The Use of WordNets for Multilingual Text Categorization: A Comparative Study

Mohamed Amine Bentaallah and Mimoun Malki

EEDIS Laboratory, Department of computer sciences
Djillali Liabes University
Sidi Bel Abbes, 22000. ALGERIA
mabentaallah@univ-sba.dz
malki-m@yahoo.com
http://www.univ-sba.dz

Abstract. The successful use of the Princeton WordNet for Text Categorization has prompted the creation of similar WordNets in other languages as well. This paper focuses on a comparative study between two WordNet based approaches for Multilingual Text Categorization. The first relates on using machine translation to access directly the princeton WordNet while the second avoids machine translation by using the WordNet associated for each language.

Key words: Multilingual, Text Categorization, WordNet, Ontology, Mapping

1 Introduction

With the rapid emergence and proliferation of Internet and the trend of globalization, a tremendous number of textual documents written in different languages are electronically accessible online. Efficiently and effectively managing these textual documents written in different languages is essential to organizations and individuals. This necessity gave birth to a new domain of research that is the Multilingual Text Categorization.

The growing popularity of the Princeton WordNet as a useful resource for English and its incorporation in natural language tasks has prompted the creation of similar WordNets in other languages as well. Indeed, WordNets for more than 50 languages are currently registered with the Global WordNet Association [1]. In this paper we try to answer the question: "***Will the use of these WordNets in Text Categorization guarantee good results better than those obtained by the Princeton WordNet ?***".

The rest of the paper is organized as follows. In section 2, we review some related works for Multilingual Text Categorization. In section 3, we describe the two approaches to be compared. Section4 presents the experiments and the results. Finally, conclusion and future works are reported in section 5.

[1] http://www.globalwordnet.org

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).

MTC is a relatively new research topic, about which not much previous work in the literature appears to be available. Most approaches have mainly addressed different translation issues to solve the problem. R.Jalam et al. presented in [1] three approaches for MTC that are based on the translation of documents toward a language of reference. Rigutini et al. used in [2] a machine translation system to bridge the gap between different languages. The major disadvantage of Machine translation based approaches is the absence of machine translation systems for many language pairs and the wide gap between the translated documents and original documents.

In order to overcome the disadvantage of using machine translation systems, many researches have been working on using linguistic resources such as bilingual dictionaries and comparable corpora to induce correspondences between two languages. 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. K.Wu et al. proposed in [3] a novel refinement framework for cross-language text categorization investigating the use of a bilingual lexicon to identify a novel model called domain alignment translation model. Their approach can achieve comparable performance with the machine translation approach using the Google translation tool, although their experiments only consider the word level but ignore the base phrase.

These last years, researches showed that using ontologies in monolingual text categorization is a promising track. J.Guyot proposed in [9] 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 [13]. 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. We have proposed in [10] a new approach for MTC based on spreading the use of WordNet in Text Categorization towards MTC in order to reduce noises introduced by machine translation.

3 Description of the two proposed approaches

As shown in figure 1, the two approaches are composed of three phases:

– Knowledge representation step;
– Training step;
– Predicting step.

For our experiments, the two approaches have the same training and prediction phases. The only difference is on the knowledge representation phase.

3.1 Knowledge representation

First approach The first approach consist on representing knowledge with the use of the Princeton WordNet. The labelled documents are mapped directly into the synsets of the princeton WordNet since they are expressed in English language. The unlabelled documents needs to be translated into the English language in order to be able to be mapped to the Princeton WordNet. The mapping into the princeton WordNet consists in replacing each term in a document by its most common meaning from the Princeton WordNet. 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. Thus the synset frequency is calculated as indicated in the following equation:

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

where:

– $tf(c_i, T')$: the sum of the frequencies of all terms $t \in T'$ in the train documents of category c_i.
– $Ref(t)$: the set of all synsets assigned to term t in WordNet.

Second approach The second approach excludes the direct use of machine translation techniques by incorporating the WordNet associated for document languages. Indeed, each term document will be firstly mapped to the WordNet synsets of the language in which the document is expressed. As result, the labelled documents and the unlabelled documents will be mapped on different taxonomies. The labelled documents will be mapped to the Princeton WordNet, and the unlabelled documents will be mapped to the WordNets associated to unlabelled documents languages. It is necessary to match the taxonomies of all the used WordNets to a common taxonomy in order to unify document representations. Since the Princeton WordNet is the richest taxonomy, we have chosen it to be the common taxonomy. This matching offers the following advantages:

– Avoiding the direct use of machine translation techniques which eliminate the problem of translation disambiguation.
– Interconnecting the different WordNets to the most rich WordNet (Princeton WordNet) which resolves the richness of some WordNets.

Formally,the synset frequency is calculated as indicated in the following equation:

$$sf(d, s) = tf(d, \{t \in T \mid match(first(Ref(t, L))) = s\}) \quad (2)$$

where:

– $tf(d, T')$: the sum of the frequencies of all terms $t \in T'$ in the unlabelled document d .
– L: The language of the unlabelled document d.
– $Ref(t, L)$: the set of all synsets assigned to term t in WordNet associated to language L.
– $match(s)$: the corresponding synset of the synset s on the Princeton WordNet.

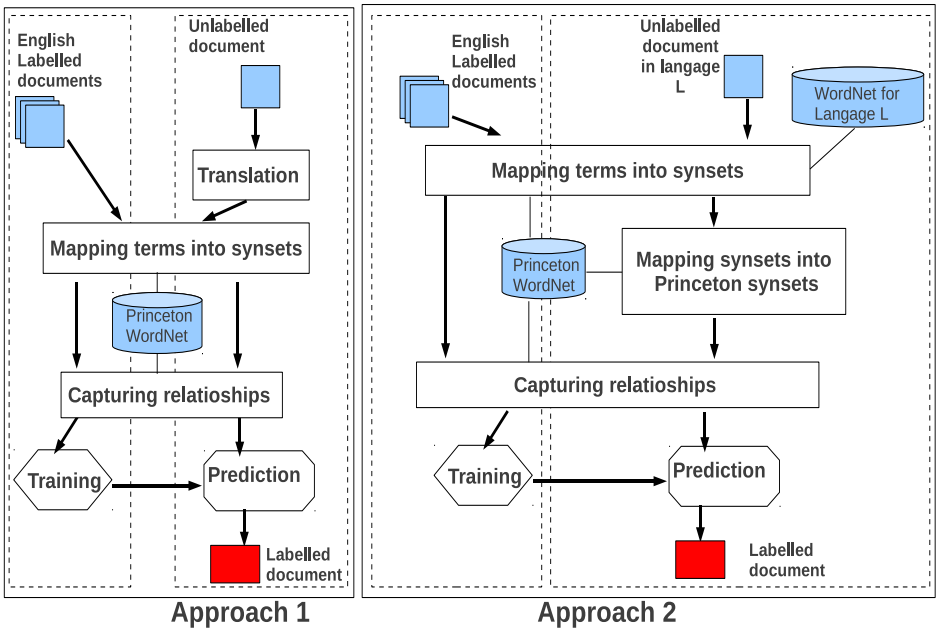

Fig. 1. The two compared approaches

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

$$sf(c_i, s) = \sum_{b \in H(s)} sf(c_i, b) \tag{3}$$

Where:

– c_i: the i^{th} category.
– b and s are synsets.
– $H(s)$: the hyponyms set of synset s

3.2 Training

The training phase consists in using the labelled documents to create conceptual categories profiles. Formally, each category will be represented by a conceptual profile which contains the K better synsets (our features) characterizing best the category compared to the others. For this purpose we used the χ_2 multivariate statistic for feature selection. The χ_2 multivariate [24], 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. Given the matrix (synsets-categories) representing the total number of occurrences of the p synsets in the m categories. The contributions of these synsets in discriminating categories are calculated as indicated in the following equation, then sorted by descending order for each category.

$$\mathcal{C}_{jk}^{\chi 2} = N\frac{(f_{jk}-f_{j.}f_{.k})^2}{f_{j.}f_{.k}} \times sign(f_{jk} - f_{j.}f_{.k}) \tag{4}$$

Where:

- $f_{jk} = \frac{N_{jk}}{N}$: the relative occurrence frequency.
- N: The total sum of the occurrences.
- N_{jk}: The frequency of the synset s_j in the category c_k.

Once the contributions of synsets are calculated and ordered for each category, the conceptual profile of each category contains the k first sorted synsets.

3.3 Prediction

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

- Weighting the conceptual categories profiles and the conceptual vector of the unlabelled document. In our experiments, we used the standard $tfidf$ (term frequency - inverse document frequency) function [25], defined as:

$$w(s_k, c_i) = tfidf(s_k, c_i) = tf(s_k, c_i) \times \log(\frac{|C|}{df(s_k)}) \tag{5}$$

  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.
- Calculating distances between the conceptual vector of the document and all conceptual categories profiles and assigning the document to the category whose profile is the closest with the document vector. In our experiments, we used the dominant similarity measure in information retrieval and text classification which is the cosine similarity that can be calculated as the normalized dot product:

$$\mathcal{S}_{i,j} = \frac{\sum_{s \in i \frown j} tfidf(s,i) \times tfidf(s,j)}{\sqrt{\sum_{s \in i} tfidf^2(s,i) \times \sum_{s \in j} tfidf^2(s,j)}} \tag{6}$$

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.

4 Experimental results

4.1 Dataset for evaluation

For our experimentations, we extracted a bilingual dataset from Reuters Corpus Vol. 1 and 2 (RCV1, RCV2) using English training (RCV1) and Spanish test documents (RCV2). Our dataset is based on topic (category) codes with a rather varying number of documents per category as shown in Table1

Table 1. The 8 used Categories of the Multilingual Reuters corpus

Code category	Category Description	English labelled documents	Spanish unlabelled documents
C183	Privatisations	200	205
GSPO	Sport	401	84
GDIS	Disaster	278	116
GJOB	labour issues	401	197
GDEF	Defence	227	83
GCRIM	Crime, Law enforcement	401	157
GDIP	International relations	401	237
GVIO	War, Civil war	401	306

4.2 Results

For comparison, we have tested the two approaches on our multilingual dataset. Experimental results reported in this section are based on the so-called "F_1 measure", which is the harmonic mean of precision and recall.

$$F_1(i) = \frac{2 \times precision \times recall}{precision + recall} \tag{7}$$

The results of the experimentations are presented in Table2, Concerning the profiles size, the best performances are obtained with size profile k = 900 for the two approaches. Indeed, the performances improve more and more by increasing the size of profiles.

Comparing the results of the two approaches, the first approach largely outperform the second approach.

Table 2. Comparison of F-score results on the two approaches

Size of profiles	Approach1	Approach2
k=100	0.586	0.201
k=200	0.608	0.213
k=400	0.621	0.219
k=500	0.509	0.222
k=700	0.634	0.221
k=900	0.639	0.268

5 Conclusion

In this paper, we have compared two approches for using WordNets for MTC, The first approach is based on using machine translation to use the Princeton WordNet while the second approach is based on replacing the use of machine translation by incorporating a WordNet for each language. The results of the experimentations show that the use of WordNets does not guarantee good results rather than those obtained by the Princeton WordNet. Future works will concern the experimentation of the second approach with differents WordNets in order to be able to confirm the obtained results.

References

1. Jalam, R., Clesh, J., Rakotomalala, R.: Cadre pour la catégorisation de textes multilingues.7 èmes Journées internationales d'Analyse statistique des Données Textuelles.Louvain-la-Neuve, Belgique (2004) 650–660
2. Rigutini, L., Maggini, M., and Liu, B.: An EM based training algorithm for Cross-Language Text Categorization. Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence. Compiegne, France. September 2005.
3. Wu, K., Lu,B.: A Refinement Framework for Cross Language Text Categorization. 4th Asia Infomation Retrieval Symposium, AIRS 2008, Harbin, China, (2008) 401–411
4. Gliozzo, A.M., Strapparava, C.: Cross Language Text Categorization by acquiring Multilingual Domain Models from Comparable Corpora.in Proceedings of the ACL Workshop on Building and Using Parallel Texts. Ann Arbor, Michigan, USA (2005) 9–16
5. Adeva, J. J., Calvo, R. A., and Ipia, D.: Multilingual Approaches to Text Categorisation. The European Journal for the Informatics Professional,Vol 6, (2005) 43-51
6. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys, (2002) 1–47
7. Peters, C., Sheridan, P.: Acces multilingue aux systèmes d'information. In: In 67th IFLA Council and General Conference. (2001)
8. Nunberg, G.: Will the Internet speak english?. The American Prospect. (2000)
9. Guyot, J., Radhouani, S., Falquet, G.: Ontology-based multilingual information retrieval. In CLEF Workhop, Working Notes Multilingual Track, Vienna, Austria (2005) 21–23

10. Bentaallah, M.A., Malki, M.: WordNet based Multilingual Text Categorization. Journal of Computer Science, Vol 6 (2007)
11. Hu, W., Jian,N., Qu, Y., Wang, Y.: Gmo: A graph matching for ontologies. In Proceedings of the K-CAP 2005Workshop on Integrating Ontologies, (2005) 41–48
12. Lacher, M.S., Groh.G.: Facilitating the exchange of explicit knowledge through ontology mappings. In Proceedings of the 14th International Florida Artificial Intelligence Research Society Conference (FLAIRS01), AAAI Press, (2001) 305–309
13. Chau, R., Yeh, C.H, Smith, K.:A Neural Network Model for Hierarchical Multilingual Text Categorization. In proceeding of ISSN-05 Second International Symposium on Neural Networks, Chongqing, China (2005) 238–245
14. Chau, R., Yeh, C.: Multilingual Text Categorization for Global Knowledge Discovery Using Fuzzy Techniques. Proceedings of the 2002 IEEE International Conference on Artificial Intelligence Systems(ICAIS), (2002) 82-86
15. Ichise, R., Hamasaki, M., Takeda, H.: Discovering relationships among catalogs. In E. Suzuki and S. Arikawa, editors, Proceedings of the 7th International Conference on Discovery Science (DS04), volume 3245 of LNCS, Springer, (2004) 371–379
16. Nottelmann, H., Straccia, U.: A probabilistic, logic-based framework for automated web directory alignment. In Zongmin Ma, editor, Soft Computing in Ontologies and the Semantic Web, Studies in Fuzziness and Soft Computing, Springer Verlag, (2006) 47–77
17. Miller, G.A.:WordNet: An On-Line Lexical Database. In Special Issue of International Journal of Lexicography, Vol 3, No. 4 (1990) 238–245
18. Furst, F., Trichet, F.: Axiom-based ontology matching. In Proceedings of the 3rd international conference on Knowledge capture (K-CAP 05), ACM Press, (2005) 195–196
19. Do, H.H., Rahm, E.: Coma - a system for flexible combination of schema matching approaches. In Proceedings of the 28th International Conference on Very Large Data Bases (VLDB 02), (2002) 610–621
20. Gonzalo, J., Verdejo, F., Chugur, I., Cigarran, J.: Indexing with WordNet synsets can improve text retrieval. In: Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems. (1998).
21. Ide, N., Veronis, J.: Introduction to the special issue on word sense disambiguation: The state of the art. Computational Linguistics. (1998).
22. Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. The VLDB Journal, 10(4), (2001) 334–350
23. Shvaiko, P., Euzenat, J.: A survey of schema-based matching approaches. Journal on Data Semantics, 4(LNCS 3730), (2005) 146–171
24. Yang, Y., Pederson, J.O.: A comparative study on feature selection in text categorization. In Proceedings of ICML-97, 14th International Conference on Machine Learning. (1997) 412–420
25. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing and Management. (1988) 513–523
26. Salton, G.: Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer.Addison-Wesley (1989)