Integrating a Lexical Database and a Training Collection for Text Categoriza tion

José María Gómez-Hidalgo, Manuel de Buenaga Rodríguez

{jmgomez,mbuenaga}@dia.ucm.es

Departamento de Informática y Automática

Universidad Complutense de Madrid
Avda. Complutense s/n, 28040 Madrid (Spain)

Abstract

Automatic text categorization is a complex and useful task for manynatural language processing applications. Recent approaches to textcategorization focus more on algorithms than on resources involved in thisoperation. In contrast to this trend, we present an approach based on the integration of widely available resources aslexical databases and training collections to overcome current limitationsof the task. Our approach[1] makes use of WordNet synonymy information toincrease evidence for bad trained categories. When testing a direct categorization, a WordNet basedone, a training algorithm, and our integrated approach, the latter exhibitsa better perfomance than any of the others. Incidentally, WordNet based approach perfomance is comparable with the trainingapproach one.

1 Introduction

Text categorization (TC) is the classification ofdocuments with respect to a set of one or more pre-existing categories. TCis a hard and very useful operation frequently applied to the assignment of subject categories to documents, toroute and filter texts, or as a part of natural language processingsystems.

In this paper we present an automatic TC approach based on theuse of several linguistic resources. Nowadays, many resources like trainingcollections and lexical databases have been successfully employed for text classificationtasks [Boguraev and Pustejovsky, 1996], but always in an isolated way. Thecurrent trend in the TC field is to pay more attention to algorithms thanto resources. We believe that the key idea for the improvement of text categorization is increasing theamount of information a system makes use of, through the integration ofseveral resources.

We have chosen the Information Retrieval vector space model for ourapproach. Term weight vectors are computed for documents and categoriesemploying the lexical database WordNet and the training subset of the testcollection Reuters-22173. We calculate the weight vectors for:

- A direct approach,
- a Wordnet based approach,
- a training collection approach,
- and finally, a technique for integrating WordNet and a training collection.

Later, we compare document-category similarity by means of a cosine-basedfunction. We have driven a series of experiments on the test subset ofReuters-22173, which yields two conclusions. First, the integrated approach performs better than any of the other ones, confirming thehypothesis that the more informed a text classification system is, thebetter it performs. Secondly, the lexical database oriented technique can rival with the training approach, avoiding the necessity ofcost-expensive building of training collections for any domain andclassification task.

2 Task Description

Given a set of documents and a set of categories, the goal of acategorization system is to decide whether any document belongs to anycategory or not. The system makes use of the information contained in adocument to compute a degree of pertainance of the document to each category. Categories are usually subject labels likeart or military, but other categories like text genres are also interesting[Karlgren and Cutting, 1994]. Documents can be news stories, e-mailmessages, reports, and so forth.

The most widely used resource for TC is the training collection. Atraining collection is a set of manually classified documents that allowsthe system to guess clues on how to classify new unseen documents. Thereare currently several TC test collections, from which a training subset and a test subset can be obtained. Forinstance, the huge TREC collection [Harman, 1996], OHSUMED [Hersh *etal,* 1994] and Reuters-22173 [Lewis, 1992] have been collected for thistask. We have selected Reuters because it has been used in other work,facilitating the comparison of results.

Lexical databases have been rarely employed in TC, but severalapproaches have demonstrated their usefulness for term classification operations like word sense disambiguation[Resnik, 1995; Agirre and Rigau, 1996]. A lexical database is a referencesystem that accumulates information on the lexical items of one o

[1] This research is supported by the Spanish Committee of Science andTechnology (CICYT TIC94-0187).

several languages In this view,machine readable dictionaries can also be regarded as primitive lexicaldatabases. Current lexical databases include WordNet [Miller, 1995], EDR[Yokoi, 1995] and Roget's Thesaurus. WordNet 's large coverage and frequent utilization has led us touse it for our experiments.

We organize our work depending on the kind and number ofresources involved. First, a direct approach in which only the categoriesthemselves are the terms used in representation has been tested. Secondly, WordNet by itself has been usedfor increasing the number of terms and so, the amount of predictinginformation. Thirdly, we have made use of the training subset of Reuters toobtain the categories representatives. Finally, we have employed both WordNet and Reuters to get a betterrepresentation of undertrained categories

3 Integrating Resources in the Vector SpaceModel

The Vector Space Model (VSM) [Salton and McGill, 1983] is a very suitableenvironment for expressing our approaches to TC: it is supported by many experiences in textretrieval [Lewis, 1992; Salton, 1989]; it allows the seamless integrationof multiple knowledge sources for text classification, and it makes it easyto identify the role of every knowledge source involved in the classification operation In the nextsections we present a straightforward adaptation of the VSM for TC, and theway we use the chosen resources for calculating several model elements.

3.1 Vector SpaceModel for Text Categorization

The bulk of the VSM for Information Retrieval (IR) is representing naturallanguage expressions as term weight vectors. Each weight measures theimportance of a term in a natural language expression, which can be adocument or a query. Semantic closeness between documents and queries is computed by the cosine of the anglebetween document and query vectors.

Exploiting an obvious analogy between queries and categories,the latters can be represented by term weight vectors Then, a category canbe assigned to a document when the cosine similarity between them exceeds acertain threshold, or when the category is highly ranked. In a closer look,and given three sets of N terms, M documents and Lcategories, the weight vector for document j is $\langle wd_{1j}, wd_{2j}, \ldots, wd_{Nj}\rangle$ and the weight vector for category k is $\langle wc_{1k}, wc_{2k}, \ldots, wc_{Nk}\rangle$. The similarity between document j and category k is obtained with the formula:

$$sim(d_j, c_k) = \frac{\sum_{i=1}^{N} wd_{ij} \cdot wc_{ik}}{\sqrt{\sum_{i=1}^{N} wd_{ij}^2 \cdot \sum_{i=1}^{N} wc_{ik}^2}}$$

Term weights for document vectors can be computed making use of wellknown formulae based on term frequency. We use the following one from[Salton, 1989]:

$$wd_{ij} = tf_{ij} \cdot \log_2 \frac{M}{df_i}$$

Where tf_{ij} is the frequency of term i in document$j$, and df_i is the number of documents in which term i occurs Now, only weights for category vectors are to be obtained. Next we will show how to do it depending on the resource used.

3.2 Direct Approach

This approach to TC makes no use of any resource apart to the documents tobe classified It tests the intuition that the name of content-basedcategories is a good predictor for the occurrence of these categories. For instance, the occurrence of the word "barley" in adocument suggests that this one should be classified in the *barley*[2] category. All the following examples are taken from the Reuters categoryset and involve words that actually occur in the documents. category. Wehave taken exactly the categories names, although classification in moregeneral categories like *strategic-metal* should rather relay on the occurrence of more specificwords like "gold" or "zinc."

In this approach, the terms used for the representation are justthe categories themselves. The weight of term i in the vector forcategory j is 1 if $i = j$ and 0 in other cases. Multiword categories imply the use of multiwordterms. For example, the expression "balance of payments" is considered as one term. When categories consist of several synonyms(like *iron-steel*), all of them are used in the representation. Since the number ofcategories in Reuters is 135, and two of them are composite, these approachproduces 137-component vectors.

3.3 WordNet based Approach

Lexical databases contain many kinds of information (concepts; synonymy andother lexical relations; hyponymy and other conceptual relations; etc.).For instance, WordNet represents concepts as synonyms sets, or synsets. We haveselected this synonymy information, performing a "categoryexpansion" similar to query expansion in IR. For any category,the synset it belongs to is selected, and any other term belonging to it is added to therepresentation. This technique increases the amount of evidence used topredict category occurrence.

Unfortunately, the disambiguation of categories with respect toWordNet concepts is required. We have performed this task manually, becausethe small number of categories in the test collection made it affordable. We are currently designing algorithms for automating this operation.

After locating categories in WordNet, a term set containing allthe category's synonyms has been built. For the 135 categories used in thisstudy, we have produced 368 terms. Although some meaningless terms

[2] All the following examples are taken from the Reuters category set, andthey involve words that actually occur in the documents

occur and could bedeleted, we have developed no automatic criteria for this at the moment.

Let us take a look to one example. The *fuel* category hasdriven us to the addition of the terms "combustible" and "combustible material," since they belong to the same synset in WordNet. In general, the termweight vector for category k is 1 for every synonym of the category an0 for any other term.

3.4 Training Collection Approach

The key asumption when using a training collection is that a term often occurring within a category and rarely within others is a good predictorfor that category. A set of predictors is typically computed from term tocategory co-ocurrence statistics, as a training step. The computation depends on the approach and algorithmselected. As Lewis [1992] has done before, we have replicated in the VSMearly Bayesian experiments that had reported good results.

Terms are selected according to the number of times they occur withincategories. Those terms which cooccur at least with the 1% and at mostwith the 10% of the categories are taken. Among them, those 286 withhighest document frequency are selected. We work the weights out in the same way as in documents vectors:

$$wc_{ik} = tf_{ik} \cdot \log_2 \frac{L}{cf_i}$$

Where tf_{ik} is the number of times that term $i$occurs within documents assigned to category k, and $cf_i$is the number of categories within term i occurs. For example, after selecting and weighting categories, the high-frequency term "export" shows its largest weight for category *trade*, but it also shows large weights for *grain* or *wheat*, andsmall weights for *belgian-franc* and *wool*. A less frequent term typically provides evidence for asmaller number of categories. For example, "private" has a large weight only for *acq* (acquisition), and medium for *earn* (earnings) and *trade*.

3.5 Integrating WordNet and a TrainingCollection

Several ways of integrating WordNet and Reuters have occurred to us. Asensible one is to use concepts instead of terms as representatives.However, and although promising, Voorhees [1993] reported no improvements with this idea.On the other side, we have realized that the shortcomings in training canbe corrected using WordNet to provide better forecast of low frequencycategories.

In general, we have linked WordNet weight vectors to training weigth vectors. First we have removed those WordNet terms not ocurring in thetraining collection. Then we have normalized both WordNet vectors andtraining vectors to separately add up across each category. This way we have smoothed training weights (much larger than WordNetones), giving equal influence to each kind of term weight. This techniqueresults in 461 term weights vectors, 185 coming from WordNet, and 286 fromtraining. Weights for terms ocurring in both sets have been summed.Examples of terms coming from training are "import" or "government," with high weights for highly frequent categories, like *acq*. Examplesof terms coming from WordNet are "petroleum" or "peanut," with weights only for the correspondingcategories *crude* and *groundnut* respectively.

We can clearly identify the role of each resource in this TCapproach. WordNet supplies information on the semantic relatedness of termsand categories when training data is no longer available or reliable It directly contributes with part of the terms used in the vector representation. On the other side, the training collection supplies terms for those categories that are better trained The problem of unavailabilityof training data is then overcome through the use of an extern resource.

4 Evaluation

Evaluation of TC and other text classification operations exhibits greatheterogeneity. Several metrics and test collections have been used fordifferent approaches or works. This results in a lack of comparability among the approaches,forcing to replicate experiments from other researchers. Trying to minimize this problem, we havechosen a set of very extended metrics and a frequently used free testcollection for our work. The metrics are recall and precision, and the testcollection is, as introduced before, Reuters-22173. Before stepping into the actual results, we provide acloser look to these elements.

4.1 Evaluation metrics

The VSM promotes recall and precision based evaluation, but there are several ways of calculating or even defining them. Wefocus on recall, being the discussion analogous for precision. First,definition can be given regarding categories or documents [Larkey andCroft, 1996]. Second, computation can be done macro-averaging or micro-averaging [Lewis, 1992].

- Recall can be defined as the number of correctly assigned documents to a category over the number of documents to becorrectly assigned to the category. But a document-oriented definition is also possible: the number of correctly assigned categories to adocument over the number of correct categories to be assigned to thedocument. This later definition is more coherent with the task, but theformer allows to identify the most problematic categories.
- Macro-averaging consists of computing recall and precision for every item (document or category) in one of both previous ways, and averaging after it. Micro-averaging is adding up all numbers of correctly assigned items, items assigned, and items to be assigned, and calculate only one value of recall and precision. When micro-averaging, no distinction about document or category orientation can be made. Macro-averaging assigns equal weight to every category, while micro-averaging is influenced by most frequent categories.

Evaluation depends finally on the category assignment strategy: probability thresholding, k-per-doc assignment, etc. Strategies define the way to produce recall/precision tables. For instance, if similarities are normalized to the [0,1] interval, eleven levels of prob-

```
PATTERN-ID 6505 TRAINING-SET
18-JUN-1987 11:44:27.20
TOPICS:      bop trade  END-TOPICS
PLACES:      italy      END-PLACES
PEOPLE:                 END-PEOPLE
ORGS:                   END-ORGS
EXCHANGES:              END-EXCHANGES
COMPANIES:              END-COMPANIES
ITALIAN BALANCE OF PAYMENTS IN DEFICIT IN MAY
    ROME, June 18 - Italy's overall balance of payments showed
a deficit of 3,211 billion lire in May compared with a surplus
of 2,040 billion in April, provisional Bank of Italy figures
how.
    The May deficit compares with a surplus of 1,555 billion
lire in the corresponding month of 1986.
   For the first five months of 1987, the overall balance of
payments showed a surplus of 299 billion lire against a deficit
of 2,854 billion in the corresponding 1986 period.
 REUTER
```

ability threshold can be set to0.0, 0.1, and so. When the system performs k-per-doc assignment, the value of k is ranged from 1 to a reasonable maximum.

Figure 1

We must assign an unknown number of categories to each document in Reuters. So, the probability thresholding approach seems the most sensible one. We have then computed recall and precision for eleven levels of threshold, both macro and micro-averaging. When macro-averaging, we have used the category-oriented definition of recall and precision. After that, we have calculated averages of those eleven values in order to get single figures for comparison.

4.2 The Test Collection

The Reuters-22173 collection consists of 22,173 newswire articles from Reuters collected during 1987. Documents in Reuters deal with financial topics, and were classified in several sets of financial categories by personnel from Reuters Ltd. and Carnegie Group Inc. Documents vary in length and number of categories assigned, from 1 line to more than 50, and from none categories to more than 8. There are five sets of categories: TOPICS, ORGANIZATIONS, EXCHANGES, PLACES, and PEOPLE. As others before, we have selected the 135TOPICS for our experiments. An example of news article classified in *bop* (balance of payments) and *trade* is shown in Figure 1. Some spurious formatting has been removed from it.

		Subcollection		
		Training	Test	Total
Docs	Number	21,450	723	22,173
0Words	Ocurrs	2,851,455	140,922	2,992,377
	DocAvg	127	195	134
Docs with	Number	11,098	566	11,664
1+ Topics	Percent	52	78	53
Topics	Ocurrs	13,756	896	14,652
	DocAvg	0 64	1 24	0 66

Table 2. Reuters-22173 statistics

When a test collection is provided, it is customary to divide it into a training subset and a test subset. Several partitions have been suggested for Reuters [Lewis, 1992], among which ones we have opted for the most general and difficult one. First 21,450news stories are used for training, and last 723 are kept for testing. We summarize significant differences between test and training sets in Table 2. These differences can bring noise into categorization, because training relies on similarity between training and test documents. Nevertheless, this 21,450/723 partition has been used before [Lewis, 1992; Hayes and Weinstein, 1990] and involves the general case of documents with no categories assigned.

We have worked with raw data provided in the Reuters distribution. Control characters, numbers and several separators like"/" have been removed, and categories different from the TOPICS set have been ignored. For disambiguating categories with respect to WordNet senses, we first had to acquire their meaning, not always self-evident. This task has been performed by direct examination of training documents.

4.3 Results and Interpretation

The results of our first series of experiments are summarized in Table 3.This table shows recall and precision averages calculated both macro and micro-averaging for a threshold-based assignment strategy. Values for the integrated approach show some general advantage over WordNet and training approaches, but results are not decisive. Training results are comparable with those from Lewis [1992], and the WordNet approach is roughly equivalent to the training one.

Threshold	*Macro-averaging*		*Micro-averaging*	
strategy	Recall	Precision	Recall	Precision
Direct	0 239302	0 242661	0.205849	0 235775
WordNet	0 324899	0 306445	0 260762	0 298363
Training	0 325586	0 188701	0 365988	0 275731
Integrated	0 373365	0 220186	0 418652	0 296423

Table 3. Overall results from our experiments

On one hand, the integrated approach shows a better performance than the WordNet one in general, although a problem of precision is detected when macro-averaging. The influence of low precision training has produced this effect. We are planning to strengthen

WordNet influence to overcome this problem. On the other hand, the integrated approach reports better general performance than the training approach.

As expected, WordNet and training both beat the direct approach. When comparing WordNet and training approaches, we observe that the former produces better results with categories of low frequency, while the latter performs better in highly frequent categories. However, both exhibit the same overall behaviour. Differences in categories are noticed by the fact that micro-averaging is influenced by highly frequent elements, while macro-averaging depends on the results of many elements of low frequency.

5 Related Work

Text categorization has emerged as a very active field of research in the recent years. Many studies have been conducted to test the accuracy of training methods, although much less work has been developed in lexical database methods. However, lexical databases and especially WordNet have been often used for other text classification tasks, like word sense disambiguation.

Many different algorithms making use of a training collection have been used for TC, including k-nearest-neighbor algorithms [Masand *et al.*, 1992], Bayesian classifiers [Lewis, 1992], learning algorithms based in relevance feedback [Lewis *et al.*, 1996] or in decision trees [Apte *et al.*, 1994], or neural networks [Wiener *et al.*, 1995]. Apart from Lewis [1992], the closest approach to ours is the one from Larkey and Croft [1996], who combine k-nearest-neighbor, Bayesian independent and relevance feedback classifiers, showing improvements over the separated approaches. Although they do not make use of several resources, their approach tends to increase the information available to the system, in the spirit of our hypothesis.

To our knowledge, lexical databases have been used only once in TC. Hearst [1994] adapted a disambiguation algorithm by Yarowsky using WordNet to recognize category occurrences. Categories are made of WordNet terms, which is not the general case of standard or user-defined categories. It is a hard task to adapt WordNet subsets to pre-existing categories, especially when they are domain dependent. Hearst's approach shows promising results confirmed by the fact that our WordNet -based approach performs at least equally to a simple training approach.

Lexical databases have been employed recently in word sense disambiguation. For example, Agirre and Rigau [1996] make use of a semantic distance that takes into account structural factors in WordNet for achieving good results for this task. Additionally, Resnik [1995] combines the use of WordNet and a text collection for a definition of a distance for disambiguating noun groupings. Although the text collection is not a training collection (in the sense of a collection of manually labeled texts for a pre-defined text processing task), his approach can be regarded as the most similar to ours in the disambiguation task. Finally, Ng and Lee [1996] make use of several sources of information inside a training collection (neighborhood, part of speech, morphological form, etc.) to get good results in disambiguating unrestricted text.

We can see, then, that combining resources in TC is a new and promising approach supported by previous research in this and other text classification operations. With more information extracted from WordNet and better training algorithms, automatic TC integrating several resources could compete with manual indexing in quality, and beat it in cost and efficiency.

6 Conclusions and Future Work

In this paper, we have presented a multiple resource approach for TC. This approach integrates the use of a lexical database and a training collection in a vector space model for TC. The technique is based on improving the language of representation construction through the use of the lexical database, which overcomes training deficiencies. We have tested our approach against training algorithms and lexical database algorithms, reporting better results than both of these techniques. We have also acknowledged that a lexical database algorithm can rival training algorithms in real world situations.

Two main work lines are open: first, we have to conduct new series of experiments to check the lexical database and the combined approaches with other more sophisticated training approaches; second, we will extend the multiple resource technique to other text classification tasks, like text routing or relevance feedback in text retrieval.

References

[Agirre and Rigau, 1996] E. Agirre and G. Rigau. Word sensedisambiguation using conceptual distance. In *Proceedings of COLING*,1996.

[Apte *et al.*, 1994] C. Apte, F Damerau, and S.W. Weiss.Automated learning of decision rules for text categorization. *ACMTransactions in Information Systems*, Vol. 12, No. 3, 1994.

[Boguraev and Pustejovsky, 1996] B. Boguraev and J. Pustejovsky, J.(Eds.). *Corpus Processing for Lexical Acquisition*. The MIT Press,1996.

[Harman, 1996] D. Harman. Overview of the Forth Text RetrievalConference (TREC-4). In *Proceedings of the Fourth Text RetrievalConference*, 1996.

[Hayes and Weinstein, 1990] P.J. Hayes and S.P. Weinstein.CONSTRUE/TIS: a system for content-based indexing of a database of newsstories. In *Proceedings of the Second Annual Conference on InnovativeApplications of Artificial Intelligence*, 1990.

[Hearst, 1994] M. Hearst. *Context and structure in automatedfull-text information access*. Ph D. Thesis, Computer Science Division,University of California at Berkeley, 1994.

[Hersh *et al.*, 1994] W. Hersh, C. Buckley, T.J. Leone, and D.Hickman. OHSUMED: an interactive retrieval evaluation and new large testcollection for research. In *Proceedings of the ACM SIGIR*, 1994.

[Karlgren and Cutting, 1994] J. Karlgren and D. Cutting. Recogninzingtext genres with simple metrics using discriminant analysis. In *Proceedings of COLING*, 1994.

[Larkey and Croft, 1996] L.S. Larkey and W.B. Croft. Combiningclassifiers in text categorization. In *Proceedings of the ACMSIGIR*, 1996.

[Lewis *et al.*, 1996] D.D. Lewis, R.E. Schapire, J.P. Callan, andR. Papka. Training algorithms for linear text classifiers. In *Proceedings of the ACM SIGIR*, 1996.

[Lewis, 1992] D.D. Lewis. *Representation and learningin information retrieval*. Ph. D. Thesis, Dept. of Computer and InformationScience, University of Massachusetts, 1992.

[Masand *et al.*, 1992] B. Masand, G. Linoff, and D. Waltz.Classifying news stories using memory based reasoning. In *Proceedingsof the ACM SIGIR*, 1992.

[Miller, 1995] G. Miller. WordNet: a lexical database for English.*Communications of the ACM*, Vol. 38, No. 11, 1995.

[Ng and Lee, 1996] H.T. Ng and H.B. Lee. Integrating multipleknowledge sources to disambiguate word sense: an exemplar based approach.In *Proceedings of the ACL*, 1996.

[Resnik, 1995] P. Resnik. Disambiguating noun groupings with respectto WordNet senses. In *Proceedings of the Third Workshop onVery Large Corpora*, 1995.

[Salton and McGill, 1983] G. Salton and M.J. McGill. *Introductionto modern information retrieval*. McGraw-Hill, 1983.

[Salton, 1989] G. Salton. *Automating text processing: thetransformation, analysis and retrieval of information bycomputer*. Addison-Wesley, 1989.

[Voorhees, 1993] E.M. Voorhees. Using WordNet to disambiguate wordsenses for text retrieval. In *Proceedings of the ACM SIGIR*,1993.

[Wiener *et al.*, 1995] E.D. Wiener, J. Pedersen and A.S. Weigend.A neural network approach to topic spotting. In *Proceedings of theSDAIR*, 1995.

[Yokoi, 1995] T. Yokoi. The EDR electronic dictionary. *Communications of the ACM*, Vol. 38, No.11, 1995.