Multilingual Text Classification using Ontologies

Gerard de Melo, Stefan Siersdorfer
{demelo,stesi}@mpi-inf.mpg.de

Max Planck Institute for Computer Science, Saarbrücken, Germany

Abstract. In this paper, we investigate strategies for automatically classifying documents in different languages thematically, geographically or according to other criteria. A novel linguistically motivated text representation scheme is presented that can be used with machine learning algorithms in order to learn classifications from pre-classified examples and then automatically classify documents that might be provided in entirely different languages. Our approach makes use of ontologies and lexical resources but goes beyond a simple mapping from terms to concepts by fully exploiting the external knowledge manifested in such resources and mapping to entire regions of concepts. For this, a graph traversal algorithm is used to explore related concepts that might be relevant. Extensive testing has shown that our methods lead to significant improvements compared to existing approaches.

1 Introduction

Text classification (TC) is the process of associating text documents with the classes considered most appropriate, thereby distinguishing topics such as particle physics from optical physics. Research in this area, despite the considerable amount of work on cross-lingual information retrieval, has almost entirely neglected cases of documents being provided in multiple languages. Apart from truly multilingual environments as in large parts of Africa, people all over the world work with a lingua franca such as English or Spanish in addition to their native languages. Therefore, most applications of TC, e.g. digital libraries, news wire filtering as well as web page and e-mail categorization, also turn out to be interesting applications of *multilingual text classification* (MLTC), where documents given in different languages are to be classified by topic or similar criteria.

In this paper, we provide linguistic arguments against existing approaches and devise a novel solution that exploits background knowledge from ontologies and lexical resources. Section 2 discusses related work in this area, followed by Section 3, which briefly recapitulates fundamental ideas in TC and delivers arguments against existing approaches. Section 4 then presents Ontology Region Mapping as an alternative, which is then evaluated in Section 5, while the concluding section outlines the implications for continued research in this area.

2 Related Work

There has been research on MLTC in the case of enough training documents being available for every language [1, 2], however such scenarios are not particularly interesting as they can be resolved with separate monolingual solutions. A more universal strategy is to use translation to ensure that all documents are available in a single language [3–5], which also corresponds to the dominant approach in *cross-lingual information retrieval* (CLIR) [6]. However, our work shows that translations alone entail suboptimal TC results. An alternative approach to monolingual TC [7–9] and CLIR [10] related to the path pursued in our work relies on ontologies or thesauri to construct concept-based representations. While some authors pay respect to hypernyms and other directly related concepts [11–13], our approach is apparently the first to use an activation spread model in TC or CLIR. Multilingual solutions based on *latent semantic analysis* (LSA) have also been studied [14, 15], however LSA differs from our approach in that it does not use formal background knowledge, but rather identifies concepts implicitly present in a set of documents, computed statistically by detecting terms with similar occurrence patterns.

3 Background

A *classification* is an assignment of class labels to objects such as text documents, and automatic text classification is the process of establishing and deploying classifiers that approximate text classifications made by human experts. When a set of pre-classified training documents is available, and an appropriate representation of their contents as numerical *feature vectors* is constructed, one of several learning algorithms can be employed to learn a classification, e.g. the *Support Vector Machine* (SVM) [16], which distinguishes two classes by using the hyperplane that maximizes the distance to the closest positive and negative examples as a binary decision surface. The conventional way of establishing the vector space for text documents involves well-known techniques such as stopword removal and stemming as preprocessing steps to computing TF-IDF values that are used to construct feature vectors based on the bag-of-words model [17].

Machine translation has been proposed as an ad hoc means of making multilingual document sets amenable to such TC processing [3]. However, certain drawbacks of the bag-of-words model then become particularly severe, e.g. when Spanish '*coche*' is generally mapped to '*car*', whereas French '*voiture*' is translated to '*automobile*', the learning algorithm remains unaware of the synonymy. Consider also that AltaVista Babel Fish [18] translates Spanish '*Me duele la cabeza*' to '*It hurts the head to me*', which does not contain the word '*headache*'.

Simple *concept mappings* have been used for alternative text representations in monolingual TC. Rather than using the original terms, one considers the *concepts* associated with their meanings as e.g. captured by WordNet [19]. Although the idea of mapping from several languages to language-neutral concepts seems particularly attractive, it may have a detrimental effect on the efficiency. For instance, lemmatizing inflected words to their base forms means that the distinct

base forms of '*protected*' and '*protection*' prevent the two from being identified. Furthermore, WordNet lists many senses of the word '*school*', of which, in TC, at least seven should be seen as a thematic cluster rather than being distinguished, including school as an educational institution, as a building, as the process of being educated, etc. The idiosyncrasies of different languages pose additional problems, e.g. the English term '*woods*' is much narrower than the French '*bois*', so the two might not be mapped to the same concept. In Japanese and Chinese, there are separate words for older and younger sisters. German (as well as several other languages) allows for almost arbitrary compounds such as '*Friedensnobelpreisträgerin*' (woman awarded the Nobel Peace Prize).

4   Ontology Region Mapping

As all of the problems mentioned above involve terms being treated as distinct despite being closely related, we present *Ontology Region Mapping* (ORM) as a novel approach, where ontologies are construed as semantic networks in which entire regions of related concepts are considered relevant, rather than just individual ones. Our approach first maps terms to the concepts they are immediately associated with and then explores further related concepts.

4.1   Ontologies and Ontology Mapping Functions

An *ontology* is a theory of what possesses being in the world or in a domain. For our purposes, the requirements are a set of concept identifiers (*concept terms*) and a function τ providing information about how they are connected. For a concept term c, $\tau(c)$ should deliver a finite set of entries (c_i, r_i, w_i), where r_i indicates the type of relation (hypernymy, antonymy, etc.), c_i is a concept term, and $w_i \in [0, 1]$ is a relation weight specifying to what degree c and c_i are related.

Additionally, we construct *ontology mapping functions* that map document terms t from languages such as Spanish to such concepts, returning a set of pairs (c, w) where c is a concept term that possibly represents t's meaning and w is c's weight, i.e. the degree of relevance of c estimated with respect to the local context in which t occurred (the words surrounding t in the document). In our implementation, the functions look up terms in the English and Spanish WordNet [20], which serve as our ontological resources, using the lemmatized base form when no entry exists for the inflected form. In order to determine to what degree the concepts listed in WordNet for a term t are relevant in a particular context, part-of-speech information determined via morphological analysis is used to eliminate certain candidates. The remaining ones are then distinguished using an existing word sense disambiguation technique [21], where additional context strings are constructed for the candidate concepts by concatenating their human language description provided by WordNet with the respective descriptions of their immediate holonyms, hyponyms, as well as two levels of hypernyms. The similarity of two context strings for a document term t and a concept, respectively, is assessed by creating feature vectors for them using bag-of-words TF-IDF weighting and

then applying the cosine measure. Our approach deviates from [21] in that we do not merely select the concept with the highest score because many related senses might be equally relevant when classifying. Instead, all candidate concepts are maintained with their cosine values, normalized with respect to the sum of all values, as their respective weights.

4.2   Weight Propagation

The mapping functions map document terms to the concept terms that immediately represent their respective meanings. ORM, however, not only maps to individual concepts but to entire regions of concepts by propagating a part of a concept's weight to related concepts. For every relation type r, an associated *relation type weight* $\beta_r \in [0, 1)$ is used, e.g. 0.8 for hypernym concept terms and 0.2 for hyponym terms. If a mapping function linked a document term to some concept term c_0 with weight 1.0, the relation type weights mentioned above would provide the direct parent hypernym of c_0 a weight of 0.8, the grandparent would obtain 0.64, and so on, until the values fall below some predetermined threshold. The amount of weight passed on is additionally also governed by the fixed relation weights stored in the ontology (cf. Section 4.1). When multiple paths from a starting concept term c_0 to another term c' exist, the path that maximizes the weight of c' is chosen. For this, Algorithm 4.1, inspired by the A* search algorithm [22], traverses the graph while avoiding cycles and suboptimal paths (see Fig. 1).
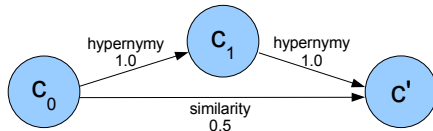

Fig. 1. Suboptimal paths: If c_0 has weight 1.0 and 80% is passed to hypernyms and 40% for similarity, then the direct path from c_0 to c' would only yield a weight of $0.5 \cdot 0.4 = 0.2$ for c', whereas for the indirect path we have $(1.0 \cdot 0.8)^2 = 0.64$.

The algorithm's objective is to determine the optimal weights for related concepts and then accordingly update global concept term counts ctc_c that represent the sum of all weights assigned to a concept while processing an entire document. A list of nodes to be visited is maintained, sorted by weight and initially only containing the starting concept c_0 in conjunction with its weight w_{c_0}. The node c with the highest weight is then repeatedly removed from this list and the counter ctc_c is incremented by c's weight. The algorithm evaluates all neighbours of c, computes their weights and adds them to the list, provided the new weight is greater than a pre-determined threshold $w_{\min}$ as well as any previously computed weight for that particular neighbour. A parameter space search heuristic can be used to empirically determine suitable values for $w_{\min}$ and the β_r values. It can be shown that this algorithm always chooses the optimal weight and terminates if the parameter constraints are fulfilled. In order to

Algorithm 4.1 Ontology-relation-based feature weighting

Input: initial concept c_0 with weight w_{c_0} from an ontology with relation function τ,
  initial term counts ctc_c for concept terms c, a relation type weight $\beta_r < 1$ for every
  relation type r, weight propagation threshold $w_{\min} > 0$

Objective: update concept term counts ctc_c for all relevant concepts c from ontology

1:   $weight_{c_0} \leftarrow w_{c_0}, \;\; weight_c \leftarrow -\infty$ for all $c \neq c_0$
2:   $open \leftarrow \{c_0\}, \; closed \leftarrow \emptyset$
3:   **while** $open \neq \emptyset$ **do**
4:      choose concept c with greatest $weight_c$ from $open$
5:      $open \leftarrow open \backslash \{c\}, \; closed \leftarrow closed \cup \{c\}$ $\triangleright$ Move c to $closed$
6:      $ctc_c \leftarrow ctc_c + weight_c$ $\triangleright$ increase concept term count
7:      **for each** relation entry $(c_i, r_i, w_i) \in \tau(c)$ **do** $\triangleright$ visit neighbours c_i of c
8:          $w \leftarrow \beta_{r_i} \cdot weight_c \cdot w_i$
9:          **if** $w \geq w_{\min}$ and $c_i \notin closed$ **then** $\triangleright$ proceed only if over threshold
10:             $open \leftarrow open \cup \{c_i\}$
11:             $weight_{c_i} \leftarrow \max\{weight_{c_i}, w\}$

---

decrease the runtime, one may add a $|closed| < k$ condition to the while-loop, causing the algorithm to visit only k highest-ranking concepts.

4.3   General Procedure

Instead of multilingual ontologies, one may also use translated documents. In both cases, each document is tokenized and stop words are removed using a fixed list, resulting in a sequence of terms $d = (d_1, \ldots, d_l)$. For each term, an appropriate mapping function then returns a list of corresponding concepts with associated weights. These are then each submitted as input to Algorithm 4.1 with their respective weight such that the concept term counts ctc_c of any additionally relevant concept terms are updated, too. Despite the non-integral values of these concept term counts, one can proceed to compute concept TF-IDF scores similar to those in conventional TC. While a normalization of the ctc_c to concept term frequencies $ctf(d, c)$ is straightforward, the notion of occurrence required for document frequencies does not emanate from our definition of concept term counts as it does in the case of conventional term counts, for it is unclear whether concept terms with a minuscule weight qualify as occurring in the document. We thus use a threshold $\alpha \in [0, 1]$ and define $ctfidf_\alpha(d, c)$ as $ctf(d, c) \cdot \log \frac{1}{cdf_\alpha(c)}$, where $cdf_\alpha(c)$ returns the fraction of all training documents for which $ctc_c > \alpha$ is obtained. The feature space is then constructed by associating each concept term with a separate dimension, and the respective $ctfidf_\alpha$ values can be used to create feature vectors, which are finally normalized. Though not ordinarily covered by mapping functions, technical terms as well as names of people or organizations, for instance, might be crucial when categorizing a document. Hence, an extended setup may be considered, where the ctc_c are combined with conventional term counts. The $ctfidf_\alpha$ values are then computed globally with respect to all such term counts and the feature space has dimensions for concept terms as well as for stems of original document terms.

5 Evaluation

5.1 Experimental Setup and Tuning

In order to evaluate our solution we performed a large number of cross-lingual tests using the SVMlight implementation [23] of Vapnik's SVM with its default settings. We imported WordNet 2.1 [19] for the English language, and additionally applied mapping tables [24] to the Spanish WordNet [20] to synchronize the two resources. A Japanese version of WordNet does not exist, so only translation-based approaches were tested in that case. All translations were performed by AltaVista's Babel Fish Translation service [18].

Two datasets were extracted from *Reuters Corpus Vol. 1* and *2* (RCV1, RCV2) using English training (RCV1) and Spanish test documents (RCV2): one based on topic codes and another one on geographical codes (industry codes could not be used due to inconsistencies between RCV1 and RCV2). An additional dataset with Japanese test documents was generated from Wikipedia [25]. As virtually all TC problems can be reduced to binary ones [17], we tested 105 binary problems per dataset, resulting from 15 randomly selected classes, with 100 training and 600 test documents (Wikipedia: 300) per setup, also selected randomly, however with equal numbers of positive and negative examples in order to avoid biased error rates. A separate validation set was generated based on the same principles as our Reuters topic dataset and then used to tune the relation type weights for hypernyms, holonyms, derivations, etc., as well as other parameters. We chose a value of 0.5 for the α in our *ctfidf*$_\alpha$ formula. For each setup, we also determined the most suitable numbers of features for feature selection based on Information Gain, which turned out to be 1000 for Ontology Region Mapping (ORM).

5.2 Results and Discussion

First of all, Table 1 shows that the conventional bag-of-words method without any translation whatsoever (B) worked surprisingly well, probably due to named entities and because of the relatedness of English and Spanish. Nonetheless, the error rates are unsatisfactory for production use and similar results cannot be achieved for arbitrary language pairs. For Japanese, in fact, this method could not be used directly as advanced tokenization heuristics would be required. As expected, the translation approach T leads to significant improvements.

Applying ORM clearly is beneficial to efficiency compared with a simple concept mapping setup without propagation (CM). The error rates depend on the ontology employed. Better results than with the English/Spanish WordNet setup (CM and ORM) may be obtained by using our ORM approach with translations (TORM), even more so by also including the document terms in the final representation (TORM+T). This is a positive result, implying that ORM with the freely available English WordNet as well as translation software, which also tends to be more available than multilingual lexical resources, suffices for MLTC, as in the case of Japanese, for which a WordNet version currently does not exist.

Table 1. Test Results for Reuters English-Spanish and Wikipedia English-Japanese datasets (micro-averaged F_1 scores in %, average error rates in % with 95% confidence intervals) where **B:** conventional bag-of-words method without translations, **CM:** simple concept mapping approach without weight propagation, **ORM:** Ontology Region Mapping, **ORM+B:** Ontology Region Mapping combined with bag-of-words, **T:** bag-of-words from English translations, **TCM/TORM/TORM+T:** same as CM/ORM/ORM+B but with English translations as input.

| | Reuters Spanish | | | |
| | Topics | | Geography | |
	F_1	error rate	F_1	error rate
B	80.97	18.61 ±0.30	81.86	18.12 ±0.30
CM	89.23	10.49 ±0.24	85.74	14.58 ±0.28
ORM	89.53	10.36 ±0.24	87.33	12.97 ±0.26
ORM+B	91.88	8.04 ±0.21	91.92	8.22 ±0.21
T	90.96	8.80 ±0.22	88.76	11.43 ±0.25
TCM	90.75	9.06 ±0.22	91.12	9.16 ±0.23
TORM	91.12	8.74 ±0.22	93.89	6.28 ±0.19
TORM+T	92.46	7.43 ±0.20	94.44	5.68 ±0.18

| | Wikipedia Japanese | |
	F_1	error rate
T	86.26	14.00 ±0.38
TCM	85.38	15.10 ±0.40
TORM	86.67	13.52 ±0.38
TORM+T	87.29	12.86 ±0.37

For news and encyclopedic articles, outperforming the T method is a rather difficult task in light of the considerable discriminatory power of the terms in the introduction paragraphs. Nonetheless, our methods delivered superior results that are statistically significant. Geographical references, in contrast, are often less explicit, so our methods pay off even more. Given that the relation type weights were tuned with respect to the Reuters topic-based validation set, we may presume that even better results than the ones indicated are achievable.

6   Conclusions and Future Work

In the past, many attempts to use natural language processing for monolingual TC have failed to deliver convincing results [17]. A linguistic analysis led us to a novel approach called Ontology Region Mapping, where related concepts, too, are taken into consideration when mapping from terms to concepts, so additional background knowledge is exploited, which is particularly useful in multilingual settings. Our experimental evaluation confirmed our intuitions.

In the future, we would like to devise strategies for constructing multilingual resources that integrate more background knowledge and better reflect the semantic relatedness of concepts than WordNet. Additionally, a more sophisticated word-to-concept mapping setup could be used that recognizes compounds and disambiguates better. Finally, it could be explored how well ORM performs for multilingual information retrieval and text clustering. Indeed, we believe our feature weighting approach or extensions of it to have a wide range of interesting applications, in multilingual as well as monolingual settings, because it captures the general meaning of a text more adequately than established schemes.

References

1. Bel, N., Koster, C.H.A., Villegas, M.: Cross-lingual text categorization. Proc. ECDL 2003 (2003) 126–139
2. García Adeva, J.J., Calvo, R.A., de Ipiña, D.L.: Multilingual approaches to text categorisation. Europ. J. for the Informatics Professional **VI**(3) (2005) 43 – 51
3. Jalam, R.: Apprentissage automatique et catégorisation de textes multilingues. PhD thesis, Université Lumière Lyon 2, Lyon, France (2003)
4. Olsson, J.S., et al.: Cross-language text classification. Proc. SIGIR 2005 (2005) 645–646
5. Rigutini, L., et al.: An EM based training algorithm for cross-language text categorization. In: Proc. Web Intelligence 2005, Washington, DC, USA (2005) 529–535
6. Oard, D.W., Dorr, B.J.: A survey of multilingual text retrieval. Technical report, University of Maryland at College Park, College Park, MD, USA (1996)
7. de Buenaga Rodríguez, M., et al.: Using WordNet to complement training information in text categorization. Proc. 2nd RANLP (1997)
8. Moschitti, A., Basili, R.: Complex linguistic features for text classification: a comprehensive study. Adv. in IR, Proc. ECIR 2004 (2004)
9. Ifrim, G., Theobald, M., Weikum, G.: Learning word-to-concept mappings for automatic text classification. Proc. 22nd ICML - LWS (2005) 18–26
10. Verdejo, F., Gonzalo, J., Peñas, A., et al.: Evaluating wordnets in cross-language text retrieval. Proceedings LREC 2000 (2000)
11. Scott, S., Matwin, S.: Text classification using WordNet hypernyms. Proc. Worksh. Usage of WordNet in NLP Systems at COLING-98 (1998) 38–44
12. Bloehdorn, S., Hotho, A.: Boosting for text classification with semantic features. Proc. Worksh. on Mining for/from the Semantic Web at KDD 2004 (2004) 70–87
13. Ramakrishnanan, G., et al.: Text representation with WordNet synsets using soft sense disambiguation. Ing. systèmes d'information **8**(3) (2003) 55–70
14. Gliozzo, A.M., et al.: Cross language text categ. by acq. multil. domain models from comp. corpora. Proc. ACL Worksh. Building and Using Parallel Texts (2005)
15. Dumais, S.T., et al.: Automatic cross-language retrieval using latent semantic indexing. AAAI Symposium on CrossLanguage Text and Speech Retrieval (1997)
16. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc., New York, NY, USA (1995)
17. Sebastiani, F.: Machine learning in automated text categorization. ACM Computing Surveys **34**(1) (2002) 1–47
18. AltaVista: Babel fish translation. http://babelfish.altavista.com/ (2006)
19. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)
20. Farreres, X., Rigau, G., Rodríguez, H.: Using WordNet for building WordNets. Proc. Conf. Use of WordNet in NLP Systems (1998) 65–72
21. Theobald, M., Schenkel, R., Weikum, G.: Exploiting structure, annotation, and ontological knowledge for automatic classification of XML data. 6th Intl. Worksh. Web and Databases (2003) 1–6
22. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach. Prentice-Hall, Englewood Cliffs, NJ, USA (1995)
23. Joachims, T.: Making large-scale support vector machine learning practical. Advances in Kernel Methods: Support Vector Machines (1999)
24. Daudé, J., et al.: Making Wordnet mappings robust. Proc. Congreso de la Sociedad Española para el Procesamiento del Lenguage Natural (SEPLN) (2003)
25. Wikimedia Foundation: Wikipedia. http://www.wikipedia.org/ (2006)