

Exploiting Thesaurus Knowledge in Rule Induction for Text Classification

Markus Junker and Andreas Abecker

German Research Center for Artificial Intelligence (DFKI GmbH)

P.O. Box 2080, D-67608 Kaiserslautern, Germany

e-mail: {junker,aabecker}@dfki.uni-kl.de

Abstract

Systems for learning text classifiers recently gained considerable interest. One technique to implement such systems is rule induction. While most other approaches rely on a relatively simple document representation and do not make use of any background knowledge, rule induction algorithms offer a good potential for improvements in both of these areas. In this paper, we show how an operator-based view of rule induction enables the easy integration of a thesaurus as background knowledge. Results with an algorithm extended by thesaurus knowledge are presented and interpreted. The interpretation shows the strengths and weaknesses of using thesaurus knowledge and gives hints for future research.

1 Introduction

Text classification deals with the task of assigning a label out of a set of predefined classes to a given text document. Example applications include classifying technical reports according to their subject research area for archiving, or analyzing incoming newswire articles wrt. their subject topic so that they can be correctly routed in a press agency. Learning approaches to text classification automatically construct classifiers for this task based on pre-classified example documents.

In today's learning approaches a document is typically represented by the set of all words it contains. This set is possibly pre-processed by a stop word list and word stemming. Various learning algorithms have been used to learn text classifiers based on such representations. One branch of approaches relies on statistical methods or ideas similar in spirit, like Bayesian classifiers, neural networks etc. (e.g. (Lewis 92; Wiener *et al.* 95; Yang 95; Sahami *et al.* 96)). Newer research in the area of text classification suggests that rule induction approaches also have the potential to achieve good results (Apté *et al.* 94; Cohen 96). Compared to other approaches, rule induction can easily be extended to more complex document representations (e.g. for the integration of word patterns). Another advantage of

rule induction is the possibility of a sound integration of background knowledge. While research on more complex document representations is underway (Finch 95; Cohen 96; Riloff 96), the proper use of background knowledge for text classification has not been investigated yet. On the other hand, text categorization heavily depends on the right understanding and context-sensitive interpretation of words. So, it seems an obvious idea to examine whether knowledge about word senses and relations between words can be exploited to improve text classification.

In this paper, we will show how an operator-based view of rule induction algorithms allows an easy integration of a thesaurus into the learning process. Through this integration, we hope to alleviate a well-known problem in learning approaches to text classification, namely the *skewed distribution of feature values* (Lewis 92): Since more complex features like word patterns will probably seldom occur in the sample set, one gets a lot of rather informative document features with very low frequency. Unfortunately, a standard learning approach cannot decide whether the low frequency of a feature is caused by this effect, by irrelevance of the feature, or by noisy example data. Thus, it will discard this feature according to some selection strategy. Using knowledge as represented by a thesaurus, however, one can cluster several rare features, which, as a cluster, gain more relevance for the classification task. Knowledge-based clustering also allows what we call *generalization beyond the training set*: Even features not occurring in the training set can —as part of a cluster— help to classify new documents.

The paper is organized as follows: First, we introduce our operator-based view on rule induction. We then briefly describe the lexical database WordNet. Section 4 deals with our extension of a simple rule learner for the integration of thesaurus knowledge. Our experiments, observations, and conclusions are discussed in section 5. In section 6, we summarize and sketch future work.

2 An Operator-based View of Rule Induction

The input of the rule induction system is a set of pre-classified sample documents represented by the document representation language. In addition, background knowledge can also be formulated in an own representation language. The output of the rule learner is a set of classification rules. While the heads of these rules indicate classes, the bodies formulate tests on documents.

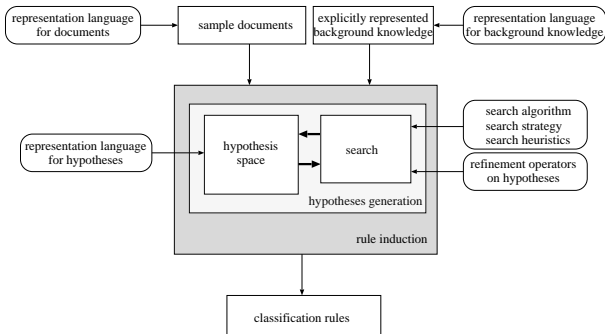


Figure 1: Parameters influencing the search for rule hypotheses

Figure 1 illustrates the basic parameters of the rule learner: The core of the systems consists of a module which generates rule hypotheses formulated in the hypothesis language. The module is starting with a set of hypotheses (e.g. a set containing only the most general hypothesis which is always true). By applying successively *refinement operators*, new hypotheses are constructed. The conjunction of a hypothesis with another hypothesis is an example for a simple refinement operator. Since, as a refinement operator, we admit hypothesis specialization as well as hypothesis generalization, we have a rather powerful general mechanism for finding appropriate classification rules. New knowledge sources can be integrated easily by adding refinement operators as we will demonstrate in section 4. The search is determined by the search algorithm (e.g. hill climbing or beam search), the search strategy (when to apply which operators), and the search heuristics (e.g. the accuracy on the training samples).

3 The WordNet Lexical Database

In order to access thesaurus knowledge we propose to employ machine-readable thesauri as available, e.g. through the WordNet lexical database developed at Princeton University (Miller 95). Based upon the differential theory of

lexical semantics, WordNet captures word meanings by linking a word form to the set of its synonyms (its *synset*), resp. a set of synsets in the case of polysemous words. WordNet provides access to a number of semantic relations between synsets of which we use hyponymy and meronymy. Figure 2 shows a simplified¹ part of the hypernym/hyponym hierarchy of WordNet. The ellipses indicate synsets, the numbers in brackets are their identifiers.

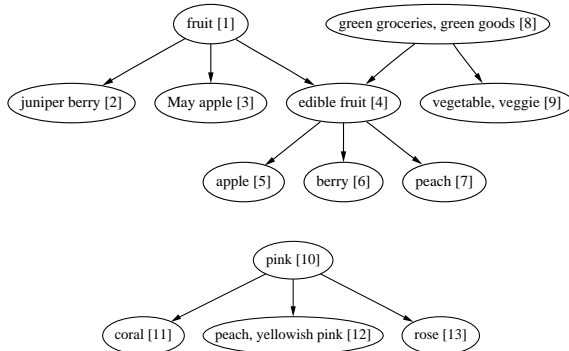


Figure 2: A simplified part of the hypernym/hyponym hierarchy of synsets

4 Rule Induction with Background Knowledge for Text Classification

We now describe a simple rule induction algorithm integrating thesaurus knowledge as an instance of our general framework (see section 2).

Documents in our algorithm are represented by the conjunction of all stemmed words they contain, e.g. a document D only containing the text “text classification” is represented as “text” \wedge “classification”. We also make use of terms built by word sequences. This representation corresponds to the widely used set representation of documents. Up to now, we do not use *explicitly* represented background knowledge.

Our hypothesis language H contains each boolean expression over words and synsets of WordNet. In the following, W denotes the set of all words ($W \subset H$) and S denotes the set of all synsets in WordNet ($S \subset H$). A WordNet synset is interpreted by the disjunction of all words it contains.

In standard rule learning algorithms hypotheses can only be refined by the specialization operator $S_{\wedge, R}$ (R denotes a subset of H , typically $R \subset W$):

¹Some words belonging to the synsets are not captured in the figure.

$$S_{\wedge, R} : H \rightarrow 2^H, h \mapsto \{h \wedge r \mid r \in R\} \cup \{h \wedge \text{not}(r) \mid r \in R\}$$

For simplicity, we integrated a restricted version $S'_{\wedge, R}$ of this operator into our algorithm:

$$S'_{\wedge, R} : H \rightarrow 2^H, h \mapsto \{h \wedge r \mid r \in R\}$$

This operator only considers the conjunction of a hypothesis with word occurrences. R was chosen as the strongest 150 words indicating the target class. This was done using a statistical significance test. So far, our algorithm does not much differ from other rule learning approaches to text classification.

For hypothesis construction, we use the separate-and-conquer strategy (Fürnkranz 96): First, the hypothesis `true` is refined to a 'good' hypothesis by applying refinement operators (*conquer step*). This is done using the hill-climbing strategy. Our search heuristic consists of two parts: The primary criterion is to maximize the accuracy of a rule hypothesis on the training set. In addition, better rules are only accepted if they fulfil a significance criterion. When a hypothesis cannot be refined anymore to a better and significant hypothesis, all positive examples covered by this hypothesis are removed and the algorithm iterates on the remaining examples (*separate step*). This is repeated until no new hypothesis can be found anymore. The resulting hypotheses of each conquer step build a disjunction which yield the final rule.

To integrate the thesaurus, we introduced five new generalization operators. They exploit the synonymy relation, the hypernymy/hyponymy relation and the meronymy relation of WordNet. Figure 3 summarizes the new operators. The following exemplary generalizations are based on the hypernym/hyponym relations shown in figure 2:

- $G_{synsets} : \text{"peach"} \mapsto \{[7], [12]\}$
- $G_1 : [4] \mapsto \{[4] \vee [5] \vee [6] \vee [7]\}$
- $G_2 : [4] \mapsto \{[1] \vee [2] \vee [3] \vee [4], [8] \vee [4] \vee [9]\}$

The thesaurus-based refinement operators were integrated in the following way: Each time a hypothesis could be improved by specializing with a word occurrence, all thesaurus-based operators were applied to all synsets of this word. Figure 4 shows the details.

¹In WordNet, a word can belong to different word categories (like noun and verb) and within each of these categories belong to different synsets. A valid assignment of a category and a synset is called interpretation, here.

5 Experimental Results

For our experiments, we relied on the well-known Reuters corpus, a collection of Reuters newswire articles the use of which has a long tradition in text classification. Since the end of 1996, a revised version of this corpus is available, called Reuters-21578². We used the "ModApte" splitting proposed in the documentation, separating the corpus into 9,603 training and 3,299 test documents. For testing our classification algorithm, we relied on the Reuters corpus organized wrt. the TOPICS set of classes, and wrt. the PLACES set of classes.

Since we expected a large collection of long, elaborated texts already would contain most relevant index terms sufficiently often, we decided to investigate first the more challenging task of classifying solely with the titles. This may appear to be a rather artificially constructed scenario in view of the fact that the complete texts were available. However, there exist relevant applications with similar properties, e.g. in multimedia information archives or literature databases, where complex documents (books, videos, images) are represented only by a short text, an abstract, or some index terms. On the other side, we first chose the task of classifying according to the PLACES because the domain of geographical relations seemed to be a rather comprehensive area of WordNet.

Learning performance for a class C was measured on the 3299 test documents, using recall (r), precision (p), and the f_β -measure (Rijsbergen 79). To give equal weight to recall and precision in f_β , we chose $\beta = 1$.

Figures 5 and 6 show an example of a rule for the class 'Netherlands' constructed without thesaurus-based operators and with such operators. The rule generated with thesaurus information is presented in an intermediate representation which allows to illustrate the construction history. In this case the refinement operator G_4 was applied to the word "rotterdam". This refinement included all other cities of the Netherlands as well as the synset denoting the Netherlands itself to the final rule. Evaluation on the test set demonstrated the superiority of this rule to the rule generated without thesaurus.

²Available on <http://www.research.att.com/~lewis>.

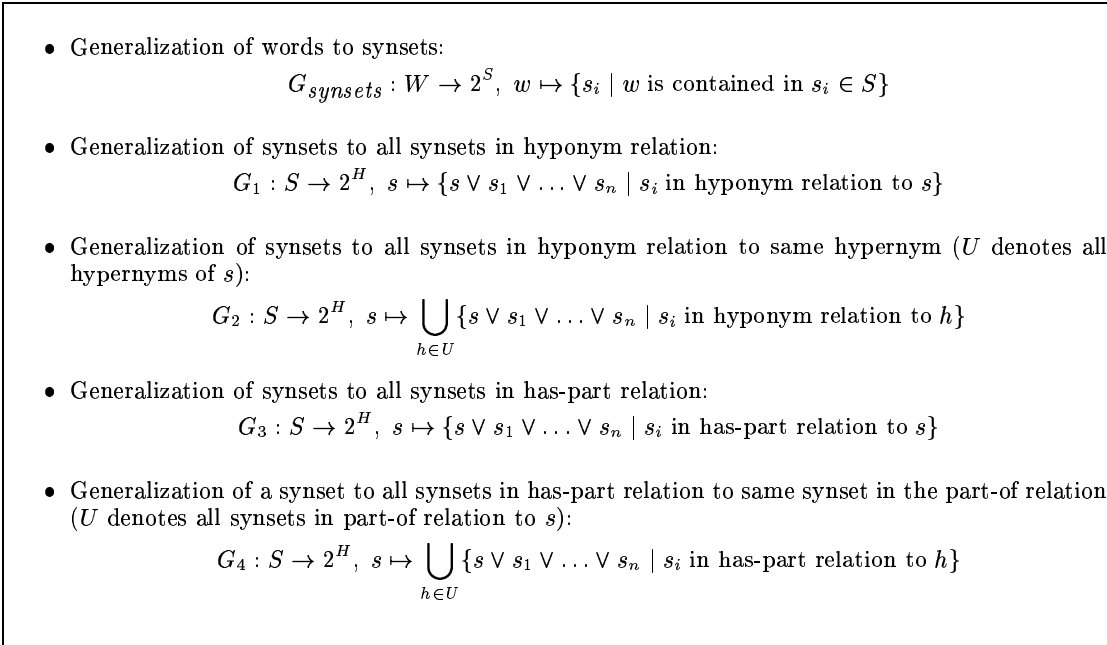


Figure 3: Thesaurus-based generalization operators

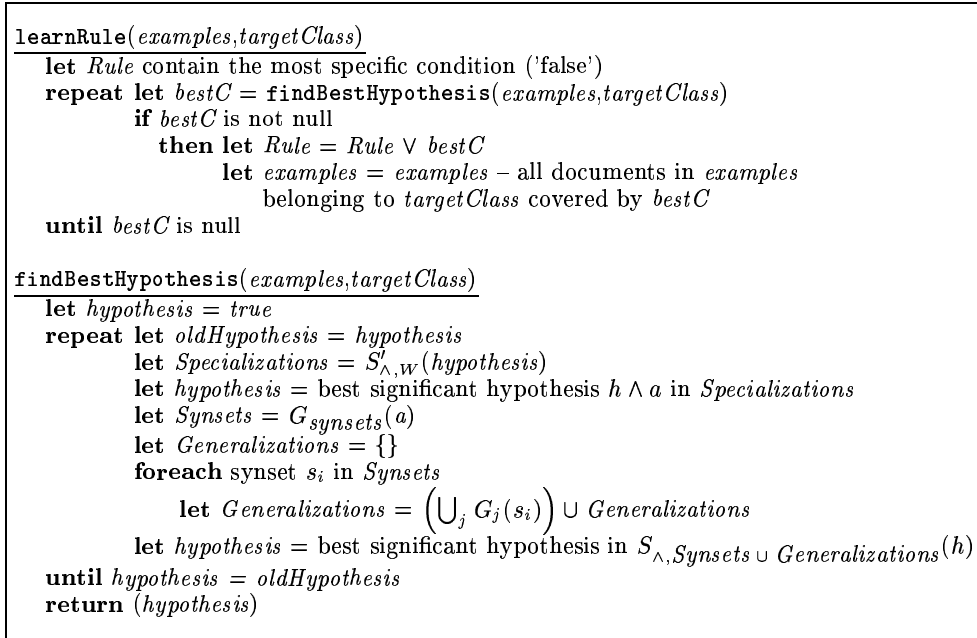


Figure 4: A simple rule induction algorithm for text classification, extended by thesaurus-based generalization operators

We now want to turn back to our motivation at the beginning. The main problem we wanted to tackle by our extension was the problem of skewed distribution of feature values. An analysis of the rule construction process shows, that the word “netherlands” only occurs once in the training set. This occurrence is in a document of class “Netherlands”. Without using the thesaurus knowledge, we would not have been able to integrate this word into the final rule. It might have occurred in this class just by chance. It reveals that the inclusion of this word into the rule helps to classify one more document of the test set correctly. In addition the inclusion of the term “the hague” supports our hope to allow generalization beyond the training set. While not occurring in the training set at all, this term also helps to classify one more document of the test set correctly.

```
[netherlands] :-
(or dutch rotterdam klm guilder akzo
  amsterdam suralco hoogovens)
```

Figure 5: Classification rule for articles about the “Netherlands” obtained without thesaurus knowledge

```
[netherlands] :-
(or dutch
  (synset noun #5565024:
    holland netherlands
    the_netherlands
  (hasparts (synset noun #5564550: amsterdam dutch_capital
    capital_of_the_netherlands)
    (synset noun #5564733: the_hague 's_gravenhage
    den_haag dutch_capital
    capital_of_the_netherlands)
  (synset noun #5564952: nijmegen)
  (synset noun #5565024: rotterdam)
  (synset noun #5565097: leyden leiden)
  (synset noun #5565176: utrecht)
  (synset noun #5746541: meuse meuse_river)
  (synset noun #5750223: rhine rhine_river)))
  klm guilder akzo suralco hoogovens)
```

Figure 6: Classification rule for articles about the “Netherlands” obtained using thesaurus-based refinement operators

Results on other classes were also promising, though thesaurus knowledge was only very rarely incorporated. Figure 7 lists some results for the classes with the best improvements. It is important to note that results as measured by f_1 never got worse in these experiments by using thesaurus knowledge. Similar results could also be observed in two other domains with similar characteristics, namely a set of emails and a set of HTML-documents (the documents were classified by personal interests).

On the contrary, we tested our algorithm on the Reuters TOPICS set of classes based upon the complete articles. Here too, we found a small number of classes where effectiveness improved. But in contrast to the observations made before, we also found classes with a decrease of effectiveness. The worse results when using the thesaurus can be explained by a too close adaptation of the rule to the thesaurus structure. We would like to demonstrate this by an example: When searching a classifier for the class “grain”, we got a rule testing on the synset denoting grain and all synsets in hyponym relation (i.e. all types of grain like corn, wheat, and rice). This rule performed worse than the rule generated without the thesaurus. The latter rule was only testing on a selection of grain types. In particular, it was excluding the synsets rice and oat which caused a number of misclassifications on the training set as well as on the test set. The example illustrates that the semantic structure of the world given by the thesaurus can affect adversely the rule construction. When concepts in the thesaurus and concepts needed for rule construction differ much, this can be identified easily by the search heuristic. When concepts are very close but still differ, there is a risk to decide erroneously in favor of the thesaurus concept.

<i>class</i>	<i>size</i>	$r \pm \Delta r$	$p \pm \Delta p$	$f_1 \pm \Delta f_1$
netherlands	32	46.9+6.2	88.2-3.2	61.2+4.2
canada	182	32.4+2.2	73.8+1.2	45.0+2.4
west-germany	97	47.4+3.1	68.7+0.3	56.1+2.2
china	41	75.6+2.4	93.9-2.5	83.8+0.4
venezuela	18	77.8±0.0	87.5±0.0	82.4±0.0
...				

Figure 7: Some performance data demonstrating the change in effectiveness when using WordNet. The number of positives examples in the test set is given by the column *size*.

We would like to summarize our experiences in the following way:

1. Domains with relatively sparse and heterogeneous training material can benefit most from using thesaurus knowledge. The thesaurus has to fit in some sense to the domain and classes. If not, there is a risk of integrating too much of the thesaurus structure into the rules. This may lead to a loss of effectiveness.
2. It is an important problem to prevent rule induction algorithms from adapting rules to

closely to the thesaurus structure. The highest risk here is represented by thesaurus concepts which lie very close to concepts beneficial for rule induction. Here, work on the search parameters seems promising.

3. In our instantiation of a rule induction algorithm the thesaurus was only rarely used. This can partly be explained by the homogeneous corpus we used for evaluation. To force the integration of thesaurus knowledge, a broader search is helpful. This leads to more improvements by the thesaurus as shown in initial experiments not documented here.

6 Summary and Future Work

In consequence of the “Information Flood”, automated text classification is a problem of growing commercial interest. It can be used for automated filtering, routing and archiving. In this paper, we demonstrated how the common-sense knowledge about word senses and relations between words can contribute to improved learning of text classification rules. We understand rule induction basically as a search in the hypothesis space. This allows an easy integration of additional knowledge by simply adding new refinement operators. We extended a simple rule induction algorithm by five new generalization operators which exploit thesaurus knowledge.

We motivated our work with the positive effects thesaurus knowledge may have on rule induction, namely the tackling of problems with skewed feature values and the chance of generalizations beyond the training set. The presented results illustrate that in fact rule induction can benefit from using a thesaurus. In particular, this holds for heterogeneous text collections with sparse training data when an appropriate thesaurus is used. However, the experiments also showed cases where effectiveness decreased when using the thesaurus. The problem in these cases was identified as a bad fitting of the concepts as needed by the rule induction and the concepts as represented by the thesaurus. We think that more elaborated search parameters can alleviate this problem. Moreover, work on search parameters can lead to a more frequent use of thesaurus knowledge.

Our future research follows two directions: The investigation of the search parameters for rule induction in presence of thesaurus-based operators and the semi-automatic construction of domain-dependent thesauri for classification problems.

The latter is currently pursued in an application project for an industrial partner who wants to improve automated routing of research reports and product information.

References

- (Apté *et al.* 94) C. Apté, F. Damerau, and S. Weiss. Towards Language Independent Automated Learning of Text Categorization Models. In *Proceedings of the 17th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (SIGIR 94)*, pages 23–30, Dublin, Ireland, July 3-6 1994.
- (Cohen 96) W.W. Cohen. Learning to Classify English Text with ILP Methods. In *Advances in Inductive Logic Programming*, pages 124–143. IOS Press, 1996.
- (Finch 95) S. Finch. Partial Orders for Document Representation: A New Methodology for Combining Document Features. In *Proceedings of the 18th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (SIGIR 95)*, pages 264–272, Seattle, Washington, USA, July 9-13 1995.
- (Fürnkranz 96) J. Fürnkranz. Pruning Algorithms for Rule Learning. Technical Report OEFAI-TR-96-07, Austrian Research Institute for Artificial Intelligence, Schottengasse 3, A-1010 Wien, Austria, 1996.
- (Lewis 92) D.D. Lewis. *Representation and Learning in Information Retrieval*. Unpublished PhD thesis, Department of Computer Science, University of Massachusetts, 1992.
- (Miller 95) G.A. Miller. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41, 1995.
- (Rijsbergen 79) Van Rijsbergen. *Information Retrieval*. Butterworth, London, England, 1979.
- (Riloff 96) E. Riloff. Using Learned Extraction Patterns for Text Classification. In *Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing*, pages 274–289. Springer, Berlin, Heidelberg, New York, 1996.
- (Sahami *et al.* 96) M. Sahami, M. Hearst, and E. Saund. Applying the Multiple Cause Mixture Model to Text Categorization. In *Machine Learning. Proceedings of the 13th International Conference (ICML 96)*, pages 435–443, Bari, Italy, July 3-6 1996.
- (Wiener *et al.* 95) E. Wiener, J.O. Pedersen, and A.S. Weigend. A Neural Network Approach to Topic Spotting. In *Proceedings of the 4th Symposium on Document Analysis and Information Retrieval (SDAIR 95)*, pages 317–332, Las Vegas, NV, USA, April 24-26 1995.
- (Yang 95) Y. Yang. Noise Reduction in a Statistical Approach to Text Categorization. In *Proceedings of the 18th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval (SIGIR 95)*, pages 256–263, Seattle, Washington, USA, July 9-13 1995.