to using WordNet focussed on improving the specific algorithms used by neural networks while retaining the bag-of-words representation of text. Our approach looks at using WordNet to change the representation of the text itself and thus may be applicable to a wider variety of machine learning systems.

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 discusses the new hypernym density representation. Section 4 presents experimental results using both bag-of-words and hypernym density and discusses the accuracy and comprehensibility of the rules learned by Ripper. Finally, section 5 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 information on the origin or style of the songs (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 use of terminology, stray from the topic, or use unusual 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.[1]

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 over *10-fold cross-validation*[2] 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.

It is worth noting that difficult classification tasks for Ripper are not necessarily difficult for humans. We classified 200 examples from each of the SONG1 and SONG2 by hand (with no special training phase) and compared our classifications to those from DigiTrad.

---

[1] 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.

[2] In *n-fold cross-validation* the articles in the corpus are split into n partitions. Then the learning algorithm is executed n times. On the k^{th} run, partition k is used as a *testing set* and all the other partitions make up the *training set*. The mean error-rate (percentage of the *testing set* wrongly classified) on the n runs is taken as an approximate measure of the real error-rate of the system on the given corpus.

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 texts in each task. "Balance" shows number of examples in each class. "Words" shows the average length of the documents in each task. "Error" show the average percentage error rates for each task using Ripper with bag-of-words *and 10-fold cross-validation.*

The error rates were approximately 1% for SONG1 and 4% for SONG2. Clearly the background knowledge and linguistic competence humans bring to a classification task enables us to overcome the difficulties posed by the text itself.

3. The *Hypernym Density* Representation

The algorithm for computing hypernym density requires three passes through the corpus.

a) During the first pass, the Brill tagger [Brill 92] 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. Infrequently occurring synsets are discarded, and those that remain form the feature set. (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.)

c) During the third pass, the density of each synset (defined as the number of occurrences of a synset in the WordNet output divided by the number of words in the document) is computed for each example resulting in a set of numerical feature vectors.

The calculations of frequency and density are influenced by the value of a parameter h that controls the *height of generalization*. This parameter can be used to limit the number of steps upward through the hypernym hierarchy for each word. At height $h=0$ only the synsets that contain the words in the corpus will be counted. At height $h>0$ the same synsets will be counted as well as all the hypernym synsets that appear up to h steps above them in the hypernym hierarchy. A special value of $h=max$ is defined as the level in which *all* hypernym synsets are counted, no matter how far up in the hierarchy they appear.

In the new representation, each feature represents a *set* of either nouns or verbs. At $h=max$, features corresponding to synsets higher up in the hypernym hierarchy represent supersets of the nouns or verbs represented by the less general features. At lower values of h, the nouns and verbs represented by a feature (synset) will be those that map to synsets up to h steps below it in the hypernym hierarchy. The best value of h for a given text classification task will depend on characteristics of the text such as use of terminology, similarity of topics, and breadth of topics. It will also depend on the characteristics of WordNet itself. In general, if the value for h is too small, the learner will be unable to generalize effectively. If the value for h is too large, the learner will suffer from overgeneralization because of the overlap between the features.

Note that no attempt is made at word sense disambiguation during the computation of hypernym density. Instead all senses returned by WordNet are judged equally likely to 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 or hyponymous words will map to common synsets, thus raising the densities of the "more relevant" synsets. 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.

(In the [Rodríguez et al. 97] text classification paper, word sense disambiguation was performed by manual inspection. 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, making manual disambiguation unfeasible. Automatic disambiguation is possible and often obtains good results as in [Yarowski 95] or [Li et al. 95], but this is left as future work.)

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) the lack of true word sense disambiguation, c) words missing from WordNet, and d) the shallowness of WordNet's semantic hierarchy in some knowledge domains.

4. Experiments and results

4.1. 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) hypernym synsets replace words. To show convincingly that improvements in accuracy are derived solely from the use of synsets, two normalizing experiments were performed using the following representations:

a) bag-of-words using only nouns and verbs, and
b) filtered normalized density vectors for nouns and verbs.

The results of these runs were compared to the bag-of-words approach using 10-fold cross-validation (see table 2) 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.

For the main experiments, average error rates over 10-fold cross-validation were compared for all six classification tasks using hypernym density representations with values of h ranging from 0 to 9 and $h=max$. For each classification task the same 10 partitions were used on every run so the results could be tested for significance using a paired-t test. Table 3 shows a comparison of three error rates: bag-of-words, hypernym density with $h=max$, and finally hypernym density using the best value for h.

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 (47% drop in number of errors for $h=9$) and USENET1 (34% drop for $h=9$). For the SONG2 and USENET2 data sets, the use of hypernyms produced error rates comparable to bag-of-words. The discussion of these results is left to section 4.3.

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

Table 2: Comparison of percentage 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.

Task	**Bag of Words**	**Hypernym Density**			
		h	**error**	***h***	**error**
REUTER1	1.75	max	2.38	0	1.75
REUTER2	3.87	max	6.13	0	4.84
SONG1	30.23	max	*22.04*	9	*16.00*
SONG2	32.64	max	34.45	4	31.04
USENET1	19.92	max	*14.36*	9	*13.11*
USENET2	37.86	max	40.00	2	36.43

Table 3: Comparison of percentage error rates over 10-fold cross-validation for the six data sets in the study. Statistically significant improvements over bag-of-words are shown in Italics.

4.2 Comprehensibility

In the machine learning community, increasing weight is given to the idea that classification hypotheses should be comprehensible to the user. Rule induction systems like Ripper are known for producing more comprehensible output than, say multi-layer perceptrons. A systematic investigation of the comprehensibility of rules produced using hypernym density versus bag-of-words 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. Figure 1 shows a comparison of the rules learned by Ripper on the USENET1 data set. The results for both bag-of-words and $h=max$ hypernym density are shown for the same fold of 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

possession(synset) $\leq$ 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

Rule learned using bag of words

Figure 1: A comparison of the rules learned by Ripper using hypernym density with h=max *(top) and bag of words (bottom) on a single fold of the USENET1 data. The bottom rule produced twice as many errors on the testing set.*

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 often used 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.

4.3 Discussion

Hypernym density has been observed to greatly improve classification accuracy in some cases, while in others the improvements are not particularly spectacular. In the case of the Reuters tasks, the lack of improvement 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 the cases of the SONG2 and USENET2 tasks are worth looking at in more detail.

In 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 density rules at *h=max* mostly contained highly abstract political synsets such as *social group* and *political unit*. It is possible that *overgeneralization* occurred when subtle differences in religious terminology (for instance between gospel hymns and political parodies of religion) were mapped to common synsets in WordNet.

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*[3]. 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 allow meaningful generalizations. Because of the shallowness of the hierarchy, *overgeneralization* quickly becomes a problem as the height of generalization increases. This is why the best error rate for USENET2 using hypernym density was found at *h=2*.

Clearly the change of representation to hypernym density works best only with an appropriate value for the parameter *h*. We have introduced a new parameter into the learning task that must somehow be set by the user. This is certainly not unheard of in the machine learning community. All currently available machine learning systems contain a large number of parameters. The only difference is that *h* modifies the *feature set* rather than the learning algorithm itself. Nevertheless, it is worth addressing the question of how this parameter could be set in practice.

[Kohavi & John 95] describe a “wrapper” method for learning algorithms that automatically selects appropriate parameters. In their system, the set of parameters is treated as a vector space that can be searched for an optimal setting. The sets of parameters are evaluated using 10-fold cross-validation on the training data, and a best-first search strategy is employed to search for the parameter set that minimizes the average error rate. This system

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

could easily be adapted to include a parameter such as *h* that modifies the feature set. Indeed [Kohavi & John 97] have already extended their method to the related problem of finding optimal feature subsets for learning.

5. Conclusions and future work.

This paper describes a method of incorporating WordNet knowledge into text representation that can lead to significant reductions in error rates on certain types of text classification tasks. The method uses the lexical and semantic knowledge embodied in WordNet to move from a *bag-of-words* representation to a representation based on *hypernym density*. The appropriate value for the *height of generalization* parameter *h* depends on the characteristics of each classification task. A side benefit of the hypernym density representation is that the classification rules induced are often simpler and more comprehensible than rules induced using the bag-of-words.

Our experience indicates that the hypernym density representation can work well for texts that use an extended or unusual vocabulary, or are written by multiple authors employing different terminologies. 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 that are broadly defined and/or semantically related (such as SONG2 and USENET2) hypernym density does not always outperform bag-of-words.

One area for future work is the incorporation of more of the relations from WordNet, such as meronymy. This will give the change of representation an even more semantic character. More sophisticated word sense disambiguation could produce more accurate hypernym density features. 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 that are semantically close and highly expert. Finally, there is the task of testing whether the improvements noted in this study generalize to machine learning systems other than Ripper.

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

[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.

[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.[4]

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

[Kohavi & John 95] Ron Kohavi and George H. John. Automatic Parameter Selection by Minimizing Estimated Error. In *Proc. ICML-95*, 1995.

[Kohavi & John 97] Ron Kohavi and George H. John. Wrapers for Feature Subset Selection. In *Artificial Intelligence Journal*, special issue on relevance, May 20, 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.

[Li et al. 95] Xiaobin Li, Stan Szpakowicz and Stan

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

Matwin. A WordNet-based Algorithm for Word Sense Disambiguation. In *Proc. IJCAI-95*, Montréal, Canada, 1995.

[Mitchell 97] Tom Mitchell, *Machine Learning*, McGraw Hill, 1997.

[Miller 90] George A. Miller. WordNet: an On-line Lexical Database. *International Journal of Lexicography* 3(4), 1990.

[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.

[Yarowski 95] David Yarowski. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *Proceedings of the 33rd Meeting of the ACL*, Cambridge, June 26-30, 1995.