Multilingual Text Classification through Combination of Monolingual Classifiers

Teresa Gonalves, Paulo Quaresma
Departamento de Informtica, Universidade de vora
7000-671 vora, Portugal
(`tcg@di.uevora.pt, pq@di.uevora.pt`*)*

Abstract. With the globalization trend there is a big amount of documents written in different languages. If these polylingual documents are already organized into existing categories one can deliver a learning model to classify newly arrived polylingual documents. Despite being able to adopt a simple approach by considering the problem as multiple independent monolingual text classification problems, this approach fails to use the opportunity offered by polylingual training documents to improve the effectiveness of the classifier. This paper proposes a method to combine different monolingual classifiers in order to get a new classifier as good as the best monolingual one having also the ability to deliver the best performance measures possible (precision, recall and F_1). The proposed methodology was applied to a corpus of legal documents – from the EUR-Lex site – and was evaluated. The obtained results were quite good, indicating that combining different mono-lingual classifiers may be a promising approach to reach the best performance for each category independently of the language.

Keywords: Multilingual text classification, Machine Learning, Support Vector Machines

1. Introduction

Current Information Technologies and Web-based services need to manage, select and filter increasing amounts of textual information. Text classification allows users, through navigation on class hierarchies, to browse more easily the texts of their interests. This paradigm is very effective both in filtering information as in the development of online end-user services.

Since the number of documents involved in these applications is large, efficient and automatic approaches are necessary for classification. A Machine Learning approach can be used to automatically build the classifiers. The construction process can be seen as a problem of supervised learning: the algorithm receives a relatively small set of labelled documents and generates the classifier. Several algorithms have been applied, such as decision trees, linear discriminant analysis and logistic regression, the nave Bayes algorithm and Support Vector Machines (SVM). Besides having a justified learning theory describing its mechanics, SVM are known to be computationally efficient, robust and accurate.

Because of the globalization trend, an organization or individual often generates, acquires and archives the same document written in different languages (i.e., polylingual documents); moreover, many countries adopt multiple languages as their official languages. If these polylingual documents are organized into existing categories one would like to use this set of pre-classified documents as training documents to build models to classify newly arrived polylingual documents.

For multilingual text classification (i.e., collections of documents written in several languages), some prior studies address the challenge of cross-lingual text classification. However, prior research has not paid much attention to using polylingual documents yet. This study is motivated by the importance of providing polylingual text classification support to organizations and individuals in the increasingly globalized and multilingual environment.

We propose a method that combines different monolingual classifiers in order to get a new classifier as good as the best monolingual one which has the ability to deliver all the best performance measures (precision, recall and F1) possible.

This methodology was applied and evaluated on a set of legal documents from the EUR-Lex site. We collected documents for two anglo-saxon languages (English and German) and two roman ones (Italian and Portuguese), obtaining four different sets. The obtained results were quite good, indicating that combining different monolingual classifiers may be a promising approach to the problem of classifying documents written in several languages.

The paper is organized as follows: Section 2 describes the main concepts and tools used in our approach, Section 3 introduces the methodology for combining monolingual classifiers and Section 4 presents the document collection used for evaluation, describes the experimental setup and evaluates the obtained results. Finally, Section 5 presents some conclusions and points out possible future work.

2. Concepts and Tools

This section introduces the Automatic Text Classification approach and the classification algorithm and software tool used in this work.

2.1. Automatic Text Classification

Originally, research in Automatic Text Classification addressed the binary problem, where a document is either relevant or not w.r.t. a given category. However, in real-world situations the great variety of different sources and hence categories usually poses a multi-class classification problem, where a document belongs to exactly one category from a predefined set. Even more general is the multi-label problem, where a document can be classified into more than one category.

In order to be fed to the learning algorithm, documents must by preprocessed to obtain a more structured representation. The most common approach is to use a bag-of-words representation (Salton, 1975), where each document is represented by the words it contains, with their order and punctuation being ignored. Normally, words are weighted by some measure of word's frequency in the document and, possibly, the corpus. In most cases, a subset of words (stop-words) is not considered, because their role is related to the structural organization of the sentences and does not have discriminating power over different classes and some works reduce semantically related terms to the same root applying a lemmatizer.

Research interest in this field has been growing in the last years. Several machine learning algorithms were applied, such as decision trees (Tong, 1994), linear discriminant analysis and logistic regression (Schütze, 1995), the naïve Bayes algorithm (Mladenić, 1999) and Support Vector Machines (SVM)(Joachims, 1999). Joachims (Joachims, 2002) says that using SVMs to learn text classifiers is the first approach that is computationally efficient and performs well and robustly in practice. There is also a justified learning theory that describes its mechanics with respect to text classification.

2.1.1. *Multilingual text classification.*

While most text classification studies focus on monolingual documents, some point to multilingual text classification. From these, the great majority address the challenge of crosslingual text classification where the classification model relies on monolingual training documents and a translation mechanism to classify documents written in another language (Bel, 2003; Rigutini, 2005; Lee, 2009). A technique that takes into account all training documents of all languages when constructing a monolingual classifier for a specific language is proposed in (Wei, 2007). Wei et al. showed that for English and Chinese a feature-based reinforcement polylingual category integration approach obtains better accuracy then monolingual ones. Our proposal is quite different because we do not use information from other languages and multilingual thesaurus to build the individual classifiers. Our aim is to combine individual classifiers in order to obtain a better classifier and not to improve individual classifiers.

2.2. Support Vector Machines

Support Vector Machines, a learning algorithm introduced by Vapnik and coworkers (Cortes, 1995), was motivated by theoretical results from statistical learning theory: it joins a kernel technique with the structural risk minimization framework.

Kernel techniques comprise two parts: a module that performs a mapping from the original data space into a suitable feature space and a learning algorithm designed to discover linear patterns in the (new) feature space. The *kernel function*, that implicitly performs the mapping, depends on the specific data type and domain knowledge of the particular data source.

The *learning algorithm* is general purpose and robust. It's also efficient since the amount of computational resources required is polynomial with the size and number of data items, even when the dimension of the embedding space grows exponentially (Shawe-Taylor, 2004). A mapping example is illustrated in Fig. 1a).

The *structural risk minimization* (SRM) framework creates a model with a minimized VC (Vapnik-Chervonenkis) dimension. This developed theory (Vapnik, 1998) shows that when the VC dimension of a model is low, the expected probability of error is low as well, which means good performance on unseen data (good generalization). In geometric terms, it can be seen as a search to find, between all decision surfaces (the $\mathcal{T}$-dimension surfaces that separate positive from negative examples) the one with maximum margin, that is, the one having a separating property that is invariant to the most wide translation of the surface. This property can be enlighten by Fig. 1b) that shows a 2-dimensional problem.

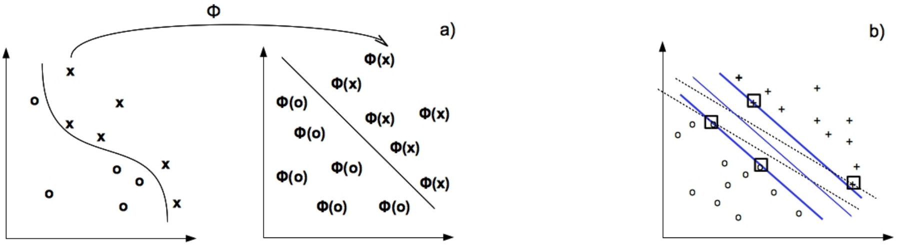

Figure 1. The SVM approach: kernel transformation and search for maximum margin.

2.2.1. *Classification software.*

As classification software we used SVM^{light} (Joachims, 1999)[1]. It is a C implementation of SVM that allows solving classification, regression and ranking problems, handles many thousands of support vectors and several hundred-thousands of training examples and supports standard kernel functions besides letting the user define its own.

3. Combining monolingual classifiers

Having documents in several languages, one can adopt a nave approach by considering the problem as multiple independent monolingual text classification problems. This simple approach only employs the training documents

[1] Available at `http://svmlight.joachims.org`

of one language to construct a monolingual classifier for that language and ignores all training documents of other languages. When a new document in a specific language arrives, one select the corresponding classifier to predict appropriate category(s) for the target document. However, the independent construction of each monolingual classifier fails to use the opportunity offered by polylingual training documents to improve the effectiveness of the classifier.

With this bearing in mind, and to get a decision for a new document, monolingual classifiers could be improved up in several ways. We propose the following strategies for the combination system:

- the sum of SVMs output values
- the F_1 weighted sum of SVMs output values
- the F_1 weighted sum of SVMs decisions

The above measures could also be used to draw decisions when considering a voting strategy of the monolingual classifiers.

4. Experiments

This section introduces the dataset, describes the experimental setup and presents the obtained results for the legal concepts classification task.

4.1. Dataset description

For testing the proposed methodology, experiments were run over a set of European Union law documents. These documents were obtained from the EUR-Lex site[2] within the "International Agreements" section, belonging to the "External Relations" subject matter. From all available agreements we chose the ones with full text (not just bibliographic notice) obtaining a set of 2714 documents (dated from 1953 to 2008).

Since agreements are available in several languages we collected them for two anglo-saxon languages (English and German) and two roman ones (Italian and Portuguese), obtaining four different corpora: `eurlex-EN`, `eurlex-DE`, `eurlex-IT` and `eurlex-PT`. Table I presents the total number and average per document of tokens (running words) and types (unique words).

Each document is classified onto several ontologies: the "EUROVOC descriptor", the "Directory code" and the "Subject matter". In all available classifications each document can be assigned to several categories. For our classification problem we used the first level of the "Directory code" classification, considering only categories with at least 50 documents. Table II shows each category along with the number of documents assigned.

[2] Available at `http://eur-lex.europa.eu/en/index.htm`

Table I. Total number and average per document of tokens and types for each corpus.

	tokens		types	
corpus	total	per doc	total	per doc
`eurlex-EN`	10699234	3942	73091	570
`eurlex-DE`	10145702	3728	133191	688
`eurlex-IT`	10665455	3929	96029	636
`eurlex-PT`	9731861	3585	86086	567

Table II. Number of documents assigned to each category.

id	*name*	*# of docs*
2	Customs Union and free movement of goods	209
3	Agriculture	390
4	Fisheries	361
7	Transport policy	81
11	External relations	2628
12	Energy	58
13	Industrial policy and internal market	55
15	Environment, consumers and health protection	138
16	Science, information, education and culture	99

4.2. Experimental setup

The experiments were done using a bag-of-words representation of documents, the SVM algorithm was run using SVM^{light} with a linear kernel and other default parameters and the model was evaluated using a 10-fold stratified cross-validation procedure with significance tests done with a 90% confidence level.

To represent each document we used the bag-of-words approach, a *vector space model* (VSM) representation where each document is represented by the words it contains, with their order and punctuation being ignored. Document's representation was obtained by mapping all numbers to the same token and using the tf-idf weighting function normalized to unit length.

To measure learner's performance we analyzed precision, recall and the F_1 measures (Salton, 1975) of the positive class. These measures are obtained from contingency table of the classification (prediction *vs.* manual classification).

4.3. Monolingual experiments

To support our claim, as baseline we have built classifiers for each language. Table III shows the average precision, recall and F_1 measures for each corpus and each category (boldface values are significantly worse than the best value obtained). Last line presents the average values over all nine classes.

Table III. Average precision, recall and F_1 values for each mono-lingual classifier.

	precision				*recall*				F_1			
id	EN	DE	IT	PT	EN	DE	IT	PT	EN	DE	IT	PT
2	**.919**	.957	**.922**	.937	.651	.665	**.580**	**.565**	.755	.778	**.702**	**.701**
3	.916	.928	.938	.943	.818	.805	**.705**	**.503**	.862	.860	**.803**	**.655**
4	**.956**	.966	.980	.971	.934	.906	.914	**.823**	.944	.934	.945	.890
7	.846	.870	**.793**	.806	.568	.543	.518	.482	.651	.640	.608	.590
11	.973	.973	.973	.973	.998	.997	.998	.997	.985	.985	.985	.985
12	.958	**.874**	**.877**	.938	.637	.700	.670	.600	.752	.765	.745	.716
13	.942	.933	.933	.967	.393	.320	.300	.320	.522	.454	.436	.461
15	.909	.922	.917	.908	.726	.732	.725	.732	.801	.813	.805	.806
16	**.862**	**.883**	.916	.947	.779	.799	.718	.647	.804	.828	.785	**.753**
avg	.828	.832	.825	.839	.650	.647	.613	.567	.708	.706	.681	.656

For the precision values we can notice that the Portuguese dataset has values with no significant difference with the "best" for all classes; all other languages perform worse for some classes (English: $c2$, $c4$ and $c16$; German: $c12$ and $c16$; Italian: $c2$, $c7$ and $c12$). With this in mind one can say that the Portuguese language generates the best precision classifiers.

Concerning recall, it's the English and German languages that consistently present the best values; Italian and Portuguese while equally good for some classes, are worse for others (Italian: $c2$ and $c3$; Portuguese: $c2$, $c3$ and $c4$).

The F_1 measure presents the same behavior as recall, being the only difference the classes where the Portuguese language performs worse ($c2$, $c3$ and $c16$).

4.4. Polylingual experiments

From all possible combiners (see Section 3), there is one that, for all classes, persistently generated the best F_1 values: the F_1 weighted sum of SVMs decisions.

Table IV shows, for each performance measure its results compared with the "best" monolingual classifiers(boldface values are significantly worse than the corresponding multilingual one): the Portuguese one for precision, and the English and German one for recall and F_1. Last line equally presents the average values over all classes.

Table IV. Average precision, recall and F_1 values compared with the combiner ones.

	precision		*recall*			F_1		
id	PT	comb	EN	DE	comb	EN	DE	comb
2	.937	.947	.651	.665	.675	.755	.778	.782
3	.943	.925	.818	.805	.813	.862	.860	.863
4	.971	.964	.934	**.906**	.928	.944	.934	.945
7	.806	.868	.568	.543	.567	.651	.640	.654
11	.973	.973	.998	.997	.998	.985	.985	.985
12	.938	.908	.637	.700	.670	.752	.765	.761
13	.967	.933	.393	.320	.340	.522	.454	.467
15	.908	.912	.726	.732	.754	.801	.813	.821
16	.947	.881	.779	.799	.779	.804	.828	.815
avg	.839	.831	.650	.647	.652	.708	.706	.709

From the average values, one can easily see that precision is higher than recall and that the best monolingual classifier depends on what performance measure one is considering. Nevertheless, the combined classifier has all performance measures very similar and never significatively worse then the best monolingual classifier.

In fact, significant tests show that, for all classes and all performance measures, there is no significant difference between the "best" monolingual classifier and the corresponding combined classifier.

5. Conclusions and Future Work

A proposal to combine monolingual classifiers was presented and evaluated. The proposed methodology uses SVM classifiers to associate concepts to legal documents and uses a decision function that combines them in order to obtain, for each class, a classifier as good as the best monolingual classifier of each performance measure.

The baseline experiments allows one to conclude that some languages generate classifiers with better precision values (Portuguese language) while others generate classifiers with better recall ones (English and German languages). In order to be able to explain and to try to generalise these results further experiments need to be done. For instance, we will need to evaluate this methodology with other collections and domains. Are these results specific for the legal domain? Or only for this collection and topics? Nevertheless, from a linguistic point of view, these results raise quite interesting questions.

By combining all classifiers one obtains a classifier as good as the best monolingual one. This combined classifier can even be considered better than the others since it has the ability to deliver all the best performance measures (precision, recall and F_1) unlike using one monolingual classifier.

As ongoing research we intend to use a deeper linguistic representation of documents and to re-evaluate this methodology. Specifically, we will use a semantic representation (based on DRS[3]) of documents and a graph kernel to create SVM models. In previous work, this approach showed to be able to improve the bag-of-words result for the Portuguese language. Another research line is to use legal thesaurus, such as the LOIS[4] lexical thesaurus, to reinforce some features/terms. With this approach we would combine our proposal with the main ideas of the Wei et al. work (Wei, 2007).

References

Bel, N., Koster, C. and Villegas, M. (2003), *Cross-lingual text categorization*, in Proceedings of ECDL'03, Proceedings of the 7th European Conference on Research and Advanced Tecnology for Digital Libraries, pp. 126–139.

Cortes, C. and Vapnik, V. (1995), *Support-vector networks*, Machine Learning, Vol. 20 No. 3, pp. 273–297.

Joachims, T. (1999a), *Making large-scale SVM learning practical*, in Schölkopf, B., Burges, C. and Smola, A. (Ed.), "Advances in Kernel Methods - Support Vector Learning", MIT Press.

Joachims, T. (2002), *Learning to Classify Text Using Support Vector Machines*, Kluwer Academic Publishers.

Lee, C.H. and Yang, H.C. (2009), *Construction of supervised and unsupervised learning systems for multilingual text categorization*, Expert Systems Applications, Vol. 36 No. 2, pp. 2400–2410.

Mladenić, D. and Grobelnik, M. (1999), *Feature selection for unbalanced class distribution and naïve Bayes*, in Proceedings of ICML'99, 16th International Conference on Machine Learning, pp. 258–267.

Rigutini, L., Maggini, M. and Liu, B. (2005), *An EM Based Training Algorithm for Cross-Language Text Categorization*, in Proceedings of WI'05, IEEE/WIC/ACM International Conference on Web Intelligence (IEEE Computer Society), pp. 529–535.

Salton, G., Wang, A. and Yang, C. (1975), *A vector space model for information retrieval*, Journal of the American Society for Information Retrieval, Vol. 18, pp. 613–620.

Schütze, H., Hull, D. and Pedersen, J. (1995), *A comparison of classifiers and document representations for the routing problem*, in Proceedings of SIGIR'95, 18th International Conference on Research and Developement in Information Retrieval (ACM), pp. 229–237.

Shawe-Taylor, J. and Cristianini, N. (2004), *Kernel Methods for Pattern Analysis*, Cambridge University Press.

[3] Discourse Representation Structures

[4] Lexical Ontologies for Legal Information Sharing

Tong, R. and Appelbaum, L.A. (1994), *Machine learning for knowledge-based document routing*, in Proceedings of TRC'94, 2nd Text Retrieval Conference.

Vapnik, V. (1998), *Statistical learning theory*, Wiley, NY.

Wei, C., Shi, H. and Yang, C. (2007), *Feature reinforcement approach to poly-lingual text categorization*, in Proceedings of the International Conference on Asia Digital Libraries (LNCS Springer), pp. 99–108.