Multilingual Document Classification via Transductive Learning

Salvatore Romeo[1], Dino Ienco[2], and Andrea Tagarelli[1]

[1] DIMES, University of Calabria, Italy
{sromeo,tagarelli}@dimes.unical.it
[2] IRSTEA - UMR TETIS, and LIRMM, Montpellier, France
dino.ienco@irstea.fr

Abstract. We present a transductive learning based framework for multilingual document classification, originally proposed in [7]. A key aspect in our approach is the use of a large-scale multilingual knowledge base, BabelNet, to support the modeling of different language-written documents into a common conceptual space, without requiring any language translation process. Results on real-world multilingual corpora have highlighted the superiority of the proposed document model against existing language-dependent representation approaches, and the significance of the transductive setting for multilingual document classification.

1 Introduction

Multilingual document collections are getting increased attention as their analysis is essential to support a variety of tasks, such as building translation resources [8, 6], detection of plagiarism in patent collections [1], cross-lingual document similarity and multilingual document categorization [4, 2]. Focusing on the latter problem, existing methods in the literature can mainly be characterized based on the language-specific resources they use to perform cross-lingual tasks. A common approach is to resort to machine translation techniques or bilingual dictionaries to mapping every document to the target language, and perform cross-lingual document similarity and categorization (e.g., [4]).

We address the multilingual document classification problem from a different perspective. First, we are not restricted to deal with bilingual corpora dependent on machine translation. In this regard, we exploit a large, publicly available knowledge base specifically designed for multilingual retrieval tasks: *BabelNet* [6]. BabelNet embeds both the lexical ontology capabilities of WordNet and the encyclopedic power of Wikipedia. Second, our view is different from the standard inductive learning setting. High-quality labeled datasets are in fact difficult to obtain due to costly and time-consuming annotation processes. This particularly holds for the multilingual scenario where the documents belong to different languages and, as a consequence, more language-specific experts need to be involved in the annotation process. Moreover, in multilingual corpora documents are often all available at the same time and the classifications for the unlabeled

instances need to be provided contextually to the learning of the current document collection. To fulfill the last two requisites, *transductive learning* offers an effective approach [4] as it requires few labels for decision making, and the learning process is tailored to the particular dataset.

Motivated by the above considerations, we present a framework for multilingual document classification under a transductive learning setting, originally proposed in [7]. By introducing a unified conceptual feature space based on BabelNet, we define a multilingual document representation model which does not require any language translation. We resort to a state-of-the-art transductive learner [5] to produce the document classification. Using RCV2 and Wikipedia document collections, we compare our proposal w.r.t. document representations typically involved in multilingual and cross-lingual analysis.

2 Transductive Multilingual Document Categorization

2.1 Text Representation Models

Bag-of-words and machine-translation based models. The classic *bag-of-words* model has been employed also in the context of multilingual documents. Hereinafter we use notation BoW to refer to the term-frequency vector representation of documents over the union of language-specific term vocabularies.

A common approach adopted in the literature is to translate all documents to a unique anchor language and then represent the translated documents with the BoW model [4]. In this work, we have considered three settings corresponding to the use of *English* (*BoW-MT-en*), *French* (*BoW-MT-fr*) or *Italian* (*BoW-MT-it*) as anchor language. We also employ a dimensionality reduction approach via Latent Semantic Analysis (LSA) [3] over the BoW representation. We will refer to this model as *BoW-LSA*.

Bag-of-synset representation. Differently from the previously discussed document representations, we propose to model a collection of multilingual documents into a unified *conceptual feature space*. Our key idea is to exploit the multilingual lexical knowledge of BabelNet [6], in order to generate document features that correspond to BabelNet synsets. More specifically, the input document collection is subject to a two-step processing phase. In the first step, each document is broken down into a set of lemmatized and POS-tagged sentences, in which each word is replaced with related lemma and associated POS-tag ($\langle w, POS(w)\rangle$). In the second step, word sense disambiguation is performed over each pair $\langle w, POS(w)\rangle$ to detect the most appropriate BabelNet synset σ_w contextually to any sentence in the document. Each document is finally modeled as a $|\mathcal{BS}|$-dimensional vector of BabelNet synset frequencies, where $\mathcal{BS}$ is the set of retrieved BabelNet synsets. We will refer to this representation model as *BoS* (i.e., *bag-of-synsets*)

2.2 Transductive Setting and Label Propagation Algorithm

A major contribution of our work is the use of a transductive learning based approach to address the problem of multilingual document classification. For

this purpose, we use a particularly effective transductive learner, named Robust Multi-class Graph Transduction ($RMGT$) [5].

RMGT implements a graph-based label propagation approach, which exploits a kNN graph built over the entire document collection to propagate the class information from the labeled to the unlabeled documents. The transductive learning scheme used by RMGT employs spectral properties of the kNN graph to spread the labeled information over the set of test documents. Specifically, the label propagation process is modeled as a constrained convex optimization problem where the labeled documents are employed to constrain and guide the final classification. After the propagation step, every unlabeled document d_i is associated to a vector representing the likelihood of the document d_i for each of the classes; therefore, d_i is assigned to the class that maximizes the likelihood.

3 Experimental Setting and Results

We evaluated our transductive learning based approach on two multilingual document collections, RCV2 and Wikipedia. Here we summarize results obtained on Wikipedia; the interested reader is referred to [7] for further details. We considered documents in three different languages, *English*, *French*, and *Italian*, covering 6 different topic-classes. Topics were selected to obtain, for each topic-language pair, the same number of documents. The resulting *balanced* dataset is comprised of 1 000 documents for each topic-language pair, with a total of 18 000 documents. The number and density of terms for the *BoW* model (resp. synsets in *BoS*) are 15 634 and 1.61E-2 (resp. 10 247 and 1.81E-2). We also produced an *unbalanced* version of the dataset by keeping the whole subset of English documents while sampling half of the French and half of the Italian subsets. To build both datasets, every document was subject to tokenization and lemmatization.[3] To setup the transductive learner, we used $k = 10$ for the kNN graph construction, and varied the percentage of labeled documents from 1% to 20% with step of 1%. To measure the classification performance, we used standard F-measure, Precision and Recall. Results were averaged over 30 runs.

Figure 1 shows a summary of the best average performances on both balanced and unbalanced corpora, and also shows results obtained on the balanced corpus for different values of the training percentage of the transductive learner. We observe that *BoS* clearly outperforms the other document representation models, including *BoW-MT-en* which in this case achieves similar (or even slightly lower) results to *BoW-MT-fr* and *BoW-MT-it*. *BoW-LSA* and *BoW* also show a performance gap from the other models. Considering the unbalanced scenario, the *BoS* results are still higher than the best competing methods. Interestingly, *BoS* performance is the same as for the balanced case, which would indicate a higher robustness of *BoS* w.r.t. the corpus characteristics. Another remark is that the proposed *BoS* not only performs significantly better than the other models, but also it exhibits a performance trend that is not affected by issues related to language specificity. In fact, the machine-translation based models have relative performance that may vary on different datasets; no preference on translation

[3] http://nlp.lsi.upc.edu/freeling/

	Balanced			*Unbalanced*		
	FM	*P*	*R*	*FM*	*P*	*R*
BoS	**0.912**	**0.915**	**0.912**	**0.912**	**0.915**	**0.912**
BoW	0.872	0.876	0.872	0.797	0.817	0.794
BoW-MT-en	0.895	0.896	0.895	0.902	0.903	0.902
BoW-MT-fr	*0.898*	*0.899*	*0.898*	0.904	0.906	0.904
BoW-MT-it	0.897	*0.899*	0.897	*0.905*	*0.907*	*0.905*
BoW-LSA	0.863	0.867	0.863	0.838	0.845	0.838

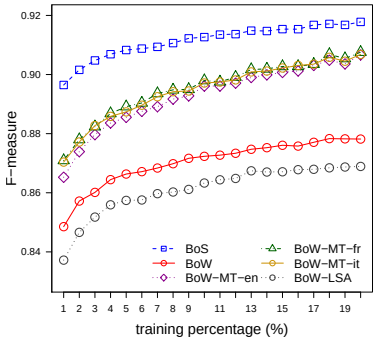

Fig. 1. Summary of best average performance results of the various representation methods (left), and average F-measure results on the balanced corpus (right).

languages can be made in advance as a language that leads to better results on a dataset can perform worse than other languages on another dataset.

4 Conclusion

We have presented a knowledge-based framework for multilingual document classification under a transductive setting. Our *BoS* document model has shown to be effective for multilingual comparable corpora, as it supports the transductive learner to obtain better classification performance than language-dependent document models, using a relatively small portion of labeled data. Future works will concentrate on the development of a richer conceptual document model that can incorporate more types of information (i.e., relations among the synsets), and on investigating hybrid solutions of transductive and active learning.

References

1. A. Barrón-Cedeño, P. Gupta, and P. Rosso. Methods for cross-language plagiarism detection. *Knowl.-Based Syst.*, 50:211–217, 2013.
2. A. Barrón-Cedeño, M. L. Paramita, P. D. Clough, and P. Rosso. A comparison of approaches for measuring cross-lingual similarity of wikipedia articles. In *ECIR*, pages 424–429, 2014.
3. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *J. Assoc. Inf. Sci.*, 41(6):391–407, 1990.
4. Y. Guo and M. Xiao. Transductive representation learning for cross-lingual text classification. In *ICDM*, pages 888–893, 2012.
5. W. Liu and S. Chang. Robust multi-class transductive learning with graphs. In *CVPR*, pages 381–388, 2009.
6. R. Navigli and S. P. Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, 2012.
7. S. Romeo, D. Ienco, and A. Tagarelli. Knowledge-based representation for transductive multilingual document classification. In *ECIR*, pages 92–103, 2015.
8. P. Vossen. EuroWordNet: A Multilingual Database of Autonomous and Language-Specific Wordnets Connected via an Inter-Lingual Index. *International Journal of Lexicography Vol.17*, 2:161–173, 2004.