Knowledge-Based Representation for Transductive Multilingual Document Classification

Salvatore Romeo, Dino Ienco, Andrea Tagarelli

Knowledge-based Representation for Transductive Multilingual Document Classification

Salvatore Romeo[1], Dino Ienco[2,3], and Andrea Tagarelli[1]

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

Abstract. Multilingual document classification is often addressed by approaches that rely on language-specific resources (e.g., bilingual dictionaries and machine translation tools) to evaluate cross-lingual document similarities. However, the required transformations may alter the original document semantics, raising additional issues to the known difficulty of obtaining high-quality labeled datasets. To overcome such issues we propose a new framework for multilingual document classification under a transductive learning setting. We exploit 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. We resort to a state-of-the-art transductive learner to produce the document classification. Results on two real-world multilingual corpora have highlighted the effectiveness of the proposed document model w.r.t. document representations usually involved in multilingual and cross-lingual analysis, and the robustness of the transductive setting for multilingual document classification.

1 Introduction

Textual data constitutes a huge, continuously growing source of information, as everyday millions of documents are generated. This is partly explained by the increased popularity of tools for collaboratively editing through contributors across the world, which eases the production of different language-written documents, leading to a new phenomenon of *multilingual information overload*. Analyzing multilingual document collections is getting increased attention as it can support a variety of tasks, such as building translation resources [20, 14], detection of plagiarism in patent collections [1], cross-lingual document similarity and multilingual document classification [18, 16, 6, 2, 5].

In this paper, we focus 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 resorting to machine translation techniques or bilingual dictionaries to map a document to the target

language, and then perform cross-lingual document similarity and categorization (e.g., [6, 9]). Some works (e.g., [16, 2]) have also used Wikipedia as benchmark or knowledge base. However, in a cross-lingual supervised setting, the classification performance can significantly vary by exchanging documents from source to target languages. The language-specific machine translation systems typically introduce noise in understanding the document semantics, thus negatively affecting the final results. Furthermore, the classification performance will depend on the number and quality of the multilingual documents obtained by a single yet non-ontological knowledge base like Wikipedia.

We address the multilingual document classification problem differently from the above mentioned approaches. 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* [14]. 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: in multilingual corpora often documents are 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. Examples of such tasks are relevance feedback, online news filtering, and reorganization of a document collection, where the system needs to automatically label documents in a collection starting from few labeled ones supplied by the user. Finally, high-quality labeled datasets are difficult to obtain due to costly and time-consuming annotation processes. This particularly holds for the multilingual scenario where language-specific experts need to be involved in the annotation process. To deal with these issues, *transductive learning* [7] offers an effective approach to supplying contextual classification of unlabeled documents by using a relatively small set of labeled ones. This learning setting fits well real-world applications and it can be very helpful in multilingual text analysis, where document labels are more difficult to obtain than in the monolingual counterpart and the classification decisions should not be made separately from learning the current data.

Motivated by the above considerations, in this work we propose a new framework for multilingual document classification under a transductive learning setting. By exploiting BabelNet, we model the multilingual documents using a common conceptual feature space. This representation model does not impose any methodological limitation on the number of languages of the documents. We then employ a state-of-the-art transductive learner [10] to produce the document classification. Using RCV2 and Wikipedia document collections, we compare our proposal w.r.t. document representations usually involved in multilingual and cross-lingual analysis. To the best of our knowledge, this is the first work that analyzes multilingual documents using a transductive learner through the lens of BabelNet. Note that transductive learning is also considered in [6], however only for bilingual analysis. Moreover, [5] also exploits BabelNet, although to propose a bilingual similarity measure, while our approach can effectively deal with comparable corpora in more than two languages.

2 Background on BabelNet

BabelNet [14] is a multilingual semantic network obtained by linking Wikipedia with WordNet, that is, the largest multilingual Web encyclopedia and the most popular computational lexicon. The linking of the two knowledge bases was performed through an automatic mapping of WordNet synsets and Wikipages, harvesting multilingual lexicalization of the available concepts through human-generated translations provided by the Wikipedia inter-language links or through machine translation techniques.

It should be noted that the large-scale coverage of both lexicographic and encyclopedic knowledge represents a major advantage of BabelNet versus other knowledge bases that could in principle be used for cross-lingual or multilingual retrieval tasks. For instance, the multilingual thesaurus EUROVOC (created by the European Commission's Publications Office) was used in [18] for document similarity purposes; however, EUROVOC utilizes less than 6 000 descriptors, which leads to evident limits in semantic coverage. Furthermore, other knowledge bases such as EuroWordNet [20] only utilize lexicographic information, while conversely studies that focus on Wikipedia (e.g., [16, 2]) cannot profitably leverage on lexical ontology knowledge.

Multilingual knowledge in BabelNet is represented as a labeled directed graph in which nodes are concepts or named entities and edges connect pairs of nodes through a semantic relation. Each edge is labeled with a relation type (is-a, part-of, etc.), while each node corresponds to a *BabelNet synset*, i.e., a set of lexicalizations of a concept in different languages. BabelNet also provides functionalities for graph-based word sense disambiguation in a multilingual context. Given an input set of words, a semantic graph is built by looking for related synset paths and by merging all them in a unique graph. Once the semantic graph is built, the graph nodes can be scored with a variety of algorithms. Finally, this graph with scored nodes is used to rank the input word senses by a graph-based approach.

3 Transductive Multilingual Document Classification

3.1 Text representation models

Bag-of-synset representation. We model the multilingual documents into a common *conceptual feature space*, which is built using the multilingual lexical knowledge of BabelNet [17]. We will refer to this representation as *BoS* (i.e., *bag-of-synsets*), since conceptual features of the documents correspond to BabelNet synsets.

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. Let us denote with $\langle w, POS(w)\rangle$ a lemma and associated POS-tag occurring in any sentence s of the document. In the second step, a word sense disambiguation (WSD) method is applied to each pair $\langle w, POS(w)\rangle$

to detect the most appropriate BabelNet synset σ_w for $\langle w, POS(w)\rangle$ contextually to s. The WSD algorithm is carried out in such a way that all words from all languages are disambiguated over the same concept space, producing a language-independent feature space for the whole multilingual corpus. Each document is finally modeled as a $|\mathcal{BS}|$-dimensional vector of BabelNet synset frequencies, being $\mathcal{BS}$ the set of retrieved BabelNet synsets.

As previously discussed in Section 2, BabelNet provides WSD algorithms for multilingual corpora. The authors in [15] suggest to use the degree ranking algorithm (i.e., given a semantic graph for the input context, it simply selects the sense of the target word with the highest vertex degree), as it has shown to yield highly competitive performance in the multilingual context. Clearly, other methods for (unsupervised) WSD, particularly PageRank-style methods (e.g., [12, 21]), can be plugged in to perform multilingual WSD based on BabelNet; however, this subject is out of the scope of this paper.

Bag-of-words and machine-translation based models. The *bag-of-words* model has been employed also in the context of multilingual documents [11]. Hereinafter we use notation *BoW* to refer to the term-frequency vector representation of documents over the union of language-specific term vocabularies.

However, in the multilingual setting, the use of *BoW* poses additional issues as it tends to exacerbate the sparsity in the document modeling, i.e., the language-specific vocabularies are generally very different, making the cross-lingual document similarity hard to compute. To overcome this issue, a common solution adopted in the literature is to translate all documents to a unique anchor language and represent the translated documents with the *BoW* model [11, 6]. In this work, we have considered three settings corresponding to the use of *English*, *French* or *Italian* as anchor language; the resulting representation models will be referred to as *BoW-MT-en*, *BoW-MT-fr* and *BoW-MT-it*, respectively. As an alternative model, we resort to a dimensionality reduction approach via Latent Semantic Indexing (LSI) [4] over the *BoW* representation. Recall that, given the document-term matrix obtained using *BoW*, LSI consists in computing the SVD decomposition of that matrix and representing the documents with low-dimensional vectors. We will refer to this model as *BoW-LSI*.

3.2 Transductive setting and label propagation algorithm

Given a document collection $\mathcal{D} = \{d_i\}_{i=1}^{N}$, let us denote with $\mathcal{L}$ the subset of $\mathcal{D}$ comprised of labeled documents, and with $\mathcal{U} = \mathcal{D} \setminus \mathcal{L}$ the subset of unlabeled documents. Note that $\mathcal{U}$ can in principle have any proportion w.r.t. $\mathcal{L}$, but in many real cases $\mathcal{U}$ is much larger than $\mathcal{L}$. Every document in $\mathcal{L}$ is assigned a label that refers to one of the known M classes $\mathcal{C} = \{C_j\}_{j=1}^{M}$. We also denote with $\mathbf{Y}$ a $N \times M$ matrix such that $\mathbf{Y}_{ij} = 1$ if C_j is the label assigned to document d_i, 0 otherwise.

The goal of a *transductive learner* is to make an inference "from particular to particular", i.e., given the classifications of the instances in the training set $\mathcal{L}$,

it aims to guess the classifications of the instances in the test set $\mathcal{U}$, rather than inducing a general rule that works out for classifying new unseen instances [19]. Transduction is naturally related to the class of case-based learning algorithms, whose most well-known algorithm is the k-nearest neighbor (kNN) [8].

To the best of our knowledge, we bring for the first time a transductive learning approach to a multilingual document classification. We use a particularly effective transductive learner, named Robust Multi-class Graph Transduction ($RMGT$) approach [10]. RMGT has shown to outperform all the other state-of-the-art transductive classifiers in the recent evaluation study by Sousa et al. [3]. Essentially, 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. In the following we describe in detail the mathematics behind RMGT.

Let $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, w \rangle$ be an undirected graph whose vertex set is $\mathcal{V} = \mathcal{D}$, edge set is $\mathcal{E} = \{(d_i, d_j) | d_i, d_j \in \mathcal{D} \wedge sim(d_i, d_j) > 0\}$, and edge weighting function is $w = sim(d_i, d_j)$. Given a positive integer k, consider the kNN graph $\mathcal{G}_k = \langle \mathcal{V}, \mathcal{E}_k, w \rangle$ derived from $\mathcal{G}$ and such that $\mathcal{E} = \{(d_i, d_j) | d_j \in N_i\}$, where N_i denotes the set of d_i's k-nearest neighbors. A weighted sparse matrix is obtained as $\mathbf{W} = \mathbf{A} + \mathbf{A}^{\mathrm{T}}$, where $\mathbf{A}$ is the weighted adjacency matrix of $\mathcal{G}_k$ and $\mathbf{A}^{\mathrm{T}}$ is the transpose of $\mathbf{A}$; the matrix $\mathbf{W}$ represents a *symmetry-favored* kNN graph [10]. Moreover, let $\mathbf{L} = \mathbf{I}_N - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ the normalized Laplacian of $\mathbf{W}$, where $\mathbf{I}_N$ is the $N \times N$ identity matrix and $\mathbf{D} = diag(\mathbf{W}\mathbf{1}_N)$. Without loss of generality, we can rewrite $\mathbf{L}$ and $\mathbf{W}$ as subdivided into four and two submatrices, respectively:

$$\mathbf{L} = \begin{bmatrix} \mathbf{\Delta}_{\mathcal{L}\mathcal{L}} & \mathbf{\Delta}_{\mathcal{L}\mathcal{U}} \\ \mathbf{\Delta}_{\mathcal{U}\mathcal{L}} & \mathbf{\Delta}_{\mathcal{U}\mathcal{U}} \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} \mathbf{Y}_{\mathcal{L}} \\ \mathbf{Y}_{\mathcal{U}} \end{bmatrix} \tag{1}$$

where $\mathbf{\Delta}_{\mathcal{L}\mathcal{L}}$ and $\mathbf{Y}_{\mathcal{L}}$ are the submatrices of $\mathbf{L}$ and $\mathbf{Y}$, respectively, corresponding to the labeled documents, and analogously for the other submatrices. The RMGT learning algorithm finally yields a matrix $\mathbf{F} \in \mathbb{R}^{N \times M}$ defined as:

$$\mathbf{F} = -\mathbf{\Delta}_{\mathcal{U}\mathcal{U}}^{-1}\mathbf{\Delta}_{\mathcal{U}\mathcal{L}}\mathbf{Y}_{\mathcal{L}} + \frac{\mathbf{\Delta}_{\mathcal{U}\mathcal{U}}^{-1}\mathbf{1}_u}{\mathbf{1}_u^{\mathrm{T}}\mathbf{\Delta}_{\mathcal{U}\mathcal{U}}^{-1}\mathbf{1}_u}(N\omega - \mathbf{1}_l^{\mathrm{T}}\mathbf{Y}_{\mathcal{L}} + \mathbf{1}_u^{\mathrm{T}}\mathbf{\Delta}_{\mathcal{U}\mathcal{U}}^{-1}\mathbf{\Delta}_{\mathcal{U}\mathcal{L}}\mathbf{Y}_{\mathcal{L}}) \tag{2}$$

where $\omega \in \mathbb{R}^M$ is the class prior probabilities.

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. The mathematical formulation given in Eq. (2) enables a closed form solution of this optimization problem. After the propagation step, every unlabeled document d_i is associated to a vector (i.e., the i-th row of $\mathbf{F}$) 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.

Algorithm 1 sketches the main steps of our multilingual document classification framework based on the RMGT learning approach. Initially, a pre-processing step is required to model every document in the collection using our proposed

Algorithm 1 Transductive classification of multilingual documents

Input: A collection of multilingual documents $\mathcal{D}$, with labeled documents $\mathcal{L}$ and unlabeled documents $\mathcal{U}$ (with $\mathcal{D} = \mathcal{L} \cup \mathcal{U}$ and $\mathcal{L} \cap \mathcal{U} = \emptyset$); a set of labels $\mathcal{C} = \{C_j\}_{j=1}^{M}$ assigned to the documents in $\mathcal{L}$; a positive integer k for the neighborhood selection.
Output: A classification over $\mathcal{C}$ for the documents in $\mathcal{U}$.
1: Model each document in $\mathcal{D}$ using *BoS* or alternative representations. */* Section 3.1 */*
2: Build the similarity graph $\mathcal{G}$ for the document collection $\mathcal{D}$.
3: Extract the k-nearest neighbor graph $\mathcal{G}_k$ from $\mathcal{G}$. */* Section 3.2 */*
4: Build the matrix $\mathbf{W}$ from $\mathcal{G}_k$, which represents the symmetry-favored k-nearest neighbor graph. */* Section 3.2 */*
5: Compute the normalized Laplacian of $\mathbf{W}$. */* Section 3.2 */*
6: Compute the *RMGT* solution $\mathbf{F}$. */* Eq. (2) */*
7: Assign document $d_i \in \mathcal{U}$ to the class C_{j^*} that maximizes the class likelihood, $j^* = \arg\max_j \mathbf{F}_{ij}$.

BoS representation or alternative representations (Line 1). Upon the computation of the similarity matrix over all documents in the collection (Line 2), the graph-based label propagation process requires the construction of the kNN graph (Line 3) and its symmetry-favored transformation (Line 4). Concerning the $sim(\cdot, \cdot)$ function, we employ the cosine similarity as standard measure in document classification, but other measures can alternatively be utilized. Moreover, the class priors (ω) used in Eq. (2) are defined as uniformly distributed.

4 Experimental Evaluation

4.1 Data

We used two document collections, built from the *RCV2* corpus[4] and from the *Wikipedia* online encyclopedia. Both datasets were constructed to contain documents in three different languages, namely *English*, *French*, and *Italian*. Six topic-classes were identified, which correspond to selected values of TOPICS Reuters field in RCV2 and to selected Wikipage titles in Wikipedia. Our choice of languages and topics allowed us to obtain a significant topical coverage in all languages. Moreover, according to [17], we considered a *balanced* way for the document assignment to each topic-language pair; specifically, 850 and 1000 documents per pair, in RCV2 and Wikipedia, respectively. RCV2 contains 15 300 documents represented over a space of 12 698 terms, for the *BoW* model, and 10 033 synsets, for the *BoS* model, with density (i.e., the fraction of non-zero entries in the document-term matrix, resp. document-synset matrix) of 4.56E-3 for *BoW* and 3.87E-3 for *BoS*. Wikipedia is comprised of 18 000 documents, with 15 634 terms and 10 247 synsets, and density of 1.61E-2 for *BoW* and 1.81E-2 for *BoS*.[5]

Note that although the two datasets were built using the same number of languages and topics, they can be distinguished by an important aspect: in RCV2, the different language-written documents belonging to the same topic-class do not necessarily share the content subjects; by contrast, the encyclopedic nature of Wikipedia favors a closer correspondence in content among the different

[4] http://trec.nist.gov/data/reuters/reuters.html.
[5] Datasets are made publicly available at http://uweb.dimes.unical.it/tagarelli/data/.

language-specific versions of articles discussing the same Wikipedia concept (although, these versions are not translation of each other). We underline that both corpora have not been previously used in a multilingual transductive scenario.

Every document was subject to lemmatization and, in the *BoS* case, to POS-tagging as well. All text processing steps were performed using the Freeling library tool.[6] To setup the transductive learner, we used $k = 10$ for the kNN graph construction, and we evaluated the classification performance by varying the percentage of labeled documents from 1% to 20% with a step of 1% for both datasets. Results were averaged over 30 runs (to avoid sampling bias) and assessed by using standard F-measure, Precision and Recall criteria [11].

4.2 Evaluation of BabelNet coverage

The extent to which our approach will actually lead to good solutions depends on the semantic coverage capabilities of the multilingual knowledge base as well as on the corpus characteristics. Therefore, we initially investigated how well BabelNet allows us to represent the concepts discussed in each of the datasets.

For every document, we calculated the BabelNet coverage as the fraction of words belonging to the document whose concepts are present as entries in BabelNet. We then analyzed the distribution of documents over different values of BabelNet coverage. Figures 1(a)–1(b) show the probability density function (pdf) of BabelNet coverage for each of the topic-classes, on RCV2 and Wikipedia, respectively; analogously, Figs. 1(c)–1(d) visualize the distributions per language.

Generally, we observe roughly bi-modal distributions in both evaluation cases and for both datasets. Considering the per-topic distributions, all of them tend to have a peak around coverage of 0.5 and a lower peak around 0.84, following the overall trend with no evident distinctions. By contrast, the per-language distributions (Fig. 1(c)–1(d)) supply more helpful clues to understand the BabelNet coverage capabilities. In fact we observe that both French and Italian documents determine the left peak of the overall distributions, actually corresponding to roughly normal distributions; on the contrary, the English documents correspond to negatively skewed (i.e., left-tailed) distributions, thus characterizing the right peak of the overall distributions. Interestingly, these remarks hold for both RCV2 and Wikipedia datasets, which indicates that BabelNet provides a more complete coverage for English documents than for French/Italian documents.

4.3 Classification performance

In this section we assess the impact of *BoS* and the other document models on the performance of our transductive multilingual classification approach. In order to inspect the models' behavior under different corpus characteristics, in this stage of evaluation we also produced unbalanced versions of the datasets, hereinafter referred to *unbalanced RCV2* and *unbalanced Wikipedia*. Specifically, in each of the two original datasets we kept the subset of English documents

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

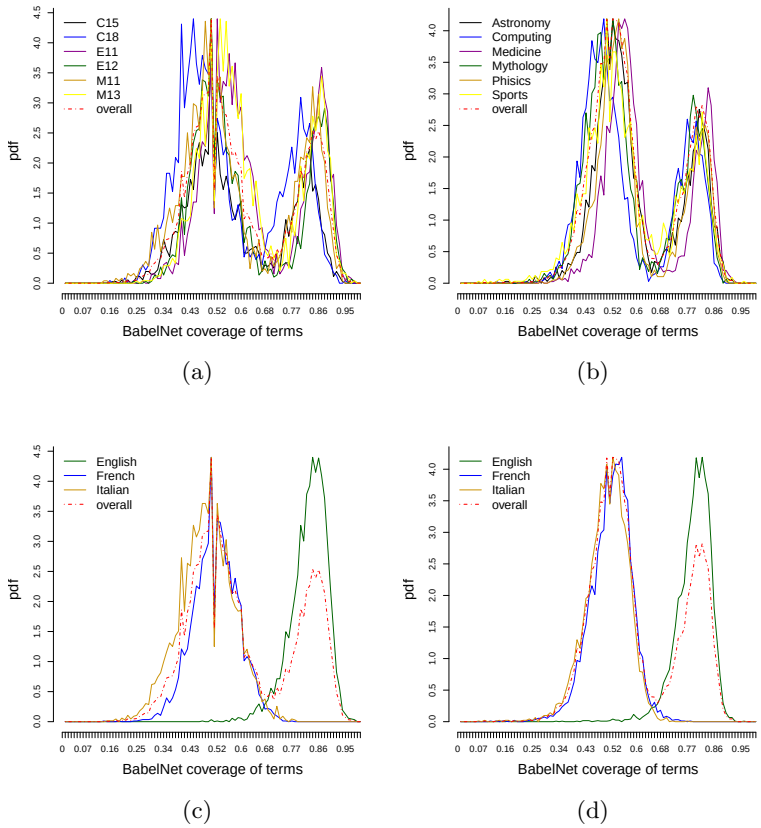


Fig. 1. BabelNet coverage: (a) RCV2, (b) Wikipedia per topic-class, and (c) RCV2, (d) Wikipedia per language. (Better viewed in the electronic version.)

while sampling half of the French and half of the Italian subsets. In the light of the remarks that stand out from the previous analysis on the BabelNet coverage, we aim here to understand how much the classification performance varies when using an English-biased multilingual corpus.

In the following, we present results obtained in the two distinguished cases of balanced and unbalanced datasets. Figure 2 shows the methods' performance (F-measure) by varying the training percentage of the transductive learning algorithm, while Table 1 summarizes the best performances in terms of F-measure, Precision and Recall. We begin with evaluation on the balanced case, which we then couple with an inspection of the intra-class and inter-class similarity of the datasets. This will allow us to unveil important aspects of the behaviors of the *BoS* model and competing ones that eventually advocate the significance of our further evaluation on unbalanced datasets.

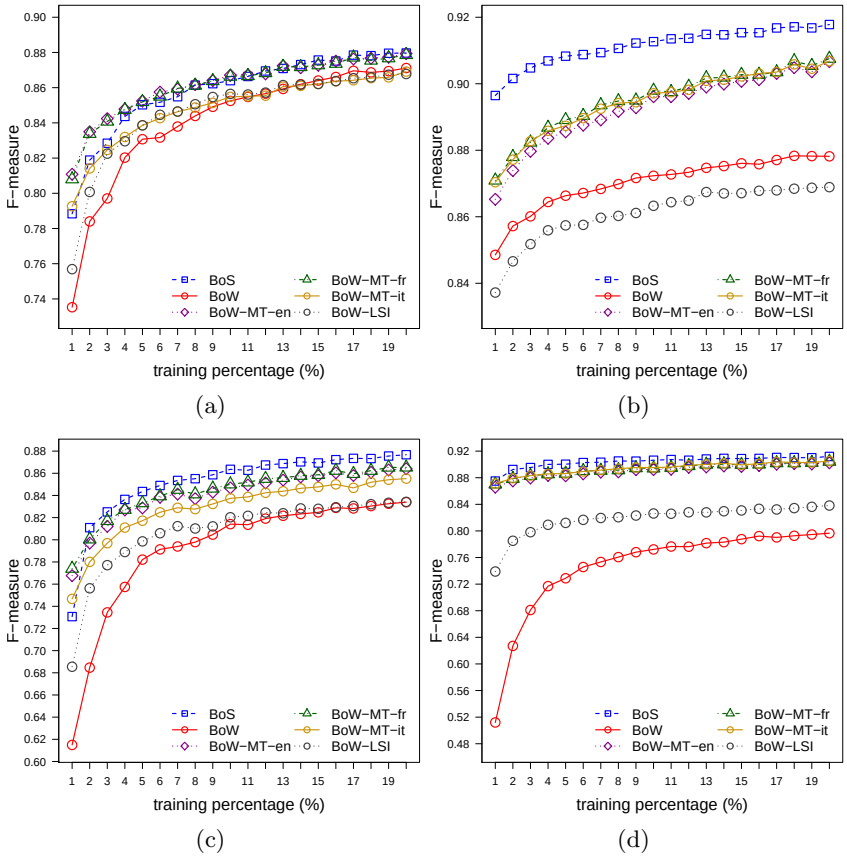


Fig. 2. F-measure for (a) RCV2, (b) Wikipedia, and (c) unbalanced RCV2, (d) unbalanced Wikipedia. (Better viewed in the electronic version.)

Evaluation on language-balanced corpora. On RCV2 (Fig. 2(a)), we observe that *BoS* follows an increasing trend, similarly to those shown by the other models and performing (for training percentage values above 4%) comparably to the best of the competing models, which are *BoW-MT-en* and *BoW-MT-fr*. The *BoW-MT-it* and *BoW-LSI* achieve lower F-measures, which become very close to the basic *BoW* for higher values of training percentage.

A different scenario is instead depicted in Fig. 2(b) for Wikipedia. *BoS* clearly outperforms the other document representation models, including *BoW-MT-en* which in this case achieves results that are similar to (or slightly lower than) those obtained by *BoW-MT-fr* and *BoW-MT-it*. *BoW-LSI* and *BoW* also show a performance gap from the other models, which is much more significant than in the RCV2 case.

As a general remark it should be noted that *BoS* not only performs comparably or significantly better than the other models—this is confirmed by the

Table 1. Summary of best performance results of the various representation methods. Bold values correspond to the best performance per dataset and assessment criterion, whereas italic values refer to *BoW* related methods.

	Balanced RCV2			*Balanced Wikipedia*			*Unbalanced RCV2*			*Unbalanced Wikipedia*		
	FM	*P*	*R*	*FM*	*P*	*R*	*FM*	*P*	*R*	*FM*	*P*	*R*
BoS	**0.880**	**0.883**	**0.881**	**0.912**	**0.915**	**0.912**	**0.877**	**0.880**	**0.878**	**0.912**	**0.915**	**0.912**
BoW	0.871	0.876	0.872	0.872	0.876	0.872	0.834	0.839	0.836	0.797	0.817	0.794
BoW-MT-en	*0.879*	*0.881*	*0.880*	0.895	0.896	0.895	0.864	*0.867*	0.865	0.902	0.903	0.902
BoW-MT-fr	*0.879*	0.879	0.879	*0.898*	*0.899*	*0.898*	*0.865*	0.866	*0.866*	0.904	0.906	0.904
BoW-MT-it	0.869	0.870	0.870	0.897	*0.899*	0.897	0.855	0.856	0.856	*0.905*	*0.907*	*0.905*
BoW-LSI	0.868	0.872	0.869	0.863	0.867	0.863	0.834	0.840	0.837	0.838	0.845	0.838

best-performance evaluation reported in Table 1—but also it exhibits a performance trend that is not affected by issues related to the language specificity. In fact, the machine-translation based models have relative performance that may vary on different datasets, since a language that leads to better results on a dataset can perform worse than other languages on another dataset.

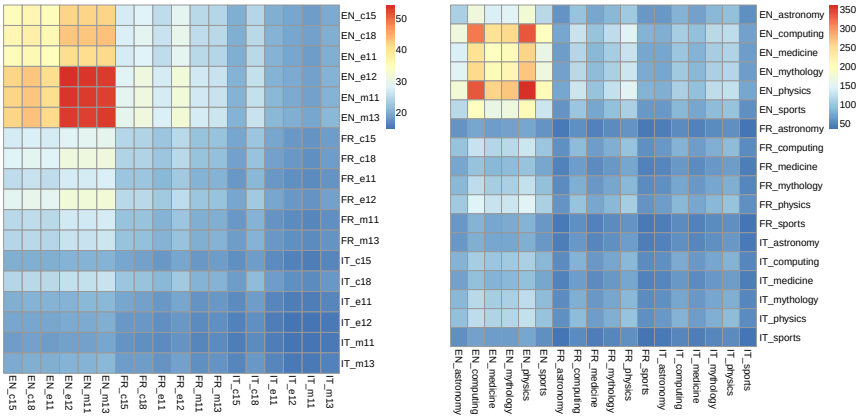
Intra-class and inter-class document similarity. The differences observed in the relative trends exhibited by *BoS* and the other models on RCV2 compared with Wikipedia, prompted us to investigate the topic homogeneity and topic separation on the datasets, over the various topic-classes and languages.

Figure 3 compares the similarity matrices for the balanced datasets obtained using *BoS*. Note that the main diagonal on each matrix corresponds to the intra-class document similarity, while the remaining cells refer to similarity between two different topic-classes (i.e., inter-class similarity). On every cell, the hue toward red (resp. blue) indicates higher (resp. lower) cosine similarity.

A first remark common to RCV2 and Wikipedia (Fig. 3(a)–(b)) is that both the intra-class and inter-class is low when only French and Italian documents are considered. This might be explained by a different support of *BabelNet* to the conceptual representation of documents in non-English languages; in particular, as discussed in [17], French and Italian documents have a significantly lower dimensional representation according to the *BoS* model, which would hence affect both intra- and inter-class document similarities.

Looking at the upper left blocks of the matrices, which correspond to English document classes, we observe that on RCV2 the intra-class similarity is high for three topics (i.e., "E12", "M11", "M13"), and, in general, higher than on Wikipedia; however, also the inter-class similarity is higher (i.e., worse) than on Wikipedia. The topic separation between English and French/Italian documents is lower on RCV2. The above findings would indicate that RCV2 appears to be a harder testbed than Wikipedia for our proposed *BoS* model.

Evaluation on language-unbalanced corpora. Here we quantify how the methods' performance change when the English written portion in the corpus varies (i.e., is double) relatively to the other languages' portions. Figure 2(c)



(a) (b)

Fig. 3. Similarity matrices of *BoS*-modeled documents grouped by language and class for balanced datasets: (a) RCV2 and (b) Wikipedia. (Better viewed in the electronic version.)

and Table 1 show that the *BoS* results are always higher (though slightly) than the best competing methods. More interestingly, the advantage taken by *BoS* is actually explained by a decreased performance of the other models, which would indicate a higher robustness of the *BoS* model w.r.t. the corpus characteristics.

Note also that on Wikipedia (Fig. 2(d)), the relative performance between *BoS* and the machine-translation based models is not changed w.r.t. the balanced case, and the finer scale-grain of the y-axis gives evidence of the decreased performance of *BoW* and *BoW-LSI*.

5 Conclusion

We have proposed a new framework for multilingual document classification under a transductive setting and with the support of the BabelNet knowledge base. Our proposed conceptual representation model for multilingual documents, *BoS*, has shown to be effective for multilingual comparable corpora: *BoS* not only leads to generally better results than various language-dependent representations, but it has also shown to preserve its performance on both balanced and unbalanced datasets. This aspect highlights the robustness of our knowledge-based representation, paving the way for future analysis of multilingual documents. Furthermore, the transductive learning approach has shown to be useful in the multilingual scenario, obtaining good classification performance with a quite small (5%) portion of labeled documents.

As future work we plan to exploit more types of information provided in BabelNet (i.e., relations among the synsets) to enrich our multilingual document model. We are also interested in combining transductive with active learning, which can aid solicit user interaction in order to guide the labeling process.

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 Proc. *ECIR*, pp. 424–429, 2014.
3. C. A. R. de Sousa, S. O. Rezende, and G. Batista. Influence of graph construction on semi-supervised learning. In Proc. *ECML-PKDD*, pp. 160–175, 2013.
4. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.
5. M. Franco-Salvador, P. Rosso, and R. Navigli. A knowledge-based representation for cross-language document retrieval and categorization. In Proc. *EACL*, pp. 414–423, 2014.
6. Y. Guo and M. Xiao. Transductive representation learning for cross-lingual text classification. In Proc. *ICDM*, pp. 888–893, 2012.
7. T. Joachims. Transductive inference for text classification using support vector machines. In Proc. *ICML*, pp. 200–209, 1999.
8. T. Joachims. Transductive Learning via Spectral Graph Partitioning. In Proc. *ICML*, 2003.
9. A. Klementiev, I. Titov, and B. Bhattarai. Inducing Crosslingual Distributed Representations of Words. In Proc. *COLING*, pp. 1459–1474, 2012.
10. W. Liu and S. Chang. Robust multi-class transductive learning with graphs. In Proc. *CVPR*, pp. 381–388, 2009.
11. C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.
12. R. Mihalcea, P. Tarau, and E. Figa. PageRank on semantic networks, with application to word sense disambiguation. In Proc. *COLING*, 2004.
13. R. Navigli and M. Lapata. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE TPAMI*, 32(4):678–692, 2010.
14. 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.
15. R. Navigli and S. P. Ponzetto. Multilingual WSD with just a few lines of code: the babelnet API. In Proc. *ACL*, pp. 67–72, 2012.
16. X. Ni, J. Sun, J. Hu, and Z. Chen. Cross lingual text classification by mining multilingual topics from wikipedia. In Proc. *WSDM*, pp. 375–384, 2011.
17. S. Romeo, A. Tagarelli, and D. Ienco. Semantic-Based Multilingual Document Clustering via Tensor Modeling. In Proc. *EMNLP*, pp. 600–609, 2014.
18. R. Steinberger, B. Pouliquen, and J. Hagman. Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC. In Proc. *CICLing*, pp. 415–424, 2002.
19. V. Vapnik. *Statistical learning theory*. Wiley, 1998.
20. 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.
21. E. Yeh, D. Ramage, C. D. Manning, E. Agirre, and A. Soroa. Wikiwalk: Random walks on wikipedia for semantic relatedness. In *Workshop on Graph-based Methods for Natural Language Processing*, pp. 41–49, 2009.