

Classifying Patent Applications with Winnow

C.H.A. Koster, M. Seutter and J. Beney

University of Nijmegen, The Netherlands.

{kees,marcs,jean}@cs.kun.nl

Abstract

The Winnow family of learning algorithms can cope well with large numbers of features and is tolerant to variations in document length, which makes it suitable for classifying large collections of large documents, like patent applications.

This note describes a number of experiments using Winnow in classifying two large corpora of patent applications, supplied by the European Patent Office, and comparing it to Rocchio.

Both the large size of the documents and the large number of available training documents for each class make this classification task qualitatively different from the classification of short documents (newspaper stories and medical abstracts) with few training examples, as exemplified by the TREC evaluations.

Keywords: Machine Learning for IR, Winnow Algorithm, Rocchio, Text Categorization.

1 Introduction

Automatic text classification (also known as *text categorization*) plays an increasing role in the automatic processing of full-text documents: in *routing* and *filtering*, but also in the contents-based disclosure of large numbers of documents on the Web. A comprehensive overview of the state-of-the-art in automatic text classification can be found in (Sebastiani, 2000) and in numerous bibliographies on Information Retrieval and Machine Learning on the Internet.

The automatic classification of patent applications is an interesting application area for automatic learning techniques, due to the large number of features (avg 5000 words per document) and the large number of pre-classified documents available. Furthermore, each document has been classified carefully by experts highly familiar with the domain at hand.

Traditional classification tasks, as exemplified by the Reuters dataset (Apté and Damerau, 1994) and the collections used in TREC, have been concerned with smaller documents (newspaper stories, medical abstracts) and classes for which fewer pre-classified examples are available (between one and a few tens). Because of these qualitative and quantitative differences, different techniques may have to be used.

The Winnow algorithm has nice mathematically guaranteed convergence properties (Grove et al, 2001). According to (Dagan et al, 1997), the balanced version of the Winnow algorithm can cope with large numbers of features and can tolerate large variations in document length because it uses both positive and negative weights. Previous experiments (Ragas and Koster, 1998) have suggested that Balanced Winnow performs better than the ubiquitous Rocchio algorithm in the case where a large number of training documents is available for each class.

We have compared the suitability of the Balanced Winnow algorithm for the task of classifying patent applications to that of Rocchio in a number of experiments: a monofclassification of 32000 documents in 16 classes and two multi-classifications of 102418 documents in 44 and 549 classes, respectively. The largest corpus contained more than 500 000 different terms.

1.1 The EPO tests

The European Patent Office (EPO) in Rijswijk, the Netherlands, has been investigating the use of Natural Language Processing (NLP) and automatic classification techniques for the pre-classification of its incoming patent applications. To that end, it has organised during the year 2000 an open competition in order to assess the state-of-the-art in supervised learning

for document classification.

Over the years, EPO has collected more than four million patent applications, primarily in English, which have been manually classified by human experts according to a detailed hierarchy of semantical categories. The EPO has about 140.000 patent applications filed each year. For the search phase, each of them must be forwarded to one of more than 2500 *patent examiners* (aka *gerants*) working in three locations (The Hague, Munich or Berlin). The technical domain for which each examiner is responsible is defined by means of ranges of European Classification (ECLA) codes.

The final objective for the pre-classification is thus that of identifying for each patent application one correct category (as ECLA range) of around 1000 possible ones.

Since this number of categories is so large that a flat categorization scheme is not suited, EPO investigated a two-step categorization process. The first step is based on the *directorate*. There are 44 directorates in The Hague covering all the technical domains, so in the first categorization step each patent application has to be assigned one out of 44 directorate codes.

The second categorization step is based on ECLA: in most cases a subclass level is sufficient, e.g. B61K, in some other cases we have to go down at main group level, e.g. A01B13, or to a range of codes, e.g. G01S13/50:G01S13/66. For a categorization procedure however the way the category is defined is not important, their total number is. Therefore the second step was simplified to 624 ECLA subclasses. The other simplifications made by EPO in setting up the classification task were:

- English language documents only.
- US patent documents only.
- Typeset quality (no OCR error issues).

In spite of these simplifications, the tasks are quite representative of the intended application.

2 The algorithms

The wellknown Rocchio algorithm (Rocchio, 1971; Cohen and Singer, 1999) is the typical workhorse for document retrieval and classification in Information Retrieval. It is based on the Vector Space model and computes a classifier essentially as a vector of term weights.

The less well-known Winnow is a heuristical learning algorithm, iterating repeatedly over the training documents in order to obtain an optimal linear separator between relevant and non-relevant documents (Littlestone, 1988).

In terms of the admirable exposition given in (Dagan et al, 1997), what we have implemented is the *Balanced Winnow* algorithm with the *thick threshold* heuristic. We have made a slight improvement to its speed of convergence by using *normalised term strengths* to initialize the Winnow weights.

3 EPO1: Mono-classification

The first EPO test concerned mono-classification of a corpus of 16000 patent applications (EPO1) into 16 disjoint classes (*clusters* of directorates) with a thousand training examples each.

Two CD-rom disks containing public patent material in English were supplied by EPO. Each disk contained 1000 abstracts and 1000 full documents for each of 16 classes. The first disk contained the training set (known classification), the second disk the test set (unknown classification). Some statistics are given in Table 1.

The material is very homogeneous in document size, and evenly distributed over the classes. The abstracts, with an average length of 129 words, are in descriptive prose without tables or formulae. The full text documents are longer (avg 4580 words) and contain some non-linguistic material like tables and formulae.

The only pre-processing applied to the documents was the replacement of capital letters by the corresponding small letters (decapitalization) and the replacement of special characters by spaces. In particular, no term stemming was performed and no stoplist was applied. We relied on the term selection to eliminate redundant features.

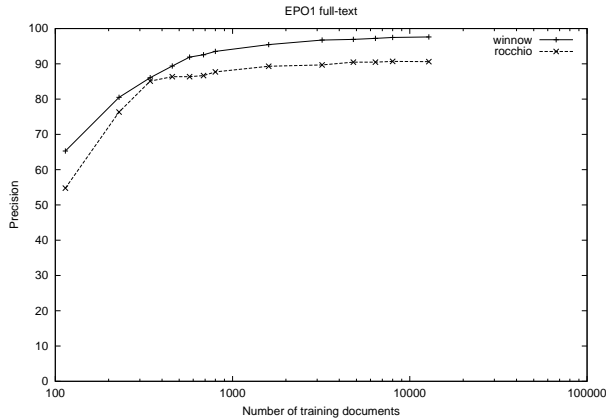
3.1 Experiments on held-out documents

From the set of full-text training documents a new train and test set was obtained by randomizing the order and splitting the set (80% training, 20% held out for testing). This new training set was trained in epochs of increasing size and the resulting classifier tested on the remaining 20%. The experiment was performed for each algorithm, and repeated three times, in

	documents	lines	words	characters
full-text	16000	7.6M	73.3M	461.1M
abstracts	16000	.5M	3.9M	25.6M

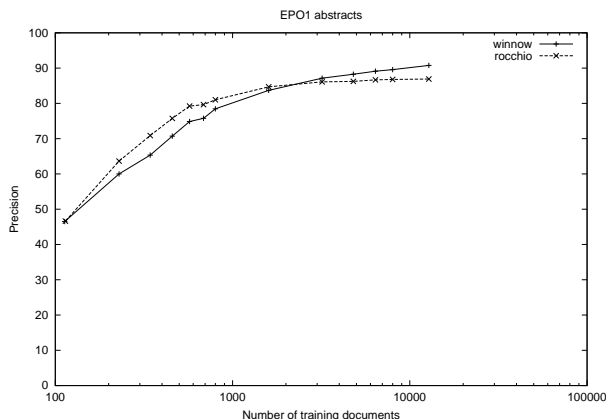
Table 1: Statistics for EPO1 corpus

order to reduce statistical variation. The results (micro-averaged Precision) are as follows:



This result is encouraging, with Precision approaching 98 percent in spite of the large size of the documents. The classes appear to be sufficiently disjoint, and the noise on the classification appears to be small. At very low numbers of training examples (229 examples means an average of 13 examples per class) the Precision is already above 65 percent. Although learning slows down after about 3000 documents trained, there is still a gradual improvement. Winnow outperforms Rocchio all the way, especially for larger numbers of training documents.

The same experiment was performed on a 80/20 split of the abstracts:



This time Rocchio outperforms Winnow for fewer than 80 examples per class, but it ap-

pears to reach a plateau whereas Winnow keeps learning from new examples. The Precision is definitely lower for abstracts than for full-text documents. This contradicts (at least for patent applications) the expectation that abstracts should be easier to classify than the corresponding documents because they contain fewer digressions and the words used are more carefully chosen, thus more pertinent.

3.2 Precision on seen documents

Another experiment was performed using 50 percent of all pre-classified documents as the train set and the same set as test set. The Precision, Recall and F1-value on “seen” documents obtained in this test were as follows:

	Precision	Recall	F1
Rocchio abstracts			
macro average	89.79	89.26	89.34
micro average	89.22	89.22	89.22
Winnow abstracts			
macro average	99.62	99.61	99.61
micro average	99.61	99.61	99.61
Rocchio full-text			
macro average	92.30	91.90	91.97
micro average	91.99	91.99	91.99
Winnow full-text			
macro average	100.00	100.00	100.00
micro average	100.00	100.00	100.00

The *macro-averaged* Precision is the average over all classes of the Precision per class, whereas the *micro-averaged* Precision is computed by lumping the classes together (and similarly for Recall and F1). In monclassification, the micro-averaged Precision and Recall are equal. The F1-value is a harmonic average of Precision and Recall:

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$$

In distinction to Rocchio, Winnow can “explain” the class given to most abstracts and all

full-text documents. This can also be seen as evidence of “overtraining” on the seen documents.

3.3 First blind test

The main test consisted of training Winnow on the data on the first disk and then testing on the 16000 test documents on disk 2. The result of the classification should give the most probable class for each document in the test set. Since the correct classes were unknown to us, the evaluation of the result was performed by EPO, which reported the following results:

- Testing on 16000 unseen documents, of the 16000 predictions made, 15497 were right and 503 were wrong, giving an error rate of 3.14%.
- Testing on 16000 unseen abstracts, of the 16000 predictions made, 14941 were right and 1059 were wrong, giving an error rate of 6.62%.

EPO also informed us that ten different systems had been tested, that our system had performed best of all and that “management was impressed”.

4 EPO2: Multi-classification

The second test involved a more realising problem: a multi-classification with many classes, training on a larger number of documents and taking as test documents a large subset out of a year’s incoming documents.

The data for the second test (EPO2 corpus) consisted of 9 CD-ROMs, containing 68418 pre-classified documents as training set and 44000 unclassified documents as test set, both in the Full Text (F) form and in the form of abstracts (A).

For these documents, different pre-classifications were provided by EPO: document:

1. 44 directorates
A directorate is an organizational unit within EPO, responsible for all patent applications in a certain domain
2. 549 examiner classes
An examiner is an expert in a certain (sub)domain working in one of the directorates. Documents are classified by a

number of examiners, sometimes from different directorates.

A third one, based on 624 ECLA subclasses was not attempted by us.

These tests were harder than the previous, due to the large size of the test material but in particular due to the fact that this test concerned *multi-classification* rather than mono-classification.

4.1 Multiclassification issues

Conceptually, multi-classification (where each document belongs to zero or more classes) is simpler than mono-classification (where each document belongs to precisely one class), since it can be seen as training an independent binary classifier for each class. In fact the training procedure is the same (apart from the fact that each document now may belong to more than one class), but the *selection* of classes for documents is more complicated: rather than assigning each document to the class for which it has the highest score, in multicassification it is necessary to determine for each class an optimal *threshold* to distinguish between relevant and non-relevant documents for that class. Winnow has a natural threshold of 1 because of the way it is trained, but for other algorithms the threshold computation is complicated (Arampatzis and van Hameren, 2001; Zhang and Callan, 2001).

The boundary condition “each document belongs to precisely one class” provides important information, as can be seen by weakening it to “each document belongs to at most one class”. In this case, tests on seen documents give the following results:

	Precision	Recall	F1
Rocchio abstracts			
macro average	94.82	76.37	84.39
micro average	94.86	76.29	84.56
Winnow abstracts			
macro average	99.87	98.83	99.35
micro average	99.87	98.83	99.35
Rocchio full-text			
macro average	97.31	74.54	84.25
micro average	97.39	74.54	84.44
Winnow full-text			
macro average	100.00	99.80	99.90
micro average	100.00	99.80	99.90

	documents	lines	words	characters
train set				
full-text	68418	44.7M	401.3M	2607.3M
abstracts	68418	1.0M	9.0M	59.5M
test set				
full-text	44000	28.2M	253.2M	1642.6M
abstracts	44000	.6M	5.9M	38.4M

Table 2: Statistics for EPO2 corpus

In comparison to the table in section 3.2, Precision increases, Recall is reduced and the F1-value also decreases. This is caused by the fact that documents are no longer forced into a class even when their score falls under the threshold of that class.

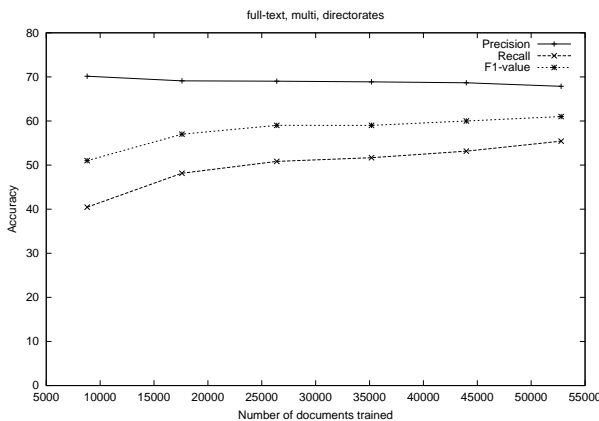
4.2 44 directorates

In the “directorate” classification, an average of 1.286 classes is assigned to each of the 68418 training documents. The total number of <document, class> pairs is 88000.

The pre-processing applied to the documents was the same as in the EPO1-test, but this time all terms with a small term frequency were eliminated (TF selection with $k = 4$ for full-text documents and $k = 3$ for the abstracts).

4.2.1 Classification of full-text documents

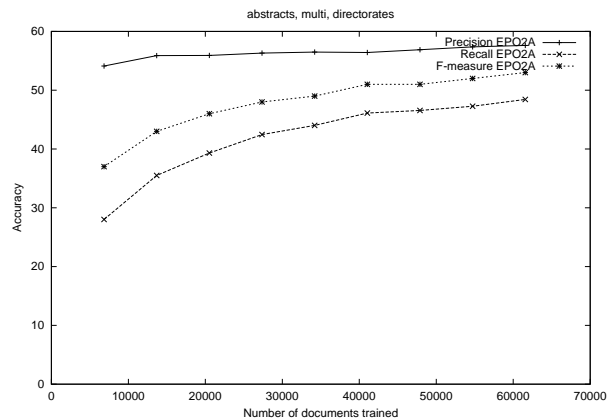
The documents were randomly split 80/20 in a train and test set, and classifiers were trained using the Winnow algorithm in nine epochs. Due to time limitations, only 6 epochs out of 9 were trained. The resulting learning curve, showing accuracy (Precision, Recall and F1-value) as a function of the number of documents trained, was as follows:



It can be seen that, as more documents are trained, the Recall improves, but the Precision actually decreases by one percent point.

4.2.2 Classification of abstracts

The abstracts were also split 80/20 in a train and test set. Classifiers were trained using the Winnow algorithm in nine epochs of 6081 abstracts each. The results were as follows:



This time both Precision and Recall improve in training. As in the first EPO test, the Precision achieved on abstracts is less than that on full-text documents.

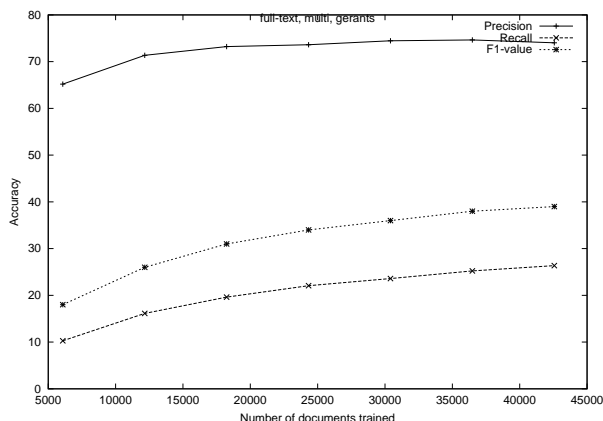
4.3 549 examiners

The hardest test concerned multi-classification with 549 “examiner” classes. In this classification, each of the 68418 training documents belonged to an average of 2.43 classes. The documents were the same as in the previous test.

4.3.1 Classification of full-text documents

The documents were split 80/20 in a train and test set, and classifiers were trained using the Winnow algorithm in nine epochs. Due to time limitations, only 6 epochs out of 9 were trained. The resulting classification was compared with the actual classification of the held-out docu-

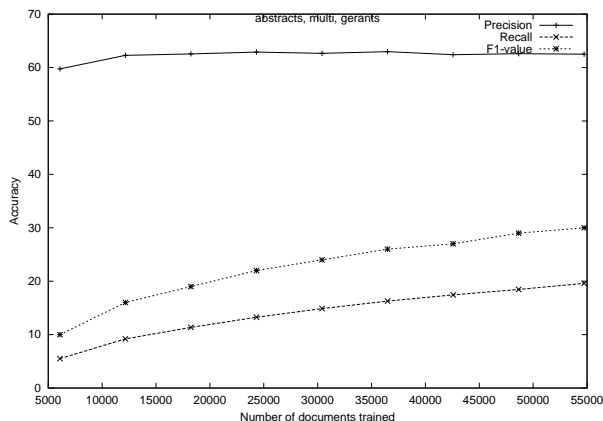
ments to compute Precision, Recall and the F1 value. The results obtained were as follows:



The Recall is disappointingly low. It improves as more documents are trained, but the Precision actually decreases by one percent point.

4.3.2 Classification of abstracts

The abstracts were also split 80/20 in a train and test set, and classifiers were trained in nine epochs. The results are as follows:



This time both Precision and Recall improve in training. As in the first EPO test, the Precision achieved on abstracts is less than that on full-text documents.

4.4 Experiments with lemmatization

It is common wisdom in IR that lemmatization of the wordforms in a document, i.e. *morphological normalization* improves Recall with at most a small reduction in Precision.

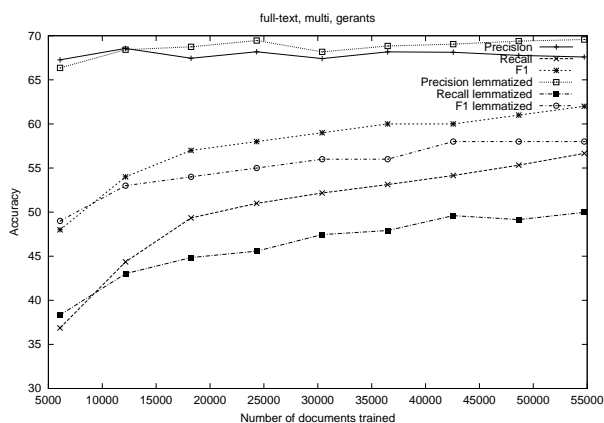
Lemmatization reduces the number of terms and raises the number of occurrences of particular terms, thereby reducing the sparsity of the feature space, which is expected to have a beneficial effect.

On the other hand, most of the words in a text do not contribute at all to the classification,

and should be eliminated rather than stemmed. Furthermore, in some cases minor differences in wordform may be important for classification¹.

In order to study the opposing effects of lemmatization, we prepared lemmatized versions of both the documents and the abstracts of EPO2/directorates, using a lemmatizer derived from Wordnet 1.6.

The following graph compares the classifications for full-text documents with and without lemmatization.

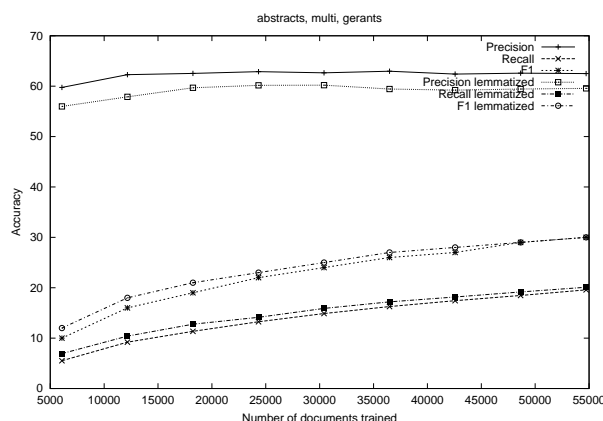


Lemmatization appears to have a small positive effect on the Precision for full-text documents, but this is offset by a large decrease in Recall. Although this result has to be interpreted with some caution (the effectiveness of the rather primitive lemmatizer on the EPO data may well be low) there does not appear to be a substantial advantage in applying lemmatization to the full-text documents.

In longer documents, the morphological variants of a word tend to occur more frequently, so that they may become significant classification terms by themselves, even without lemmatization. By the same argument, the benefits of lexical normalization should be somewhat larger for abstracts.

The learning curve for the abstracts with and without lemmatization is as follows:

¹In a striking example due to Riloff, documents containing the word 'murders' were likely to be about terrorism, whereas those containing 'murder' were not; Bob Krovetz has found that "accelerator" refers to either a car accelerator or a nuclear particle accelerator in newspaper text, but "accelerators" referred only to nuclear particle accelerators.



For the (shorter) abstracts, the Precision decreases by lemmatization with a very small rise in Recall. In this respect, there appears to be a difference between long and short documents.

4.5 Second blind test

After training on all full-text training documents, the test documents provided by EPO were classified and the resulting ranked list for each class sent to EPO. This concerned both the classification on directorates (44 classes) and the classification on examiners (549 classes). The final results as communicated to us by EPO are shown in the Appendix (see also (Krier and Zaccà, 2001)). We were participant B, which came out best in both tests.

5 Conclusions and further work

The experiments described above show that the Winnow algorithm is indeed suitable for the classification of large collections of large documents, such as patent applications. In mono-classification, Winnow is preferable to Rocchio for large trainsets.

The accuracy on full-text documents was much larger than that on abstracts, contradicting the expectation that abstracts would be easier to classify, because of the more expressive terminology used and the lack of digressions.

The difference in Precision and Recall between mono-classification and multi-classification is vexing. Even though our system was the most accurate out of all participants in the EPO tests, which proves that it is good at ranking the documents, in multi-classification it is not very good at separating relevant from non-relevant documents. The assignation of documents to

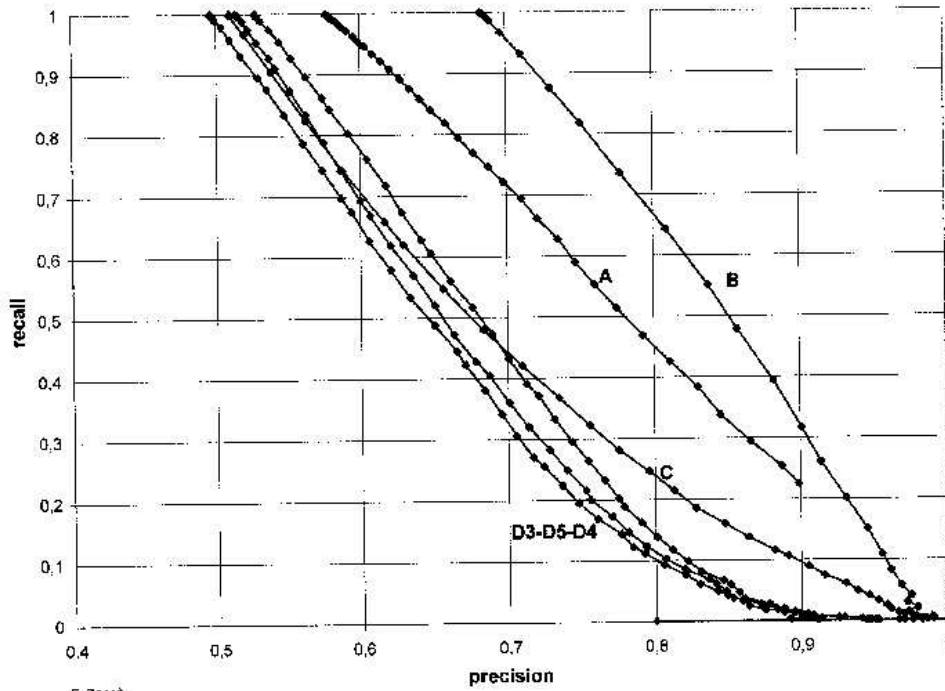
classes in multi-classification requires further investigation.

The documents were represented quite conventionally, without any linguistic preprocessing. In particular, no semantical ontology was used, and discourse structure was completely ignored. We are presently investigating more informative document representations.

References

- Apté, C. and Damerau, F. (1994) Automated learning of decision rules for text categorization. In: *ACM Transactions on Information Systems* 12(3):233-251, 1994.
- A. Arampatzis and A. van Hameren (2001), The Score-Distributional Threshold Optimization for Adaptive Binary Classification Tasks. In: W. Bruce Croft et al (Eds.), *Proceedings of SIGIR 2001*, ACM Press, pp. 267-275.
- W.W. Cohen and Y. Singer (1999), Context-sensitive learning methods for text categorization. *ACM Transactions on Information Systems* 13, 1, 100-111.
- I. Dagan, Y. Karov, D. Roth (1997), Mistake-Driven Learning in Text Categorization. In: *Proceedings of the Second Conference on Empirical Methods in NLP*, pp. 55-63.
- A. Grove, N. Littlestone, and D. Schuurmans (2001), General convergence results for linear discriminant updates. *Machine Learning* 43(3), pp. 173-210.
- M. Krier and F. Zacca (2001), Automatic Categorisation Applications at the European Patent Office, International CHEMical Information Conference, Nimes, October 2001, 10 pp.
- N. Littlestone (1988), Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2, pp. 285-318.
- H. Ragas and C.H.A. Koster, Four classification algorithms compared on a Dutch corpus. *Proceedings SIGIR 98*, pp. 369-370.
- J.J. Rocchio (1971), Relevance feedback in Information Retrieval, In: Salton, G. (ed.), *The Smart Retrieval system - experiments in automatic document processing*, Prentice - Hall, Englewood Cliffs, NJ, pp 313-323.
- F. Sebastiani (2001), Machine learning in automated text categorization. Technical Report IEI-B4-31-1999, Istituto di Elaborazione dell'Informazione, Consiglio Nazionale delle Ricerche, Pisa, IT, 1999. Submitted for publication to ACM Computing Surveys.
- Y. Zhiang and J. Callan (2001), Maximum Likelihood Estimation for Filtering Thresholds, In: W. Bruce Croft et al (Eds.), *Proceedings of SIGIR 2001*, ACM Press, pp. 294-302.

Precision-Recall Diagram at Examiner Level



Precision-Recall diagram at Directorate Level

