A Knowledge-based Representation for Cross-Language Document Retrieval and Categorization

Marc Franco-Salvador[1,2], **Paolo Rosso**[2] and **Roberto Navigli**[1]
[1] Department of Computer Science
Sapienza Università di Roma, Italy
{francosalvador,navigli}@di.uniroma1.it
[2] Natural Language Engineering Lab - PRHLT Research Center
Universitat Politècnica de València, Spain
{mfranco,prosso}@dsic.upv.es

Abstract

Current approaches to cross-language document retrieval and categorization are based on discriminative methods which represent documents in a low-dimensional vector space. In this paper we propose a shift from the supervised to the knowledge-based paradigm and provide a document similarity measure which draws on BabelNet, a large multilingual knowledge resource. Our experiments show state-of-the-art results in cross-lingual document retrieval and categorization.

1 Introduction

The huge amount of text that is available online is becoming ever increasingly multilingual, providing an additional wealth of useful information. Most of this information, however, is not easily accessible to the majority of users because of language barriers which hamper the cross-lingual search and retrieval of knowledge.

Today's search engines would benefit greatly from effective techniques for the cross-lingual retrieval of valuable information that can satisfy a user's needs by not only providing (Landauer and Littman, 1994) and translating (Munteanu and Marcu, 2005) relevant results into different languages, but also by reranking the results in a language of interest on the basis of the importance of search results in other languages.

Vector-based models are typically used in the literature for representing documents both in monolingual and cross-lingual settings (Manning et al., 2008). However, because of the large size of the vocabulary, having each term as a component of the vector makes the document representation very sparse. To address this issue several approaches to dimensionality reduction have been proposed, such as Principal Component Analysis (Jolliffe, 1986), Latent Semantic Indexing (Hull, 1994), Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and variants thereof, which project these vectors into a lower-dimensional vector space. In order to enable multilinguality, the vectors of comparable documents written in different languages are concatenated, making up the document matrix which is then reduced using linear projection (Platt et al., 2010; Yih et al., 2011). However, to do so, comparable documents are needed as training. Additionally, the lower dimensional representations are not of easy interpretation.

The availability of wide-coverage lexical knowledge resources extracted automatically from Wikipedia, such as DBPedia (Bizer et al., 2009), YAGO (Hoffart et al., 2013) and BabelNet (Navigli and Ponzetto, 2012a), has considerably boosted research in several areas, especially where multilinguality is a concern (Hovy et al., 2013). Among these latter are cross-language plagiarism detection (Potthast et al., 2011; Franco-Salvador et al., 2013), multilingual semantic relatedness (Navigli and Ponzetto, 2012b; Nastase and Strube, 2013) and semantic alignment (Navigli and Ponzetto, 2012a; Matuschek and Gurevych, 2013). One main advantage of knowledge-based methods is that they provide a human-readable, semantically interconnected, representation of the textual item at hand (be it a sentence or a document).

Following this trend, in this paper we provide a knowledge-based representation of documents which goes beyond the lexical surface of text, while at the same time avoiding the need for training in a cross-language setting. To achieve this we leverage a multilingual semantic network, i.e., BabelNet, to obtain language-independent representations, which contain concepts together with semantic relations between them, and also include semantic knowledge which is just implied by the input text. The integration of our multilingual graph model with a vector representation enables us to obtain state-of-the-art results in comparable

Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 414–423,
Gothenburg, Sweden, April 26-30 2014.

document retrieval and cross-language text categorization.

2 Related Work

The mainstream representation of documents for monolingual and cross-lingual document retrieval is vector-based. A document vector, whose components quantify the relevance of each term in the document, is usually highly dimensional, because of the variety of terms used in a document collection. As a consequence, the resulting document matrices are very sparse. To address the data sparsity issue, several approaches to the reduction of dimensionality of document vectors have been proposed in the literature. A popular class of methods is based on linear projection, which provides a low-dimensional mapping from a high dimensional vector space. A historical approach to linear projection is Principal Component Analysis (PCA) (Jolliffe, 1986), which performs a singular value decomposition (SVD) on a document matrix D of size $n \times m$, where each row in D is the term vector representation of a document. PCA uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components, which make up the low-dimensional vector. Latent Semantic Analysis (LSA) (Deerwester et al., 1990) is very similar to PCA but performs the SVD using the correlation matrix instead of the covariance matrix, which implies a lower computational cost. LSA preserves the amount of variance in an eigenvector $\vec{v}$ by maximizing its Rayleigh ratio: $\frac{\vec{v}^T C \vec{v}}{\vec{v}^T \vec{v}}$, where $C = D^T D$ is the correlation matrix of D.

A generalization of PCA, called Oriented Principal Component Analysis (OPCA) (Diamantaras and Kung, 1996), is based on a noise covariance matrix to project the similar components of D closely. Other projection models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003) are based on the extraction of generative models from documents. Another approach, named Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007), represents each document by its similarities to a document collection. Using a low domain specificity document collection such as Wikipedia, the model has proven to obtain competitive results.

Not only have these methods proven to be successful in a monolingual scenario (Deerwester et al., 1990; Hull, 1994), but they have also been adapted to perform well in tasks at a cross-language level (Potthast et al., 2008; Platt et al., 2010; Yih et al., 2011). Cross-language Latent Semantic Indexing (CL-LSI) (Dumais et al., 1997) was the first linear projection approach used in cross-lingual tasks. CL-LSI provides a cross-lingual representation for documents by reducing the dimensionality of a matrix D whose rows are obtained by concatenating comparable documents from different languages. Similarly, PCA and OPCA can be adapted to a multilingual setting. LDA was also adapted to perform in a multilingual scenario with models such as Polylingual Topic Models (Mimno et al., 2009), Joint Probabilistic LSA and Coupled Probabilistic LSA (Platt et al., 2010), which, however, are constrained to using word counts, instead of better weighting strategies, such as log(tf)-idf, known to perform better with large vocabularies (Salton and McGill, 1986). Another variant, named Canonical Correlation Analysis (CCA) (Thompson, 2005), uses a cross-covariance matrix of the low-dimensional vectors to find the projections. Cross-language Explicit Semantic Analysis (CL-ESA) (Potthast et al., 2008; Cimiano et al., 2009; Potthast et al., 2011), instead, adapts ESA to be used at cross-language level by exploiting the comparable documents across languages from Wikipedia. CL-ESA represents each document written in a language L by its similarities with a document collection in the same language L. Using a multilingual document collection with comparable documents across languages, the resulting vectors from different languages can be compared directly.

An alternative unsupervised approach, Cross-language Character n-Grams (CL-CNG) (Mcnamee and Mayfield, 2004), does not draw upon linear projections and represents documents as vectors of character n-grams. It has proven to obtain good results in cross-language document retrieval (Potthast et al., 2011) between languages with lexical and syntactic similarities.

Recently, a novel supervised linear projection model based on Siamese Neural Networks (S2Net) (Yih et al., 2011) achieved state-of-the-art performance in comparable document retrieval. S2Net performs a linear combination of the terms of a document vector $\vec{d}$ to obtain a reduced vector $\vec{r}$, which is the output layer of a neural network. Each element in $\vec{r}$ has a weight which is a linear combination of the original weights of $\vec{d}$, and captures relationships between the original terms.

However, linear projection approaches need a high number of training documents to achieve state-of-the-art performance (Platt et al., 2010; Yih et al., 2011). Moreover, although they are good at identifying a few principal components,

the representations produced are opaque, in that they cannot explicitly model the semantic content of documents with a human-interpretable representation, thereby making the data analysis difficult. In this paper, instead, we propose a language-independent knowledge graph representation for documents which is obtained from a large multilingual semantic network, without using any training information. Our knowledge graph representation explicitly models the semantics of the document in terms of the concepts and relations evoked by its co-occurring terms.

3 A Knowledge-based Document Representation

We propose a knowledge-based document representation aimed at expanding the terms in a document's bag of words by means of a knowledge graph which provides concepts and semantic relations between them. Key to our approach is the use of a graph representation which does not depend on any given language, but, indeed, is multilingual. To build knowledge graphs of this kind we utilize BabelNet, a multilingual semantic network that we present in Section 3.1. Then, in Section 3.2, we describe the five steps needed to obtain our graph-based multilingual representation of documents. Finally, we introduce our knowledge graph similarity measure in Section 3.3.

3.1 BabelNet

BabelNet (Navigli and Ponzetto, 2012a) is a multilingual semantic network whose concepts and relations are obtained from the largest available semantic lexicon of English, WordNet (Fellbaum, 1998), and the largest wide-coverage collaboratively-edited encyclopedia, Wikipedia, by means of an automatic mapping algorithm. BabelNet is therefore a multilingual "encyclopedic dictionary" that combines lexicographic information with wide-coverage encyclopedic knowledge. Concepts in BabelNet are represented similarly to WordNet, i.e., by grouping sets of synonyms in the different languages into multilingual synsets. Multilingual synsets contain lexicalizations from WordNet synsets, the corresponding Wikipedia pages and additional translations output by a statistical machine translation system. The relations between synsets are collected from WordNet and from Wikipedia's hyperlinks between pages.

We note that, in principle, we could use any multilingual network providing a similar kind of information, e.g., EuroWordNet (Vossen, 2004). However, in our work we chose BabelNet because of its larger size, its coverage of both lexicographic and encyclopedic knowledge, and its free availability.[1] In our work we used BabelNet 1.0, which encodes knowledge for six languages, namely: Catalan, English, French, German, Italian and Spanish.

3.2 From Document to Knowledge Graph

We now introduce our five-step method for representing a given document d from a collection D of documents written in language L as a language-independent knowledge graph.

Building a Basic Vector Representation Initially we transform a document d into a traditional vector representation. To do this, we score each term $t_i \in d$ with a weight w_i. This weight is usually a function of term and document frequency. Following the literature, one method that works well is the log tf-idf weighting (Salton et al., 1983; Salton and McGill, 1986):

$$w_i = log_2(f_i + 1)log_2(n/n_i). \quad (1)$$

where f_i is the number of times term i occurs in document d, n is the total number of documents in the collection and n_i is the number of documents that contain t_i. We then create a weighted term vector $\vec{v} = (w_1, ..., w_n)$, where w_i is the weight corresponding to term t_i. We exclude stopwords from the vector.

Selecting the Relevant Document Terms We then create the set T of base forms, i.e., lemmas[2], of the terms in the document d. In order to keep only the most relevant terms, we sort the terms T according to their weight in vector $\vec{v}$ and retain a maximum number of K terms, obtaining a set of terms T_K.[3] The value of K is calculated as a function of the vector size, as follows:

$$K = (log_2(1 + |\vec{v}|))^2, \quad (2)$$

The rationale is that K must be high enough to ensure a good conceptual representation but not too high, so as to avoid as much noise as possible in the set T_K.

Populating the Graph with Initial Concepts Next, we create an initially-empty knowledge graph $G = (V, E)$, i.e., such that $V = E = \emptyset$.

We populate the vertex set V with the set S_K of all the synsets in BabelNet which contain any term in T_K in the document language L, that is:

[1] http://babelnet.org

[2] Following the setup of (Platt et al., 2010), our initial data is represented using term vectors. For this reason we lemmatize in this step.

[3] Since the vector $\vec{v}$ provides weights for all the word forms, and not only lemmas, occurring in d, we take the best weight among those word forms of the considered lemma.

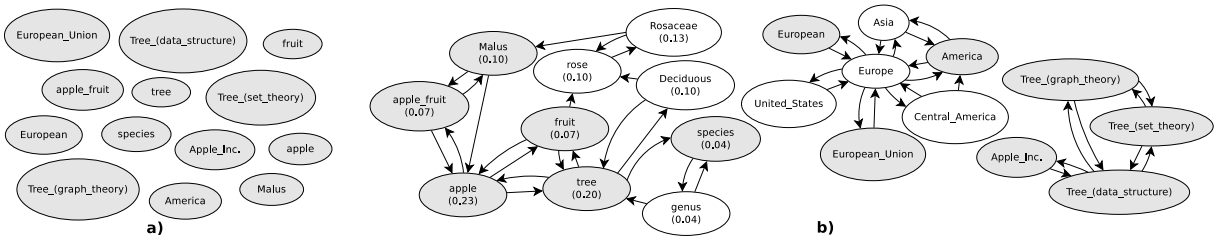

Figure 1: (a) initial graph from $T_K = \{$"European", "apple", "tree", "Malus", "species", "America"$\}$; (b) knowledge graph obtained by retrieving all paths from BabelNet. Gray nodes are the original concepts.

$$S_K = \bigcup_{t \in T_K} Synsets_L(t), \tag{3}$$

where $Synsets_L(t)$ is the set of synsets in BabelNet which contain a term t in the language of interest L. For example, in Figure 1(a) we show the initial graph obtained from the set $T_K = \{$"European", "apple", "tree", "Malus", "species", "America"$\}$. Note, however, that each retrieved synset is multilingual, i.e., it contains lexicalizations for the same concept in other languages too. Therefore, the nodes of our knowledge graph provide a language-independent representation of the document's content.

Creating the Knowledge Graph Similarly to Navigli and Lapata (2010), we create the knowledge graph by searching BabelNet for paths connecting pairs of synsets in V. Formally, for each pair $v, v' \in V$ such that v and v' do not share any lexicalization[4] in T_K, for each path in BabelNet $v \rightarrow v_1 \rightarrow \ldots \rightarrow v_n \rightarrow v'$, we set: $V := V \cup \{v_1, \ldots, v_n\}$ and $E := E \cup \{(v, v_1), \ldots, (v_n, v')\}$, that is, we add all the path vertices and edges to G. After prototyping, the path length is limited to maximum length 3, so as to avoid an excessive semantic drift.

As a result of populating the graph with intermediate edges and vertices, we obtain a knowledge graph which models the semantic context of document d. We point out that our knowledge graph might have different isolated components. We view each component as a different interpretation of document d. To select the main interpretation, we keep only the largest component, i.e., the one with the highest number of vertices, which we consider as the most likely semantic representation of the document content.

Figure 1(b) shows the knowledge graph obtained for our example term set. Note that our approach retains, and therefore weights, only the subgraph focused on the "apple fruit" meaning.

[4]This prevents different senses of the same term from being connected via a path in the resulting knowledge graph.

Knowledge Graph Weighting The final step consists of weighting all the concepts and semantic relations of the knowledge graph G. For weighting relations we use the original weights from BabelNet, which provide the degree of relatedness between the synset end points of each edge (Navigli and Ponzetto, 2012a). As for concepts, we weight them on the basis of the original weights of the terms in the vector $\vec{v}$. In order to score each concept in our knowledge graph G, we applied the topic-sensitive PageRank algorithm (Haveliwala et al., 2003) to G. While the well-known PageRank algorithm (Page et al., 1998) calculates the global importance of vertices in a graph, topic-sensitive PageRank is a variant in which the importance of vertices is biased using a set of representative "topics". Formally, the topic-sensitive PageRank vector $\vec{p}$ is calculated by means of an iterative process until convergence as follows: $\vec{p} = cM\vec{p} + (1-c)\vec{u}$, where c is the damping factor (conventionally set to 0.85), $1-c$ represents the probability of a surfer randomly jumping to any node in the graph, M is the transition probability matrix of graph G, with $M_{ji} = degree(i)^{-1}$ if an edge from i to j exists, 0 otherwise, $\vec{u}$ is the random-jumping transition probability vector, where each u_i represents the probability of jumping randomly to the node i, and $\vec{p}$ is the resulting PageRank vector which scores the nodes of G. In contrast to vanilla PageRank, the "topic-sensitive" variant gives more probability mass to some nodes in G and less to others. In our case we perturbate $\vec{u}$ by concentrating the probability mass to the vertices in S_K, which are the synsets corresponding to the document terms T_K (cf. Formula 3).

3.3 Similarity between Knowledge Graphs

We can now determine the similarity between two documents $d, d' \in D$ in terms of the similarity of their knowledge graph representations G and G'.

Following the literature (Montes y Gómez et al., 2001) we calculate the similarity between the vertex sets in the two graphs using Dice's coefficient (Jackson et al., 1989):

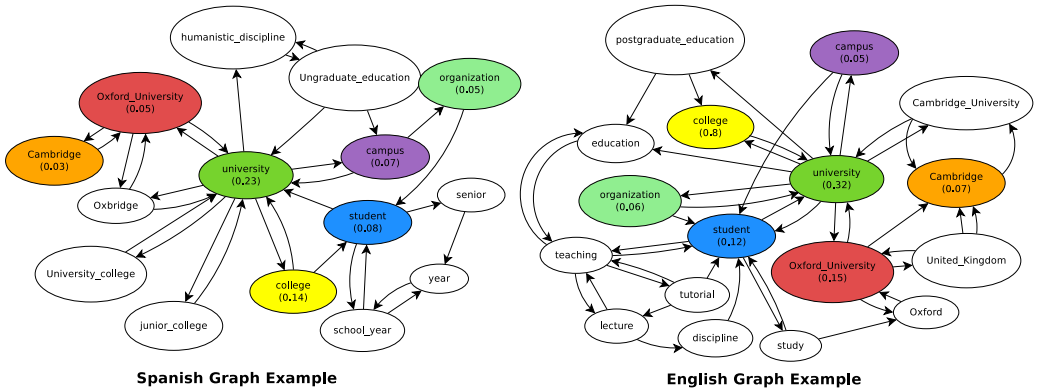

Figure 2: Knowledge graph examples from two comparable documents in different languages.

$$S_c(G, G') = \frac{2 \cdot \sum_{c \in V(G) \cap V(G')} w(c)}{\sum_{c \in V(G)} w(c) + \sum_{c \in V(G')} w(c)}, \quad (4)$$

where $w(c)$ is the weight of a concept c (see Section 3.2). Likewise, we calculate the similarity between the two edge sets as:

$$S_r(G, G') = \frac{2 \cdot \sum_{r \in E(G) \cap E(G')} w(r)}{\sum_{r \in E(G)} w(r) + \sum_{r \in E(G')} w(r)}, \quad (5)$$

where $w(r)$ is the weight of a semantic relation edge r.

We combine the two above measures of conceptual (S_c) and relational (S_r) similarity to obtain an integrated measure $S_g(G, G')$ between knowledge graphs:

$$S_g(G, G') = \frac{S_c(G, G') + S_r(G, G')}{2}. \quad (6)$$

Notably, since we are working with a language-independent representation of documents, this similarity measure can be applied to the knowledge graphs built from documents written in any language. In Figure 2 we show two knowledge graphs for comparable documents written in different languages (for clarity, labels are in English in both graphs). As expected, the graphs share several key concepts and relations.

4 A Multilingual Vector Representation

4.1 From Document to Multilingual Vector

Since our knowledge graphs will only cover the most central concepts of a document, we complement this core representation with a more traditional vector-based representation. However, as we are interested in the cross-language comparison of documents, we translate our monolingual vector $\vec{v}_L$ of a document d written in language L into its corresponding vector $\vec{v}_{L'}$ in language L' using BabelNet as our multilingual dictionary. We detail the document-vector translation process in Algorithm 1.

Algorithm 1 Dictionary-based term-vector translation.

Input: a weighted document vector $\vec{v}_L = (w_1, \ldots, w_n)$, a source language L and a target language L'
Output: a translated vector $\vec{v}_{L'}$

1: $\vec{v}_{L'} \leftarrow (0, \ldots, 0)$ of length n
2: **for** i = 1 to n
3: **if** $w_i = 0$ **continue**
4: // let t_i be the term corresponding to w_i in $\vec{v}_L$
5: $S_L \leftarrow \mathit{Synsets}_L(t_i)$
6: **for each** synset $s \in S_L$
7: $T \leftarrow \mathit{getTranslations}(s, L')$
8: **if** $T \neq \emptyset$ **then**
9: **for each** $tr \in T$
10: $w_{new} = w_i \cdot \mathit{confidence}(tr, t_i)$
11: // let $\mathit{index}(tr)$ be the index of tr in $\vec{v}_L$
12: **if** $\exists\ \mathit{index}(tr)$ **then**
13: $v_{L'}(\mathit{index}(tr)) = w_{new}$
14: **return** $\vec{v}_{L'}$

The translated vector $\vec{v}_{L'}$ is obtained as follows: for each term t_i with non-zero weight in v_L we obtain all the possible meanings of t_i in BabelNet (see line 5) and, for each of these, we retrieve all the translations (line 7), i.e., lexicalizations of the concept, in language L' available in the synset. We set a non-zero value in the translation vector $\vec{v}_{L'}$,[5] in correspondence with each such translation tr, proportional to the weight of t_i in the original vector and the confidence of the translation (line 10), as provided by the BabelNet semantic network.[6]

In order to increase the amount of information available in the vector and counterbalance possible wrong translations, we avoid translating all vectors to one language. Instead, in the present work we create a multilingual vector representation of a

[5] To make the translation possible, while at the same time keeping the same number of dimensions in our vector representation, we use a shared vocabulary which covers both languages. See Section 6 for details on the experimental setup.

[6] Non-English lexicalizations in BabelNet have confidence 1 if originating from Wikipedia inter-language links and ≤ 1 if obtained by means of statistical machine translation (Navigli and Ponzetto, 2012a).

document d written in language L by concatenating the corresponding vector $\vec{v}_L$ with the translated vector $\vec{v}_{L'}$ of d for language L'. As a result, we obtain a multilingual vector $\vec{v}_{LL'}$, which contains lexicalizations in both languages.

4.2 Similarity between Multilingual Vectors

Following common practice for document similarity in the literature (Manning et al., 2008), we use the cosine similarity as the similarity measure between multilingual vectors:

$$S_v(\vec{v}_{LL'}, \vec{v}'_{LL'}) = \frac{\vec{v}_{LL'} \cdot \vec{v}'_{LL'}}{||\vec{v}_{LL'}|| \, ||\vec{v}'_{LL'}||}. \quad (7)$$

5 Knowledge-based Document Similarity

Given a source document d and a target document d', we calculate the similarities between the respective knowledge-graph and multilingual vector representations, and combine them to obtain a knowledge-based similarity as follows:

$$KBSim(d, d') = c(G)S_g(G, G') + (1 - c(G))S_v(\vec{v}_{LL'}, \vec{v}'_{LL'}), \quad (8)$$

where $c(G)$ is an interpolation factor calculated as the edge density of knowledge graph G:

$$c(G) = \frac{|E(G)|}{|V(G)|(|V(G)| - 1)}. \quad (9)$$

Note that, using the factor $c(G)$ to interpolate the two similarities in Eq. 8, we determine the relevance for the knowledge graphs and the multilingual vectors in a dynamic way. Indeed, $c(G)$ makes the contribution of graph similarity depend on the richness of the knowledge graph.

6 Evaluation

In this section we compare our knowledge-based document similarity measure, KBSim, against state-of-the-art models on two different tasks: comparable document retrieval and cross-lingual text categorization.

6.1 Comparable Document Retrieval

In our first experiment we determine the effectiveness of our knowledge-based approach in a comparable document retrieval task. Given a document d written in language L and a collection $D_{L'}$ of documents written in another language L', the task of comparable document retrieval consists of finding the document in $D_{L'}$ which is most similar to d, under the assumption that there exists one document $d' \in D_{L'}$ which is comparable with d.

6.1.1 Corpus and Task Setting

Dataset We followed the experimental setting described in (Platt et al., 2010; Yih et al., 2011) and evaluated KBSim on the Wikipedia dataset made available by the authors of those papers. The dataset is composed of Wikipedia comparable encyclopedic entries in English and Spanish. For each document in English there exists a "real" pair in Spanish which was defined as a comparable entry by the Wikipedia user community. The dataset of each language was split into three parts: 43,380 training, 8,675 development and 8,675 test documents. The documents were tokenized, without stemming, and represented as vectors using a log(tf)-idf weighting (Salton and Buckley, 1988). The vocabulary of the corpus was restricted to 20,000 terms, which were the most frequent terms in the two languages after removing the top 50 terms.

Methodology To evaluate the models we compared each English document against the Spanish dataset and vice versa. Following the original setting, the results are given as the average performance between these two experiments. For evaluation we employed the averaged top-1 accuracy and Mean Reciprocal Rank (MMR) at finding the real comparable document in the other language. We compared KBSim against the state-of-the-art supervised models S2Net, OPCA, CCA, and CL-LSI (cf. Section 2). In contrast to these models, KBSim does not need a training step, so we applied it directly to the testing partition.

In addition we also included the results of CL-ESA[7], CL-C3G[8] and two simple vector-based models which translate all documents into English on a word-by-word basis and compared them using cosine similarity: the first model (CosSim_E) uses a statistical dictionary trained with Europarl using Wavelet-Domain Hidden Markov Models (He, 2007), a model similar to IBM Model 4; the second model (CosSim_{BN}) instead uses Algorithm 1 to translate the vectors with BabelNet.

6.1.2 Results

As we can see from Table 1,[9] the CosSim_{BN} model, which uses BabelNet to translate the document vectors, achieves better results than CCA and CL-LSI. We hypothesize that this is due to these linear projection models losing information during the projection. CosSim_E yields results similar to CosSim_{BN}, showing that BabelNet is a good alternative statistical dictionary. In contrast to CCA

[7] Document collections with sizes higher than 10^5 provide high performance (Potthast et al., 2008). Here we used 15k documents from the training set to index the test documents.

[8] CL-C3G is CL-CNG using character 3-grams, which has proven to be the best length (Mcnamee and Mayfield, 2004).

[9] In this work, statistically significant results according to a χ^2 test are highlighted in bold.

Model	Dimension	Accuracy	MMR
S2Net	2000	**0.7447**	0.7973
KBSim	N/A	**0.7342**	0.7750
OPCA	2000	**0.7255**	0.7734
CosSim$_E$	N/A	0.7033	0.7467
CosSim$_{BN}$	N/A	0.7029	0.7550
CCA	1500	0.6894	0.7378
CL-LSI	5000	0.5302	0.6130
CL-ESA	15000	0.2660	0.3305
CL-C3G	N/A	0.2511	0.3025

Table 1: Test results for comparable document retrieval in Wikipedia. S2Net, OPCA, CosSim$_E$, CCA and CL-LSI are from (Yih et al., 2011).

Model	Dim.	EN News Accuracy	ES News Accuracy
KBSim	N/A	0.8189	**0.6997**
Full MT	50	**0.8483**	0.6484
CosSim$_{BN}$	N/A	0.8023	**0.6737**
OPCA	100	**0.8412**	0.5954
CCA	150	**0.8388**	0.5323
CL-LSI	5000	**0.8401**	0.5105
CosSim$_E$	N/A	0.8046	0.4481

Table 2: Test results for cross-language text categorization. Full MT, OPCA, CCA, CL-LSI and CosSim$_E$ are from (Platt et al., 2010).

and CL-LSI, OPCA performs better thanks to its improved projection method using a noise covariance matrix, which enables it to obtain the main components in a low-dimensional space.

CL-C3G and CL-ESA obtain the lowest results. Considering that English and Spanish do not have many lexical similarities, the low performance of CL-C3G is justified because these languages do not share many character n-grams. The reason behind the low results of CL-ESA can be explained by the low number of intersecting concepts between Spanish and English in Wikipedia, as confirmed by Potthast et al. (2008). Despite both using Wikipedia in some way, KBSim obtains much higher performance than CL-ESA thanks to the use of our multilingual knowledge graph representation of documents, which makes it possible to expand and semantically relate its original concepts. As a result, in contrast to CL-ESA, KBSim can integrate conceptual and relational similarity functions which provide more accurate performance. Interestingly, KBSim also outperforms OPCA which, in contrast to our system, is supervised, and in terms of accuracy is only 1 point below S2Net, the supervised state-of-the-art model using neural networks.

6.2 Cross-language Text Categorization

The second task in which we tested the different models was cross-language text categorization. The task is defined as follows: given a document d_L in a language L and a corpus $D'_{L'}$ with documents in a different language L', and C possible categories, a system has to classify d_L into one of the categories C using the labeled collection $D'_{L'}$.

6.2.1 Corpus and Task Setting

Dataset To perform this task we used the Multilingual Reuters Collection (Amini et al., 2009), which is composed of five datasets of news from five different languages (English, French, German, Spanish and Italian) and classified into six possible categories. In addition, each dataset of news is translated into the other four languages using the Portage translation system (Sadat et al., 2005). As a result, we have five different multilingual datasets, each containing source news documents in one language and four sets of translated documents in the other languages. Each of the languages has an independent vocabulary. Document vectors in the collection are created using TFIDF-based weighting.

Methodology To evaluate our approach we used the English and Spanish news datasets. From the English news dataset we randomly selected 13,131 news as training and 1,875 as test documents. From the Spanish news dataset we selected all 12,342 news as test documents. To classify both test sets we used the English news training set. We performed the experiment at cross-lingual level using Spanish and English languages available for both Spanish and English news datasets, therefore we classified each test set selecting the documents in English and using the Spanish documents in the training dataset, and vice versa. We followed Platt et al. (2010) and averaged the values obtained from the two comparisons for each test set to obtain the final result. To categorize the documents we applied k-NN to the ranked list of documents according to the similarity measure employed for each model. We evaluated each model by estimating its accuracy in the classification of the English and Spanish test sets.

We compared our approach against the state-of-the-art supervised models in this task: OPCA, CCA and CL-LSI (Platt et al., 2010). In addition, we include the results of the CosSim$_{BN}$ and CosSim$_E$ models that we introduced in Section 6.1.1, as well as the results of a full statistical machine translation system trained with Europarl and post-processed by LSA (Full MT), as reported by Platt et al. (2010).

6.2.2 Results

Table 2 shows the cross-language text categorization accuracy. CosSim_E obtained the lowest results. This is because there is a significant number of untranslated terms in the translation process that the statistical dictionary cannot cover. This is not the case in the CosSim_{BN} model which achieves higher results using BabelNet as a statistical dictionary, especially on the Spanish news corpus.

On the other hand, however, the linear projection methods as well as Full MT obtained the highest results on the English corpus. The differences between the linear projection methods are evident when looking at the Spanish corpus results; OPCA performed best with a considerable improvement, which indicates again that it is one of the most effective linear projection methods. Finally, our approach, KBSim, obtained competitive results on the English corpus, performing best among the unsupervised systems, and the highest results on the Spanish news, surpassing all alternatives.

Since KBSim does not need any training for document comparison, and because it based, moreover, on a multilingual lexical resource, we performed an additional experiment to demonstrate its ability to carry out the same text categorization task in many languages. To do this, we used the Multilingual Reuters Collection to create a 3,000 document test dataset and 9,000 training dataset[10] for five languages: English, German, Spanish, French and Italian. Then we calculated the classification accuracy on each test set using each training set. Results are shown in Table 3.

The best results for each language were obtained when working at the monolingual level, which suggests that KBSim might be a good untrained alternative in monolingual tasks, too. In general, cross-language comparisons produced similar results, demonstrating the general applicability of KBSim to arbitrary language pairs in multilingual text categorization. However, we note that German, Italian and Spanish training partitions produced low results compared to the others. After analyzing the length of the documents in the different datasets we discovered that they have different average lengths in words: 79 (EN), 76 (FR), 75 (DE), 60 (ES) and 55 (IT). German, Spanish and especially Italian documents have the lowest average length, which makes it more difficult to build a representative knowledge graph of the content of each document when it is performing at cross-language level.

[10]Note that training is needed for the k-NN classifier, but not for document comparison.

Testing datasets	Training datasets				
	DE	EN	ES	FR	IT
DE	0.8053	0.6872	0.5373	0.6417	0.5920
EN	0.5827	0.8463	0.5540	0.6530	0.5820
ES	0.5883	0.6153	0.8707	0.6237	0.7010
FR	0.6867	0.7103	0.6667	0.8227	0.6887
IT	0.5973	0.5487	0.6263	0.5973	0.8317

Table 3: KBSim accuracy in a multilingual setup.

7 Conclusions

In this paper we introduced a knowledge-based approach to represent and compare documents written in different languages. The two main contributions of this work are: i) a new graph-based model for the language-independent representation of documents based on the BabelNet multilingual semantic network; ii) KBSim, a knowledge-based cross-language similarity measure between documents, which integrates our multilingual graph-based model with a traditional vector representation.

In two different cross-lingual tasks, i.e., comparable document retrieval and cross-language text categorization, KBSim has proven to perform on a par or better than the supervised state-of-the-art models which make use of linear projections to obtain the main components of the term vectors. We remark that, in contrast to the best systems in the literature, KBSim does not need any parameter tuning phase nor does it use any training information. Moreover, when scaling to many languages, supervised systems need to be trained on each pair, which can be very costly.

The gist of our approach is in the knowledge graph representation of documents, which relates the original terms using expanded concepts and relations from BabelNet. The knowledge graphs also have the nice feature of being human-interpretable, a feature that we want to exploit in future work. We will also explore the integration of linear projection models, such as OPCA and S2Net, into our multilingual vector-based similarity measure. Also, to ensure a level playing field, following the competing models, in this work we did not use multi-word expressions as vector components. We will study their impact on KBSim in future work.

Acknowledgments

The authors gratefully acknowledge the support of the ERC Starting Grant MultiJEDI No. 259234, EC WIQ-EI IRSES (Grant No. 269180) and MICINN DIANA-Applications (TIN2012-38603-C02-01). Thanks go to Yih et al. for their support and Jim McManus for his comments.

References

Massih-Reza Amini, Nicolas Usunier, and Cyril Goutte. 2009. Learning from multiple partially observed views - an application to multilingual text categorization. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*, pages 28–36.

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia - a crystallization point for the web of data. *J. Web Sem.*, 7(3):154–165.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Philipp Cimiano, Antje Schultz, Sergej Sizov, Philipp Sorg, and Steffen Staab. 2009. Explicit versus latent concept models for cross-language information retrieval. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, volume 9, pages 1513–1518.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Konstantinos I. Diamantaras and Sun Y. Kung. 1996. *Principal component neural networks*. Wiley New York.

Susan T. Dumais, Todd A. Letsche, Michael L. Littman, and Thomas K. Landauer. 1997. Automatic cross-language retrieval using latent semantic indexing. In *Proc. of AAAI Spring Symposium on Cross-language Text and Speech Retrieval*, pages 18–24.

Christiane Fellbaum. 1998. *WordNet: An electronic lexical database*. Bradford Books.

Marc Franco-Salvador, Parth Gupta, and Paolo Rosso. 2013. Cross-language plagiarism detection using a multilingual semantic network. In *Proc. of the 35th European Conference on Information Retrieval (ECIR'13)*, volume LNCS(7814), pages 710–713. Springer-Verlag.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proc. of the 20th International Joint Conference on Artifical Intelligence (IJCAI)*, pages 1606–1611.

Taher Haveliwala, Sepandar Kamvar, and Glen Jeh. 2003. An analytical comparison of approaches to personalizing pagerank. Technical Report 2003-35, Stanford InfoLab, June.

Xiaodong He. 2007. Using word dependent transition models in hmm based word alignment for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 80–87. Association for Computational Linguistics.

Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. 2013. Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61.

Eduard H. Hovy, Roberto Navigli, and Simone Paolo Ponzetto. 2013. Collaboratively built semi-structured content and Artificial Intelligence: The story so far. *Artificial Intelligence*, 194:2–27.

David Hull. 1994. Improving text retrieval for the routing problem using latent semantic indexing. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 282–291. Springer.

Donald A. Jackson, Keith M. Somers, and Harold H. Harvey. 1989. Similarity coefficients: measures of co-occurrence and association or simply measures of occurrence? *American Naturalist*, pages 436–453.

Ian T. Jolliffe. 1986. *Principal component analysis*, volume 487. Springer-Verlag New York.

Thomas K. Landauer and Michael L. Littman. 1994. Computerized cross-language document retrieval using latent semantic indexing, April 5. US Patent 5,301,109.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

Michael Matuschek and Iryna Gurevych. 2013. Dijkstra-WSA: A graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics (TACL)*, 1:151–164.

Paul Mcnamee and James Mayfield. 2004. Character n-gram tokenization for european language text retrieval. *Information Retrieval*, 7(1-2):73–97.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 880–889. Association for Computational Linguistics.

Manuel Montes y Gómez, Alexander F. Gelbukh, Aurelio López-López, and Ricardo A. Baeza-Yates. 2001. Flexible comparison of conceptual graphs. In *Proc. of the 12th International Conference on Database and Expert Systems Applications (DEXA)*, pages 102–111.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Vivi Nastase and Michael Strube. 2013. Transforming wikipedia into a large scale multilingual concept network. *Artificial Intelligence*, 194:62–85.

Roberto Navigli and Mirella Lapata. 2010. An experimental study of graph connectivity for unsupervised word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):678–692.

Roberto Navigli and Simone Paolo Ponzetto. 2012a. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli and Simone Paolo Ponzetto. 2012b. BabelRelate! a joint multilingual approach to computing semantic relatedness. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*, pages 108–114, Toronto, Canada.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1998. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project.

John C. Platt, Kristina Toutanova, and Wen-tau Yih. 2010. Translingual document representations from discriminative projections. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 251–261.

Martin Potthast, Benno Stein, and Maik Anderka. 2008. A wikipedia-based multilingual retrieval model. In *Advances in Information Retrieval*, pages 522–530. Springer.

Martin Potthast, Alberto Barrón-Cedeño, Benno Stein, and Paolo Rosso. 2011. Cross-language plagiarism detection. *Language Resources and Evaluation*, 45(1):45–62.

Fatiha Sadat, Howard Johnson, Akakpo Agbago, George Foster, Joel Martin, and Aaron Tikuisis. 2005. Portage: A phrase-based machine translation system. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, USA.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

Gerard Salton and Michael J. McGill. 1986. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA.

Gerard Salton, Edward A. Fox, and Harry Wu. 1983. Extended boolean information retrieval. *Communications of the ACM*, 26(11):1022–1036.

Bruce Thompson. 2005. Canonical correlation analysis. *Encyclopedia of statistics in behavioral science*.

Piek Vossen. 2004. EuroWordNet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual index. *International Journal of Lexicography*, 17(2):161–173.

Wen-tau Yih, Kristina Toutanova, John C. Platt, and Christopher Meek. 2011. Learning discriminative projections for text similarity measures. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 247–256.