WordNet based Multilingual Text Categorization

BENTAALLAH MOHAMED AMINE[1]
MALKI MIMOUN[2]

EEDIS laboratory, department of computer science
Djillali Liabes university
Sidi Bel Abbes 22000, ALGERIA
[1]mabentaallah@univ-sba.dz
[2]malki_m@yahoo.com

Abstract. This article is essentially dedicated to the problem of Multilingual Text Categorization, that consists in classifying documents in different languages according to the same classification tree. The proposed approach is based on the idea to spread the utilization of WordNet in Text Categorization towards Multilingual Text Categorization. Experimental results of the bi-lingual classification of the ILO corpus (with the documents in English and Spanish) show that the idea we describe are promising and deserve further investigation.

Keywords: Multilingual, Text Categorization, WordNet, ILO corpus, Reuters-21578.

(Received February 27, 2007 / Accepted July 13, 2007)

1 Introduction

Text categorization(TC) is the task of assigning a Boolean value to each pair $< d_j, c_i > \in \mathcal{D} \times \mathcal{C}$, where $\mathcal{D}$ is the domain of documents and $\mathcal{C} = \{c_1, ..., c_{|c|}\}$ is a set of predefined categories. A value of T(True) assigned to $< d_j, c_i >$ indicates a decision to file d_j under c_i, while a value of F(False) indicates a decision not to file d_j under c_i[18].

During this last decade, research paid an important attention for the treatment of multilingual data due to the following several grounds:

- The availability of document collections distributed to the worldwide level created new needs to find information, whatever is the language and the storage support[13].

- The domination of the English language on the worldwide network is moving back to open the way toward a multilingual worldwide network [11]. Statistics showed that between 1998 and 2002, the population of which the English is the maternal language moved back from 60 to 36.5% [1].

- The time of globalization is coming. Many countries have been unified. The European project to unify European countries is a very important example in order to eliminate broader for the cooperation, global and large market, real international and free business. The high-developed technologies in network infrastructure and Internet set the platform of the cooperation and globalization. Thus, the issues of the multilinguality arise and should be addressed as soon as possible in order to overcome the remaining technical barriers that still separate countries and cultures.

The presented grounds gave birth above to a new domain of research that is the Multilingual Text Categorization. In this article, we propose a new approach for Multilingual Text Categorization that consists in spreading the use of WordNet in text categorization to categorize documents coming from different languages.

The paper is organized as follows. In section 2, a

[1]These statistics are extracted from: http://www.nua.com/surveys/

definition of Multilingual Text Categorization is presented. In Section 3, we briefly review some related work for Multilingual Text Categorization. Section 4 is dedicated to presenting WordNet. We describe the proposed approach with all its stages in section 5. In section 6, we will evaluate our approach on two different datasets. Finally, conclusion and future works are reported in section 7.

2 Multilingual Text Categorization

Multilingual Text Categorization(MTC) is a new area in text categorization in which we have to cope with two or more languages(e.g English, Spanish and Italian). In MTC, three scenarios can be distinguished:

- **Poly-lingual training:** In this scenario, the system is trained using training examples from all the different languages. A single classifier is build using a set of labelled training documents in all languages, which will classify documents from different languages. This scenario exclude the use of translation strategies, therefore, no distortion of information nor loss is made.

- **Cross-lingual training:** The system use labelled training for only one language to classify documents in other languages. This approach is what we are interested in this paper. To solve this problem, we can use the translation in different ways:

  - **Training-Set Translation:** In this approach, the labelled set is translated into the target language which then is used to train a classifier for this language. So, the Cross-lingual training became a Poly-lingual training.
  - **Test-Set Translation:** This approach consists in translating the unlabelled documents into one language (L_1). To classify the unlabelled translated documents, the system is trained using the labelled training set for language (L_1). So, the Multilingual Text Categorization became Monolingual.

- **Esperanto language:** This approach uses an universal reference language which all documents are translated to. This universal language should contain all properties of the languages of interest and be organized in a semantic way.

3 Related Work

When we embarked on this line of research, we have noticed a lack on works addressing directly the area of Multilingual Text Categorization. The majority of research works essentially comes of the Multilingual Text Retrieval. Indeed, the two areas are based on the same aspects (similarity between texts, comparison of documents with queries or class profiles).

R.Jalam et al. presented in [8] three approaches for Multilingual Text Categorization that are based on the translation of documents towards a language of reference. The authors claimed to have got good enough results.

A.Gliozzo and C.Strapparava propose in [4] a new approach to solve the Multilingual Text Categorization problem based on acquiring Multilingual Domain Models from comparable corpora to define a generalized similarity function (i.e. a kernel function) among documents in different languages, which is used inside a Support Vector Machines classification framework. The results show that the approach largely outperforms a baseline.

These last years, researches showed that using ontologies in monolingual text categorization is a promising track. J.Guyot proposed in [6] a new approach that consists in using a multilingual ontology for Information Retrieval, without using any translation. He tried only to prove the feasibility of the approach. Nevertheless, it still has some limits because the used ontology is incomplete and dirty.

Intelligent methods for enabling concept-based hierarchical Multilingual Text Categorization using neural networks are proposed in [1]. These methods are based on encapsulating the semantic knowledge of the relationship between all multilingual terms and concepts in a universal concept space and on using a hierarchical clustering algorithm to generate a set of concept-based multilingual document categories, which acts as the hierarchical backbone of a browseable multilingual document directory. The concept-based multilingual text classifier is developed using a three-layer feed-forward neural network to facilitate the concept-based Multilingual Text Categorization.

4 WordNet & Text Categorization

WordNet [10] is a lexical inheritance ontology gifted with many different pointers that aim to represent some aspects of the semantics of the lexicon, and the relationships of different lexicalized concepts. Princetons WordNet has been under construction for over a decade and several versions were proposed. The last version (WordNet 2.1) contains more than 155000 word forms organized in 117597 word meanings. The word forms in WordNet are divided by part of speech into nouns, verbs, adjectives, and adverbs. The nouns are orga-

nized as a hierarchy of nodes, where each node is a word meaning or, as it is termed in WordNet, a synset. A synset is simply a set of words that express the same meaning in at least one context. For example,{accession, addition } is a synset which express the meaning of adding to something. Table1 shows the distribution of the synsets on the four data bases (noun, verb, adjective, adverb) in WordNet 2.1 [2].

Table 1: Number of words&synsets in WordNet 2.1.

POS	Word forms	Synsets
Noun	117097	81426
Verb	11488	13650
Adjective	22141	18877
Adverb	4601	3644
Totals	155327	117597

Synsets are connected to each other through various semantic relations. The most important relations between nouns are the relations of hyponymy and hypernymy, which are transitive relations between synsets. The hypernymy relationship between synsets A and B means that B is a kind of A. Hypernymy and hyponymy are inverse relationships, so if A is a hypernym of B, then B is a hyponym of A. For example the synset {computer, computing machine, computing device, data processor, electronic computer, information processing system} is a hypernym of the synset {home computer}. Usually each synset has only one hypernym, therefore this relation organizes WordNet into a hierarchical structure. Another pair of inverse relations that hold between nouns are the meronymy and the holonymy relations. If A is a holonym of B (or in other words B is a meronym of A), it means that B is a part of A. For example, synset {keyboard} is a meronym of the synset {computer, computing machine, computing device, data processor, electronic computer, information processing system}.

The wide coverage of WordNet and its free availability has promoted its utilization for a variety of text classification tasks, including IR and TC. While WordNet usage for text classification has not proven widely effective [17, 19], some works in which WordNet synsets are used as indexing terms for IR and TC are very promising [3, 5, 14, 9].

5 Our Approach

Specially dedicated to the problem of "Cross-lingual training", the proposed approach is based on the translation of documents to be categorized towards the English language in order to be able to use the WordNet ontology thereafter. This hybridization between using machine translation and WordNet offers the following advantages:

- Without using machine translation, it becomes necessary to construct a WordNet ontology for every language. This construction is very expensive in terms of times and personals.

- The use of an ontology well constructed and rich as WordNet is going to permit to correct mistakes of the translation while using hypernymy relation and others.

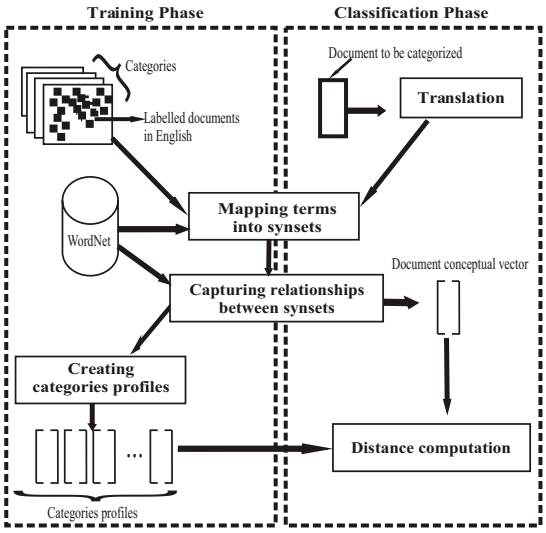

Figure 1: Architecture of the proposed approach

As indicated in Figure1, the suggested approach is composed of two phases. The first relates on the training phase that consists in creating conceptual categories profiles. These conceptual profiles will contain concepts that characterize best a category with regard to the other categories. The second phase is the classification, it consists in using machine translation techniques to translate documents into English language in order to generate its conceptual vector using WordNet. After this, we have to weight document vector and categories profiles, then to calculate distances between them in order to be able to decide the adherence of the document to a category or an another one.

[2]These statistics are extracted from: http://wordnet.princeton.edu/

5.1 Training phase

The first issue that needs to be addressed in this phase is: "how to represent texts so as to facilitate machine manipulation but also to retain as much information as needed ?". The commonly used text representation is the Bag-Of-Words, which simply uses a set of words with their number of occurrences to represent documents and categories [12]. This representation was disadvantaged by the ignorance of any relation between words, thus learning algorithms are restricted to detect patterns in the used terminology only, while conceptual patterns remain ignored. In our approach, the training phase consists of using WordNet to create profiles categories which will contain concepts (synsets in WordNet) that characterize best one category with regard to the other categories. For this purpose, three steps are required:

- Mapping terms into synsets using WordNet;
- Capturing relationships between synsets;
- Using features selection method to select the characteristic concepts that will form the conceptual categories profiles.

5.1.1 Mapping terms into synsets

The most straightforward representation of documents relies on term vectors. The major drawback of this basic approach for document representation is the size of the feature vectors, usually more than 10,000 terms. In the application of text categorization, however, completely different terms may represent the same concepts. In some cases, terms with different concepts can even be replaced with only one higher level concept without negative effect on performance of the classifier. Obviously, mapping terms to concepts is an effective and reasonable method to reduce the dimensionality of the vector space. In the most case, one word may have several meanings and thus one word may be mapped into several synsets which may add noise to the representation and may induce a loss of information. In this case, we need to determine which meaning is being used, which is the problem of sense disambiguation [7]. For this purpose, WordNet returns an ordered list of synsets for each term. Thereby, the ordering is supposed to reflect how common it is that a term reflects a synsets in "standard" English language. More common term meanings are listed before less common ones.

While there is a whole field of research dedicated to word sense disambiguation, it has not been our intention to determine which one could be the most appropriate, but simply whether word sense disambiguation is needed at all. In our approach, we used a simple disambiguation strategy that consists of considering only the most common meaning of the term (first ranked element) as the most appropriate. So our mapping process consists in replacing each term by its most common meaning. Thus the synset frequency is calculated as indicated in the following equation:

$$sf(c_i, s) = tf(c_i, \{t \in T \mid first(Ref_s(t)) = s\}) \quad (1)$$

where:

- c_i: the i^{th} category.
- $tf(c_i, T')$: the sum of the frequencies of all terms $t \in T'$ in the train documents of category c_i.
- $Ref_s(t)$: the set of all synsets assigned to term t in WordNet.

5.1.2 Capturing relationships between synsets

After mapping terms into synsets, this step consists in using the WordNet hierarchies to capture some useful relationships between synsets (hypernymy in our case). The synset frequencies will be updated as indicated in the following equation:

$$sf(c_i, s) = \sum_{b \in H(s)} sf(c_i, b) \quad (2)$$

Where:

- c_i: the i^{th} category.
- b and s are synsets.
- $H(s)$ contains the synsets that have the synset s as hypernym.

5.1.3 Creating conceptual categories profiles

Selection methods for dimensionality reduction take as input a set of features and output a subset of these features, which are relevant for discriminating among categories [2]. Controlling the dimensionality of the vector space is essential for two reasons. The complexity of many learning algorithms depends crucially not only on the number of training examples but also on the number of features. Thus, reducing the number of features may be necessary to make these algorithms tractable. Also, although more features can be assumed to carry more information and should, thus, lead to more accurate classifiers, a larger number of features with possibly many of them being irrelevant may actually hinder a learning algorithm constructing a classifier. For our approach, a feature selection technique is necessary in order to reduce the big dimensionality by creating the

conceptual categories profiles. For this purpose we used the χ_2 multivariate statistic for feature selection.

The χ_2 multivariate [21], noted $\chi_2^{multivariate}$ is a supervised method allowing the selection of features by taking into account not only their frequencies in each category but also interaction of features between them and interactions between features and categories. In our case, it consists of extracting, for each category, the K better synsets (our features) characterizing best the category compared to the others. With this intention, the matrix (synsets-categories) representing the total number of occurrences of the p synsets in the m categories is calculated. The total sum of the occurrences is noted N. The values N_{jk} represent the frequency of the synset s_j in the category c_k. Then, contributions of these synsets in discriminating categories are calculated as indicated in equation(3), then sorted by descending order for each category.

$$\mathcal{C}_{jk}^{\chi 2} = N\frac{(f_{jk}-f_{j.}f_{.k})^2}{f_{j.}f_{.k}} \times sign(f_{jk} - f_{j.}f_{.k}) \quad (3)$$

Where: $f_{jk} = \frac{N_{jk}}{N}$ representing the relative frequencies of the occurrences.

The evaluation of the sign in the equation (3) makes it possible to determine the direction of the contribution of the synset in discriminating the category. A positive value indicates that it is the presence of the synset which contribute in the discrimination while a negative value reveals that it is its absence which contribute in it.

5.2 Classification Phase

The classification phase consists on using the conceptual categories profiles in classifying unlabelled documents in different languages. Our classification phase consists of:

- Translating the document to be categorized and generating a conceptual vector;
- Weighting the conceptual categories profiles and the conceptual vector of the unlabelled document;
- Calculating distance between the conceptual vector of the document and all conceptual categories profiles.

5.2.1 Translation and generation of the conceptual vector

The translation of the text to be classified in the language of training corpus is also a paramount stage. The objective here is not to produce a translated text accurately recalling the semantic properties of the original text, but to provide a text ensuring a sufficient quality of classification. It is obvious that the obtained result will depend on the used translator. For that, we used JWT [3](Java Web Translator) library which provides automatic language translation for 14 languages including English, Spanish, French, Italian, Deutsch, Greek, Chinese, Japanese, Russian, ect.

After translating document, we have to use WordNet in order to generate a conceptual vector for the document (mapping terms into synsets and capturing relationships between synsets).

5.2.2 Weighting

This stage consists of weighting conceptual categories profiles and conceptual vector of the unlabelled document. Each weight $w(s, c)$ expresses the importance of synset s in vector of c with respect to its frequency in all training documents. The objective of using a feature weight rather than plain frequencies is to enhance classification effectiveness. In our experiments, we used the standard $tfidf$ (term frequency - inverse document frequency) function [16], defined as:

$$w(s_k, c_i) = tfidf(s_k, c_i) = tf(s_k, c_i) \times \log(\frac{|C|}{df(s_k)}) \quad (4)$$

Where:

- $tf(s_k, c_i)$ denotes the number of times synset s_k occurs in category c_i.
- $df(s_k)$ denotes the number of categories in which synset s_k occurs.
- $\mid C \mid$ denotes the number of categories.

5.2.3 Distance computation

The similarity measure is used to determine the degree of resemblance between two vectors. To achieve reasonable classification results, a similarity measure should generally respond with larger values to documents that belong to the same class and with smaller values otherwise. The dominant similarity measure in information retrieval and text classification is the cosine similarity between two vectors. Geometrically, it evaluates the cosine of the angle between two vectors $d1$ and $d2$ and is, thus, based on angular distance [15]. This allows us to abstract from varying vector length. The cosine similarity can be calculated as the normalized dot product:

$$\mathcal{S}_{i,j} = \frac{\sum_{s \in i \frown j} tfidf(s,i) \times tfidf(s,j)}{\sqrt{\sum_{s \in i} tfidf^2(s,i) \times \sum_{s \in j} tfidf^2(s,j)}} \quad (5)$$

[3]This package is available on: http://sourceforge.net/projects/webtranslator

With:
s: a synset,
i and j: the two vectors (profiles) to be compared.
$tfidf(s,i)$: the weight of the synset s in i.
$tfidf(s,j)$: the weight of the synset s in j.

Which can be translated in the following way: "If the two vectors are very alike their corresponding angle should be very small and approaching zero (cosine value approaching 1). On the other hand, if the angle is high, let say, 90 degrees, the vectors would be perpendicular(orthogonal) and the cosine value would be 0. In such case, the two vectors are not related". In our approach, this similarity measure is used to calculate distances between the conceptual vector of the unlabelled document and all categories profiles. As a result, the document will be assigned to the category whose profile is the closest with the document vector.

6 Experimental Results

6.1 Datasets for evaluation

6.1.1 Monolingual dataset

The Reuters dataset has been used in many text categorization experiments; the data was collected by the Carnegie group from the Reuters newswires in 1987. There are now at least five versions of the Reuters datasets widely used in TC community.

Table 2: The 10 used categories of Reuters-21578 corpus

Category	♯ Training	♯ Test
Earn	2877	1087
Acquisition	1650	719
Money-fx	538	179
Grain	433	149
Crude	389	189
Trade	369	118
Interest	347	131
Wheat	212	71
Ship	197	89
Corn	182	56
Total	7194	2788

In our experiments, we used the 10 most frequent categories from the Modapte [4] version as our monolingual dataset for training and testing (as shown in Table2).

6.1.2 Multilingual dataset

The ILO corpus is a collection of full-text documents, each labelled with one category (mono-classification) which can be downloaded from the ILOLEX website of the International Labour Organisation[5]. ILOLEX describes itself as a trilingual database containing ILO Conventions and Recommendations, ratification information, comments of the Committee of Experts and the Committee on Freedom of Association, representations, complaints, interpretations, General Surveys, and numerous related documents. The languages concerned are English, Spanish and French. In our experiments, we used a bilingual version (with documents in English and Spanish) mono-classified in 10 categories with a rather varying number of documents per category as shown in Table3.

Table 3: The 10 used categories of the ILO corpus

Category	♯ English	♯ Spanish
Special prov. by Sector of Econ. Act.	108	121
Conditions of employment	397	86
Conditions of work	299	71
Economic and social development	22	23
Employment	410	448
Labour Relations	276	278
Labour Administration	85	81
Health and Labour	98	86
Social Security	150	148
Training	79	20
Total:	1924	1362

6.2 Evaluation method

Experimental results reported in this section are based on the so-called "F_1 measure", which is the harmonic mean of precision and recall.

$$F_1(i) = \frac{2 \times precision \times recall}{precision + recall} \quad (6)$$

In the above formula, precision and recall are two standard measures widely used in text categorization literature to evaluate the algorithms effectiveness on a given category [20]. We also use the macroaveraged F_1 to evaluate the overall performance of our approach on given datasets. The macroaveraged F_1 computes the F_1 values for each category and then takes the average over the per-category F_1 scores. Given a training dataset with m categories, assuming the F_1 value for the $i-th$ category is $F_1(i)$, the macroaveraged F_1 is defined as :

$$MacroAveragedF_1 = \frac{\sum_{i=1}^{m} F_1(i)}{m} \quad (7)$$

[4] http://www.daviddlewis.com/ressources/testcollections

[5] http://ilolex.ilo.ch:1567/Spanish/index.htm

6.3 Results

In order to be able to show the utility of the use of WordNet in Multilingual Text Categorization, we tested the suggested approach on both monolingual and multilingual datasets. It is necessary to specify here, that our objective is not to compare the two used datasets but to show if it is possible to spread the use of WordNet in Multilingual Text Categorization with using machine translation techniques.

Table 4: Comparison of MacroAveraged F_1 results on the two used datasets

Size of profiles	Monolingual dataset	Multilingual dataset
k=100	0.664	0.416
k=200	0.663	0.455
k=300	0.663	0.504
k=400	0.666	0.534
k=500	0.667	0.554
k=600	0.668	0.560
k=700	0.667	0.561
k=800	0.668	0.573
k=900	0.668	0.573

The results of the experimentations are presented in Table4. Concerning the profiles size, it is noted that for the two datasets, the best performances are obtained with size profile $k = 800$. Indeed, the performances improve more and more by increasing the size of profiles.

In addition, it is noticed that the performances of multilingual text categorization are very close to those of monolingual text categorization. Indeed, the differences of the error rates obtained in multilingual categorization (after translation) compared with those obtained in monolingual categorization are not significant.

7 Conclusion

In this paper, we proposed a new approach for Multilingual Text Categorization which is based on the one hand on the use of WordNet and on the other hand on the use of machine translation techniques. The obtained results are encouraging and carry out us to confirm that the use of WordNet in Multilingual Text Categorization is a promising track.

Our future works will concern the use of WordNet in distance computation in order to be able to test the use of the semantic distances instead of the statistics distances. Another track consists in using other disambiguation, selection and weighting techniques.

References

[1] Chau, R., Yeh, C.-H., and Smith, K. A. A Neural Network Model for Hierarchical Multilingual Text Categorization. *Proceeding of ISSN-05 Second International Symposium on Neural Networks, Chongqing, China*, pages 238–245, 2005.

[2] Dash, M. and Liu, H. Feature Selection for Classification. *Intelligent Data Analysis*, 1(3):131–156, 1997.

[3] Fukumoto, F. and Suzuki, Y. Learning lexical representation for text categorization. *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*, 2001.

[4] Gliozzo, A. and Strapparava, C. Cross Language Text Categorization by acquiring Multilingual Domain Models from Comparable Corpora. *Proceedings of the ACL Workshop on Building and Using Parallel Texts. Ann Arbor, Michigan, USA*, pages 9–16, 2005.

[5] Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J. Indexing with WordNet synsets can improve text retrieval. *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, 1998.

[6] Guyot, J., Radhouani, S., and Falquet, G. Ontology-based multilingual information retrieval. *In CLEF Workhop, Working Notes in Multilingual Textual Document Retrieval Track. Vienna, Austria*, 2005.

[7] Ide, N. and Veronis, J. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):2–40, 1998.

[8] Jalam, R., Clesh, J., and Rakotomalala, R. Cadre pour la catégorisation de textes multilingues. *7èmes Journées Internationales d'Analyse Statistique des Données Textuelles, Louvain-la-Neuve, Belgique*, pages 650–660, 2004.

[9] Mihalcea, R. and Moldovan, D. Semantic indexing using WordNet senses. *Proceedings of ACL Workshop on IR and NLP*, 2000.

[10] Miller, G. A. WordNet: An On-Line Lexical Database. *In Special Issue of International Journal of Lexicography, Chongqing, China*, 3(4), 1990.

[11] Nunberg, G. Will the Internet Always Speak English? *The American Prospect*, 11(10), 2000.

[12] Peng, X. and Choi, B. Document Classifications Based On Word Semantic Hierarchies. *Proceedings of the IASTED International Conference on Artificial Intelligence and Applications*, pages 362–367, 2005.

[13] Peters, C. and Sheridan, P. Accès multilingue aux systèmes d'information. *In 67th IFLA Council and General Conference*, 2001.

[14] Petridis, V., Kaburlasos, V., Fragkou, P., and Kehagias, A. Text classification using the sigma-FLNMAP neural network. *Proceedings of the 2001 International Joint Conference on Neural Networks*, 2001.

[15] Salton, G. Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer. *Addison-Wesley*, 1989.

[16] Salton, G. and Buckley, C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

[17] Scott, S. Feature engineering for a symbolic approach to text classification. *Masters thesis. Computer Science Dept, University of Ottawa, Ottawa, CA*, 1998.

[18] Sebastiani, F. Machine learning in automated text categorization. *ACM Computing Surveys*, pages 1–47, 2002.

[19] Voorhees, E. Using WordNet for text retrieval. *In: WordNet: An Electronic Lexical Database. MIT Press*, 1998.

[20] Yang, Y. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, 1(1-2):69–90, 1999.

[21] Yang, Y. and Pedersen, J. O. A comparative study on feature selection in text categorization. *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 412–420, 1997.