



Wikipedia-based cross-language text classification



Marcos Antonio Mouriño García*, Roberto Pérez Rodríguez, Luis Anido Rifón

Department of Telematics Engineering, Telecommunication Engineering School, University of Vigo, Campus Lagoas-Marcosende, Vigo, 36310, Spain

ARTICLE INFO

Article history:

Received 14 March 2016

Revised 27 March 2017

Accepted 10 April 2017

Available online 13 April 2017

Keywords:

Cross-language text classification

Wikipedia Miner

Bag of concepts

Bag of words

Hybrid

Document representation

ABSTRACT

This paper presents the application of a Wikipedia-based bag of concepts (WikiBoC) document representation to cross-language text classification (CLTC). Its main objective is to alleviate the major drawbacks of the state-of-the-art CLTC approaches – typically based on the machine translation (MT) of documents, which are represented as bags of words (BoW). We propose a technique called cross-language concept matching (CLCM), to convert concept-based representations of documents from one language to another using Wikipedia correspondences between concepts in different languages and thus not relying on automated full-text translations. We describe two proposals: the first proposal consists in the use of the WikiBoC representation in conjunction with the CLCM technique (WikiBoC-CLCM) to classify documents written in a language L_1 by using a SVM algorithm that was trained with documents written in another language L_2 ; the second proposal consists of a hybrid model for representing documents that combines WikiBoC-CLCM with the classic BoW-MT approach. To evaluate the two proposals we conducted several experiments with three cross-lingual corpora: the JRC-Acquis corpus and two purpose-built corpora composed of Wikipedia articles. The first proposal outperforms state-of-the-art approaches when training sequences are short, achieving performance increases up to 233.33%. The second proposal outperforms state-of-the-art approaches in the whole range of training sequences, achieving performance increases up to 23.78%. Results obtained show the benefits of the WikiBoC-CLCM approach, since concepts extracted from documents add useful information to the classifier, thus improving its performance.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

Text classification consists in the algorithmic assignation of text documents to predefined classes. It has multiple applications, such as sentiment analysis and spam filtering. Document classification can be modelled as a machine learning problem: a classification algorithm is trained with a labelled set of examples, and it is later applied to a set of unlabelled documents [36]. Usually, the larger the training set, the higher the performance of the classifier is [23]. Therefore, algorithmic classification may perform poorly when we do not count with a large enough set of documents to train the classifier [16]. It is in this scenario when cross-language text classification (CLTC) becomes relevant: it consists in training a classifier with a labelled set of documents written in a language L_1 – where large training sequences are available – to classify a set of unlabelled documents written in a different language L_2 .

* Corresponding author.

E-mail addresses: marcos@gist.uvigo.es (M.A. Mouriño García), roberto.perez@gist.uvigo.es (R. Pérez Rodríguez), lanido@gist.uvigo.es (L. Anido Rifón).

Text documents have to be represented in a way that classifiers can understand and relate them. These representations are based on the extraction of features from natural language, such as the frequency of occurrence of words or the structure of the language used. The most used representation is the bag of words (BoW) model, where a document is represented by a set of words and the frequency of occurrence of these words in the document.

Cross-language classification of text documents has traditionally been approached by using the bag of words representation along with machine translation (MT) techniques, either translating the documents before extracting the set of features [23,33,45], or translating the features themselves [28,37,47]. Both approaches have a number of drawbacks related to both the bag of words representation and machine translation techniques. On the one hand, despite being one of the traditionally used representations in document classification tasks, the BoW representation is suboptimal because it only accounts for word frequency in text, which involves the emergence of two problems of language that affect the classification performance: redundancy (synonymy problem) and ambiguity (polysemy problem) [19,27,46]. On the other hand, machine translation techniques have two major drawbacks: lexical and structural ambiguity [20,37], which negatively affect the quality of translations. Thus, if an incorrect translation is selected, it can distort the precision of the classifier due to the introduction of erroneous features into the classifier. Therefore, when the bag of words representation is combined with machine translation techniques the disadvantages of each one add up, which leads to an increased error probability.

With the aim of solving the problems inherent to bag of words representations, several authors explored a new paradigm: the bag of concepts (BoC) representation of documents, being a concept a “unit of meaning” [5,46]. Concepts are non-ambiguous by definition, which alleviates the problems introduced by synonymy and polysemy. In accordance with the bag of concepts representation, a document is represented by a set of concepts and their associated weights, which indicate their relevance in the text. Several previous studies demonstrate that this representation provides good results in classification tasks [35,46], clustering [18] and information retrieval [9]. The literature hosts different ways to create BoC representations, such as Latent Semantic Analysis (LSA) [7], Latent Dirichlet Allocation (LDA) [3], Explicit Semantic Analysis (ESA) [12], word embeddings (WE) [2], and semantic annotators [26].

We consider that there exists a research gap in the application of bag of concepts representations (leveraging encyclopedic knowledge without requiring any kind of extra linguistic resources such as thesauri or dictionaries) to building cross language classifiers of text documents. This article aims at bridging this gap by describing the foundations and reporting the evaluation results of a cross-language classifier that leverages Wikipedia knowledge to represent text documents as bags of concepts, and that performs classification without relying on machine translation. To that end, we propose a technique called *cross-language concept matching* (CLCM), that converts the bag of concepts representation of a document in a language L_1 to a language L_2 , by leveraging Wikipedia interlanguage links. This feature has also been leveraged by other authors to perform cross-language text classification [40]. We call the proposed classifier WikiBoC-CLCM. Furthermore, we also propose a hybrid model that combines the BoW-MT and WikiBoC-CLCM approaches, by enriching the bag of words representation of each document with concepts extracted from the document itself.

In order to evaluate the system we conducted several classification experiments with our two proposals, and compared the performance with four state-of-the-art approaches: the BoW-MT, the ESA-concept-based approach, the Bilingual LDA (BiLDA), and bilingual word embeddings (BWEs). In order to perform a comprehensive evaluation of the proposed classifiers we used three datasets covering different domains. We expressly created the first two corpora – Wikipedia Corpus and Wikipedia Human Medicine corpus – composed of Wikipedia documents about general and biomedical topics respectively. Besides, we used the JRC-Acquis corpus [38] that comprises European Union documents of legal nature.

The remainder of this article is organised as follows. Section 2 reviews the most relevant state-of-the-art approaches to performing CLTC. Section 3 presents the *Wikipedia Miner* algorithm – the semantic annotator selected to create the concept representation of documents used in our approach – the representations of documents employed, the cross-language concept matching (CLCM) technique, and the description and the generation process of the corpora. Section 4 exposes the two approaches proposed: the WikiBoC-CLCM and the hybrid model. Section 5 describes the experiments conducted and shows the results obtained. Section 6 discusses and analyses the results gathered. Section 7 shows the limitations of the research. Finally, Section 8 presents some conclusions.

2. Literature review

Cross-language text classification is a relatively recent research area, and the available literature is scarce on this subject [33]. In this section, we briefly review published studies, grouped in accordance with the method followed to represent documents: bag of words or bag of concepts.

2.1. Classifiers based on bag of words representations

This kind of classifiers use weighted word vectors to represent documents, basing the calculation of weights on the occurrences of words in text. Their main approaches to cross-language text classification are Cross-Lingual Training and Multi-View Learning.

Cross-Lingual Training classifiers are trained on a corpus of translated documents [23,44] – leveraging machine translation [8,20] tools such as Google Translate¹ and Bing Translator² (see Fig. 1(a)). A variation of this approach consists in translating the features extracted from documents instead [28,37,47], either during the training or the classification phase (see Fig. 1(b)).

Multi-View Learning classifiers are trained with both the original and translated documents, and they regard each language as one independent view of the data (see Fig. 1(c)). Amini et al. [1] and Guo and Xiao [15] train classifiers with multilingual collections in which not all documents are available in all languages, concluding that additional views obtained through machine translation may significantly increase classification performance when there is only a low number of available labelled documents. Wan [45] proposes a co-training approach to sentiment classification of Chinese product reviews by using labelled English and unlabelled Chinese reviews as training data. Lu et al. [24] propose a model for joint bilingual sentiment classification that augments labelled data in each language with unlabelled parallel data – obtained through machine translation – reporting improved performance for English and Chinese. Finally, Fortuna and Shawe-Taylor [11] use machine translation and Canonical Correlation Analysis (CCA) to leverage an existing corpus of annotated documents from one language in order to classify documents in a different language.

2.2. Classifiers based on bag of concepts representations

This type of classifiers represent documents as weighted concept vectors using, to this end, techniques such as Latent Semantic Analysis, Latent Dirichlet Allocation, Explicit Semantic Analysis, and word embeddings.

The Latent Semantic Analysis (LSA) technique [7] is based on the distributional hypothesis [34], which states that words that occur in similar contexts tend to have similar meanings. This is line with the theory of language of Wittgenstein [48]: “meaning is use”. In the LSA model, a concept is a vector that represents the context in which a term occurs. This representation overcomes synonymy, but it does not combat polysemy. Gliozzo and Strapparava [14] report promising results using this technique in combination with a SVM algorithm to classify Italian documents when only English ones were available for training, and vice versa.

Latent Dirichlet Allocation (LDA) [3] presupposes that each document within a collection comprises a small number of topics, each one of them “generating” words. Thus, LDA automatically finds topics in text by “going back” from the document and finds the set of topics that may have generated it. There are some studies – such as the ones by De Smet et al. [6] and Ni et al. [30] – that (1) use LDA to extract multilingual topic models from parallel and comparable corpora, (2) represent the documents in the space of multilingual topics extracted, and (3) perform cross-lingual classification across languages in which there are no labelled documents available.

Gabrilovich and Markovitch [12] propose Explicit Semantic Analysis (ESA), a technique that leverages external knowledge sources such as Wikipedia or the Open Directory Project to generate features from text documents. ESA performs an explicit semantic analysis of the input text, identifying topics that are explicitly present in background knowledge bases, instead of latent topics. The main disadvantage of this approach is its tendency toward generating outliers [9] – concepts that are related to the annotated document only marginally – which hampers its usefulness for classification. Maleshkova et al. [25] classify a set of APIs written in Czech making use of a set of APIs written in English by representing the documents as ESA vectors and converting these concept vectors between the two languages.

Word embeddings are dense real-valued vectors, also known as distributed representations of words [2,43]. They have recently been proposed, serving as rich and coherent word representations. The extension from monolingual to multilingual settings [22,40] allows to learn embeddings for words denoting similar concepts that are very close in the shared bilingual embedding space – e.g., the representations for the English word *car* and the Spanish word *coche* should be very similar. Then, these bilingual word embeddings (BWEs) can be used in several application scenarios such as cross-lingual semantic word similarity, cross-lingual information retrieval and cross-language text classification [41,43].

3. Materials and methods

This section briefly presents: the semantic annotator that we leverage for extracting concepts from text, *Wikipedia Miner*; the representation of documents; the cross-language concept matching technique; and the corpora used in the evaluation of the different approaches.

3.1. Semantic annotator: Wikipedia Miner algorithm

We rely on a semantic annotator for obtaining weighted concept vectors representing text documents: *Wikipedia Miner* [26], which is a general-purpose semantic annotator that builds on natural language processing and machine learning techniques, and uses Wikipedia as its knowledge base. *Wikipedia Miner* (1) identifies only concepts that are actually present in the documents, thus avoiding the generation of irrelevant annotations; (2) performs word-sense disambiguation, tackling

¹ <https://translate.google.com> (Accessed 25 January 2017).

² <https://www.bing.com/translator> (Accessed 25 January 2017).

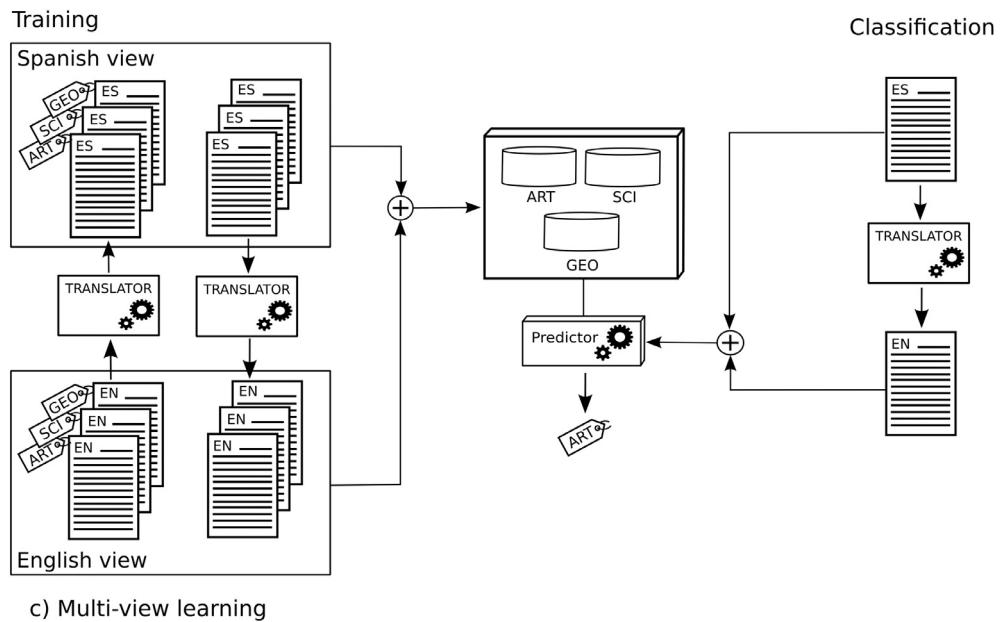
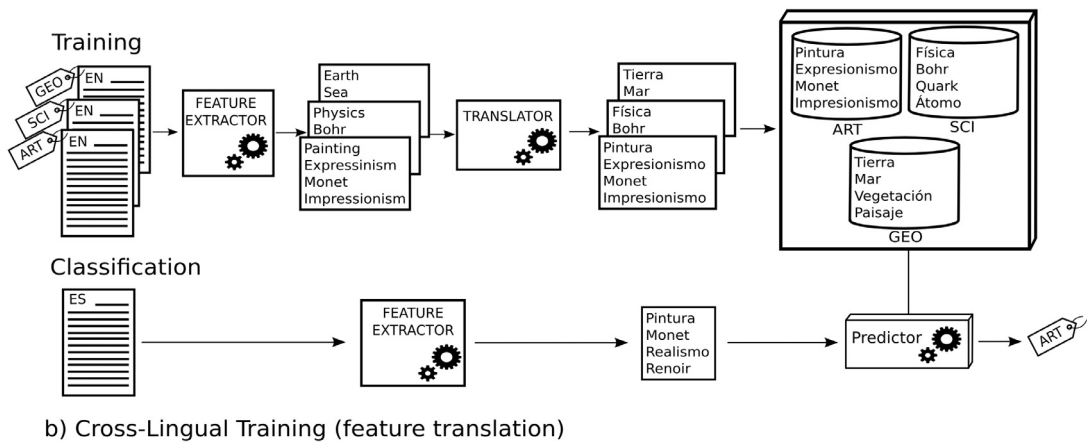
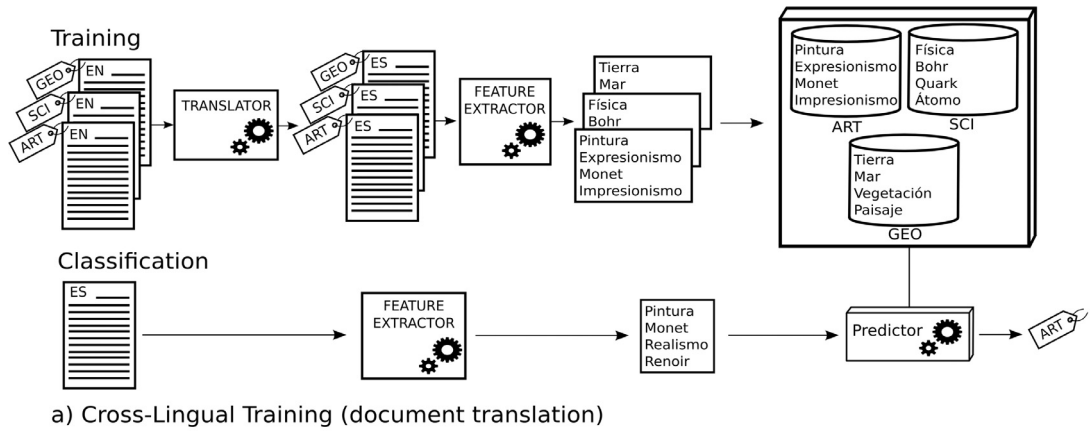


Fig. 1. State-of-the-art approaches to perform CLTC with the BoW paradigm.

the synonymy and polysemy problems; (3) links concepts to Wikipedia entries; and (4) calculates weights for extracted concepts in accordance with their relevance in the text.

3.2. Document representations

Document representations are based on the extraction of features from natural language text, being the most widely used the vector space model, in which each document in a collection is represented as a point in space and each dimension is a feature. There exist different representations depending on which kind of features we extract from natural language. In particular, we used the following three representations in our study. The first one is the bag of words (BoW) model [35], in which a document is represented as a vector $\overline{d}^{L_j} = (ww_1, ww_2, \dots, ww_{|\mathbb{W}^{L_j}|})$, where ww_i are the weights – frequency of occurrence – of words in the document. The domain of features \mathbb{W}^{L_j} is composed of the set of all words in a particular language L_j , ignoring “stop words” after applying the stemming algorithm of Porter [32]. The second one is the bag of concepts (BoC) model [35], in which a document is represented as a vector $\overline{d}^{L_j} = (cw_1, cw_2, \dots, cw_{|\mathbb{C}^{L_j}|})$, where cw_i are the weights – relevance – of concepts in the document. The domain of features \mathbb{C}^{L_j} is composed of all Wikipedia articles in language L_j . In order to extract the features – concepts – we make use of the *Wikipedia Miner* algorithm previously described. Finally, we used also a combination of the previous two: the hybrid model [18,35].

3.3. Cross-language concept matching technique

The cross-language concept matching (CLCM) technique converts the BoC representation of a document in a language L_m to a different language L_n . The base of this technique lies in *interlanguage links*, a feature of Wikipedia that provides a correspondence between all versions of the same article in the different languages in which they are available. For instance, starting from the Spanish Wikipedia article *Mercurio_(elemento)*, we can easily get the corresponding article in English (*Mercury_(element)*), French (*Mercure_(chimie)*), or German (*Quecksilber*). Then, to convert a BoC representation from a language L_m to a language L_n it is only necessary to obtain the equivalent concept – article – in the language L_n for each concept of the BoC in the language L_m . Fig. 2 shows the conversion process of the BoC representation of a document written in Spanish to its equivalent in English. The table represents a linear transformation function $\tilde{\mathcal{M}}$, which maps the features of a language L_m – in this case Spanish – to another different language L_n – in this case English. Thus, for each pair of features of both languages, if there is a correspondence between them, it will be marked as true (T), and, if there is no correspondence, it will be marked as false (F). The transformation matrix \mathbf{M} is obtained from the aforementioned table, being “1” the value of each component if there is a correspondence, and being “0” if there is not a correspondence. The Spanish bag of concepts is represented as a vector – \overline{d}^{L_m} – as defined in Section 3.2. Finally, the vector \overline{d}^{L_m} is multiplied by the \mathbf{M} matrix in order to obtain the vector \overline{d}^{L_n} that represents the BoC converted from Spanish to English.

The main advantage of this technique is that there is no semantic weight loss between languages, due to the correspondence between different language versions of the same article. Thus, the performance of the classifier will not be affected by context, synonymy, or polysemy, as it is the case of translation-based approaches.

It is worth noting that the space of features – concepts – by which a document can be represented is limited to the set of Wikipedia articles available for each language. More explicitly, a text document written in English is represented by a subset of the articles – i.e., concepts – available in the English Wikipedia, and a text document written in Spanish is represented by a subset of the articles present in the Spanish Wikipedia. Consequently, a Spanish-converted-to-English document is represented in the space of articles resulting from the intersection of Wikipedia English and Wikipedia Spanish articles: the set of articles that are mapped between Spanish and English editions of Wikipedia.


3.4. Corpora

Shi et al. [37] state that there is no standard evaluation benchmark available for cross-lingual text classification. This is consistent with the review of the state-of-the-art conducted for our study, where we observed that different researchers use different corpora, which are expressly created in the majority of cases. Therefore, in order to perform a comprehensive evaluation of the system proposed, we used three datasets covering different domains.³ We expressly created the first two corpora, Wikipedia Corpus and Wikipedia Human Medicine corpus, so we can consider them as additional contributions of this work. The first corpus comprises Wikipedia documents about general topics, while the second one is composed of Wikipedia documents in the biomedical domain. We selected Wikipedia articles to create the corpora because they are written in the language of the Wikipedia edition where they have been extracted, that is to say, they are not translations. Finally, we also used the JRC-Acquis corpus [38], composed of documents of legal nature.

³ The corpora are available for free at http://www.itec-sde.net/cross_language_corpora.zip (Accessed 25 January 2017).

	Novel	Literature	Drama	Prose	Mark_Adlard	Taste_(sociology)	John_Almon
Novela	T	F	F	F	F	F	F
Literatura	F	T	F	F	F	F	F
Poema	F	F	F	F	F	F	F
Prosa	F	F	F	T	F	F	F
Obra_literaria	F	F	F	F	F	F	F
Gusto_artístico	F	F	F	F	F	T	F

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

	concept	relevance
	Novela	0.43
	Poema	0.75
	Prosa	0.35
	Gusto_artístico	0.65

$\rightarrow d^{L^m} = (0.43, 0, 0.75, 0.35, 0, 0.65)$

$$\text{conversion}(d^{L^m}) = d^{L^n} = d^{L^m} \times \mathbf{M}$$

$$d^{L^n} = (0.43, 0, 0.75, 0.35, 0, 0.65) \times \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

$$d^{L^n} = (0.43, 0, 0, 0.35, 0, 0.65, 0)$$


	concept	relevance	
$d^{L^n} = (0.43, 0, 0, 0.35, 0, 0.65, 0) \rightarrow$	Novel	0.43	
	Prose	0.35	
	Taste_(sociology)	0.65	

Fig. 2. Conversion process of the BoC representation of a document written in Spanish to its equivalent in English.

3.4.1. Wikipedia Corpus

Wikipedia *Portal:Contents*⁴ provides a navigation system to help browsing content in the encyclopedia. Among the categories in which it is organised, we selected three semantically distant categories to create the corpus: Culture and the arts, Geography and places and Mathematics and logic.

In order to create the training sequence we automatically selected approximately 1000 articles from the English edition of Wikipedia for each of the three categories and parsed the HTML code in order to extract the relevant information (the title and the full body) of each article. In order to create the test set of the corpus we followed the Wikipedia interlanguage links to the Spanish equivalent (if available) for each article in the training set and parsed the HTML code to extract the relevant information. As a result, we obtained a corpus formed by a training sequence that comprises 3019 Wikipedia articles written in English, and a test sequence composed of 832 articles written in Spanish.

3.4.2. Wikipedia Human Medicine corpus

The creation process of Wikipedia Human Medicine corpus is the same as for creating Wikipedia Corpus. We first selected the category “Human medicine”, placed under “Health and fitness” *Portal:Content* category. It has 22 subcategories: Alternative medicine, Cardiology, Endocrinology, Forensics, Gastroenterology, Human Genetics, Geriatrics, Gerontology, Gyne-

⁴ <https://en.wikipedia.org/wiki/Portal:Contents/Categories> (Accessed 25 January 2017).

cology, Hematology, Nephrology, Neurology, Obstetrics, Oncology, Ophthalmology, Orthopedic surgical procedures, Pathology, Pediatrics, Psychiatry, Rheumatology, Surgery and Urology.

To create the training set of the corpus we selected from the English edition of Wikipedia all the articles classified under each of the aforementioned categories and parsed their HTML code in order to extract the title and the full body of each of them. To create the test set we followed the interlanguage links to the Spanish equivalent (if available) for each article in the training set and parsed the HTML code to extract the relevant information they contain. As a result, we obtained a corpus formed by a training sequence composed of 2143 Wikipedia articles written in English, and a test sequence that comprises 469 articles written in Spanish.

3.4.3. JRC-Acquis corpus

The JRC-Acquis [38] is a multi-label and multilingual corpus that comprises European Union (EU) documents of mostly legal nature. The documents are available in 22 official EU languages, so this corpus is particularly suitable to carry out all types of cross-language research. Most texts have been manually classified into subject domains according to the EuroVoc thesaurus [10]. The highest level in the EuroVoc hierarchy contains 21 subjects: Politics, International Relations, European Union, Law, Economics, Trade, Finance, Social Questions, Education and Communications, Science, Business and Competition, Employment and Working Conditions, Transport, Environment, Agriculture, Forestry and Fisheries, Agri-Foodstuffs, Production, Technology and Research, Energy, Industry, Geography and International Organisations.

In order to build the training sequence of the corpus we first downloaded the English and Spanish documents from year 1995,⁵ to later parse the XML code of each document in order to extract the title and the full body. As a result, we obtained a corpus formed by a training sequence that comprises 20411 documents written in English, and a test sequence composed of 20507 documents written in Spanish.

4. Approach

This section presents the two approaches we propose to perform cross-lingual text classification: WikiBoC-CLCM and the hybrid model. First of all, we briefly describe the main structural components involved in the approaches proposed.

4.1. Structural components

Wikipedia Miner concept extractor. This is the component responsible for extracting the concepts from documents and creating the BoC representation of each of them. In our approach we make use of *Wikipedia Miner* concept extractor, that is explained in more detail in Section 3.1.

Cross-language concept matcher. The element responsible for converting the BoC representations of documents from a language L_m to a different language L_n in accordance with Section 3.3.

Bag of Words extractor. The item in charge of obtaining the BoW representations of documents, after removing “stop words” and applying the Porter stemming algorithm.

MT Tool. The element responsible for performing automatic translations of documents between languages. It is based on Google Translate.

Classifier. It is the element responsible for predicting the most suitable category which a document belongs to. We selected the Support Vector Machines (SVM) algorithm [17] because it is one of the most relevant and best performing machine learning algorithms in automatic text classification tasks [33]. To implement the SVM algorithm we used the *Scikit-learn* library for Python [31], in particular, the *sklearn.svm.LinearSVC* class.

4.2. WikiBoC-CLCM cross-language text classification (Proposal 1)

The WikiBoC-CLCM approach is based only on Wikipedia concepts. During the training phase, each labelled document – written in a language L_m – passes through the *Wikipedia Miner* concept extractor in order to obtain its BoC in language L_m , to later train the SVM classification algorithm. In the classification phase, the documents to classify – written in a language L_n – are first passed through the *Wikipedia Miner* concept extractor in order to obtain their BoC representations in language L_n . Since they are BoC representations in language L_n , it is necessary to convert them to language L_m through the cross-language concept matcher. Finally, the converted BoC representations are input into the classifier to predict the category which each document belongs to.

Fig. 3 shows a particular example of the aforementioned, where the classifier is trained with the Wikipedia-based BoC representations of English documents, in order to classify the Wikipedia-based BoC representation of Spanish documents converted to English by using the cross-language concept matcher.

⁵ The Version 3.0 of the JRC-Acquis corpus is freely available on <http://optima.jrc.it/Acquis/JRC-Acquis.3.0/corpus/> (Accessed 25 January 2017).

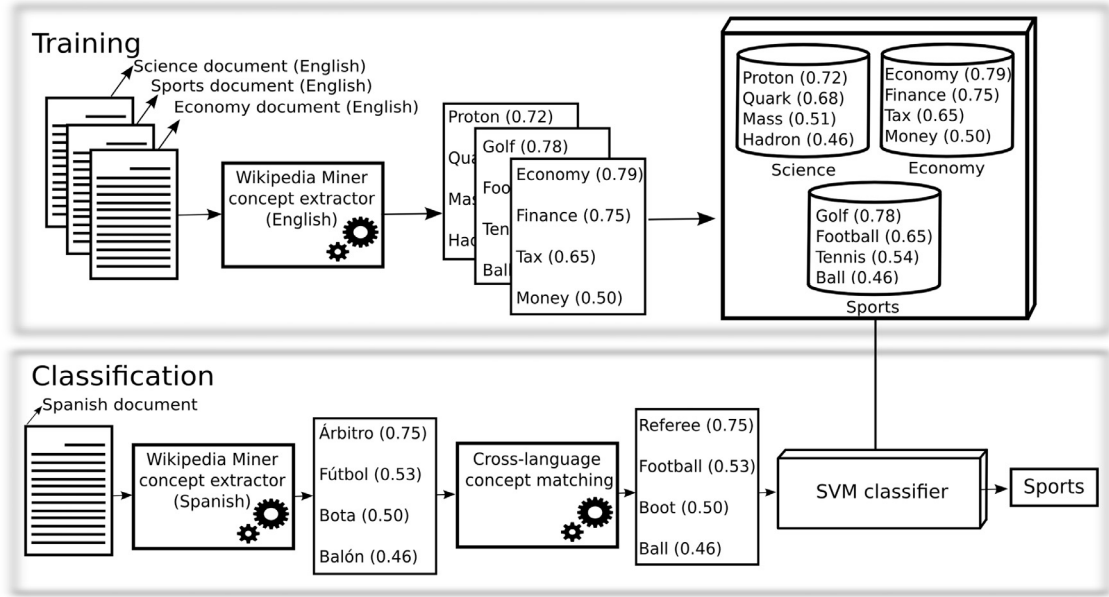


Fig. 3. Architecture of the WikiBoC-CLCM cross-language text classifier.

4.3. Hybrid cross-language text classification (Proposal 2)

In order to leverage the benefits of the traditional BoW representation and Machine Translation, along with the benefits of the Wikipedia-based BoC representation and the CLCM technique, we propose a combination of both approaches.

The implementation of the hybrid approach consists in enriching the bag of words of each document with those concepts extracted by the *Wikipedia Miner* algorithm. Fig. 4 shows the full process graphically. Given that training documents are written in English, all that is required is to obtain the bag of words of each document and enrich it with the concepts extracted by the English version of *Wikipedia Miner*. However, regarding documents to classify, they are written in Spanish and therefore some intermediate steps are needed: (1) translating the documents written in Spanish into English by using Google Translate; (2) obtaining the English bag of words for each document translated; (3) extracting the concepts for each document by using the Spanish version of *Wikipedia Miner*; (4) converting to English each Spanish BoC representation by using the CLCM technique; (5) enriching the BoW of each document with the concepts of the BoC representation converted to English. Finally, these hybrid representations of documents are used to train and test an SVM classification algorithm.

5. Experiments and results

This section presents the experiments conducted to verify the performance of the approaches presented, as well as the results obtained.

5.1. Experimental settings

The experiments we performed consist in – for each one of the corpora – training the SVM classification algorithm with a set of English documents in order to classify a different set of documents written in Spanish, using to this end the two approaches presented: the WikiBoC-CLCM and the hybrid model. Due to the non-existence of a standard evaluation benchmark, it is not possible to compare in a fair way the results obtained to those reported in the literature with the different state-of-the-art approaches. So, we conducted the same experiments with four state-of-the-art approaches: Machine Translation along with the classical BoW document representation (BoW-MT), the ESABoC document representation along with the CLCM approach proposed (ESABoC-CLCM), the Bilingual LDA (Bi-LDA) approach, and bilingual word embeddings (BWEs).

In order to implement the traditional BoW-MT approach, first, the classification algorithm is trained with the BoW representation of the English documents. Then, Spanish documents are automatically translated into English – by using Google Translate – and the BoWs obtained from translated documents are input in the algorithm to predict which category they belong to.

The study by Maleshkova et al. [25], is to the best of our knowledge, the only one that uses an ESABoC representation to perform cross-lingual text classification, but their approach is based on semantic similarity, so their proposal is not comparable with the approaches we propose. Even so, the authors propose a technique conceptually similar to CLCM to convert concept-representations between languages. As the ESA concept-based approach also represents documents as Wikipedia

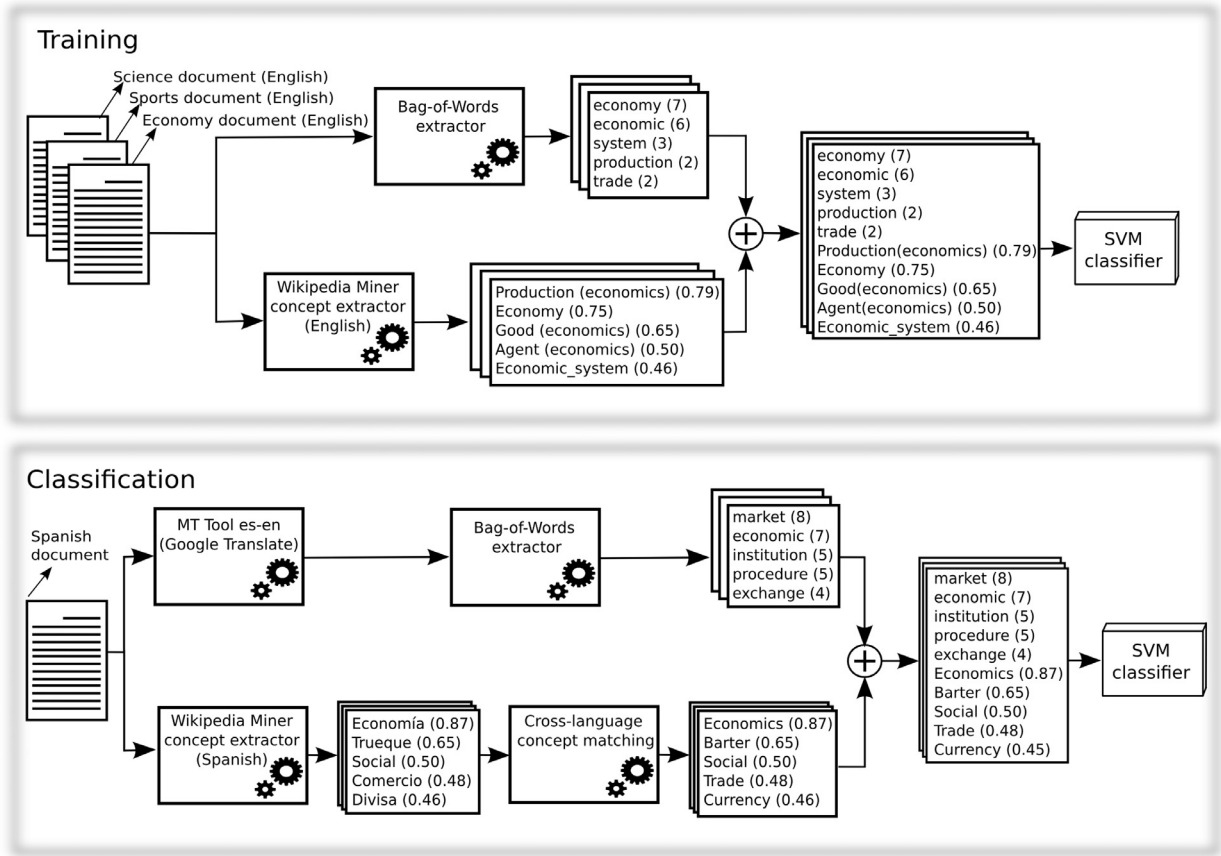


Fig. 4. Architecture of the hybrid cross-language text classifier.

articles, it is suitable for application of the CLCM approach. To implement the ESABoC-CLCM approach, we follow the same methodology used to implement the WikiBoC-CLCM. First, each English training document passes through an English-ESA instance in order to obtain its ESABoC. Next, the SVM classifier is trained with these concept-based representations. Next, during the classification phase, each Spanish document passes through a Spanish-ESA instance in order to obtain its ESABoC. Given that they are BoC representations in Spanish, it is necessary to convert them to English. To that end we use the CLCM technique. Finally, the converted-to-English ESABoC representations are input into the classifier to predict which category documents belong to.

In order to implement the Bi-LDA approach, we follow the methodology proposed and described by De Smet et al. [6], Ni et al. [30] and Vulić et al. [42]. A bilingual probabilistic topic model is first learnt on a general English–Spanish comparable corpora obtained from Wikipedia.⁶ Then, the learnt cross-lingual topics are used to represent the labelled collection of documents written in English and the unlabelled collection of documents written in Spanish. Finally, the SVM algorithm is trained on the set of English labelled documents to later classify the set of Spanish unlabelled documents. The Bi-LDA parameters were selected in accordance with the works by De Smet et al. [6], Ni et al. [30] and Vulić et al. [42]. Then, we selected $K = 400$ bilingual topics and 200 Gibbs Sampling iterations.

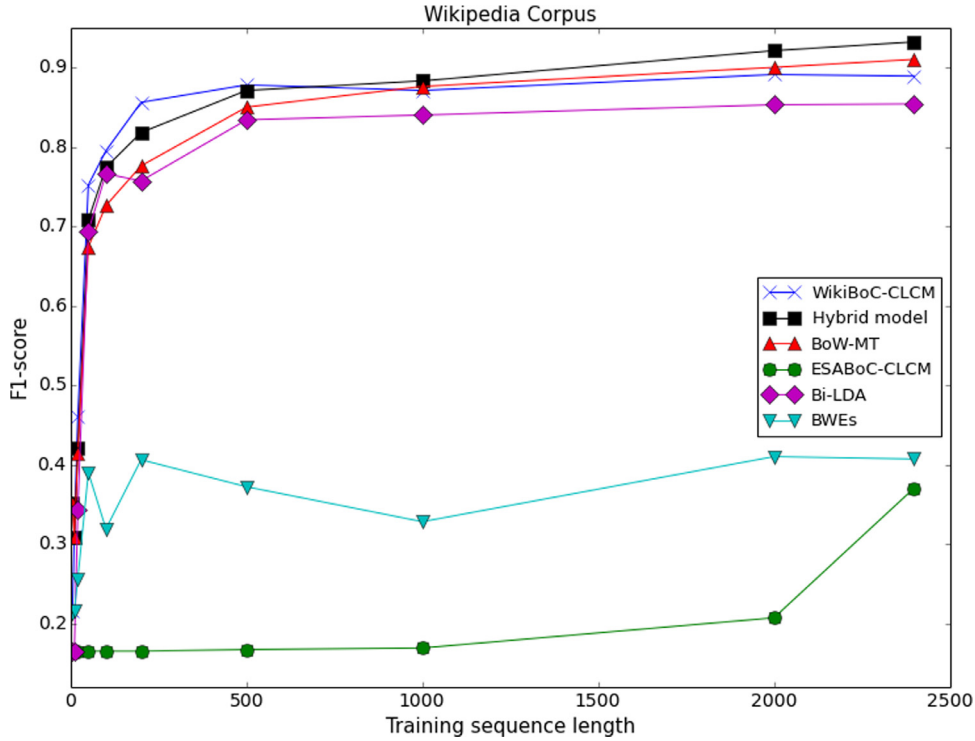
The BWEs proposal is implemented following the methodological directives outlined by Upadhyay et al. [41] and Vulić and Moens [43]. The bilingual word embeddings are learnt from the same English–Spanish comparable corpus used with the Bi-LDA approach. To that end, each pair of comparable documents are merged into a single pseudo-bilingual document using a deterministic strategy based on the length ratio of two documents. Then, a skip-gram model is trained on the corpus of pseudo-bilingual documents in order to create the bilingual word embeddings. Table 1 shows top ten semantically similar words from combined Spanish and English languages for three Spanish words. Hereafter, the labelled set of English documents and the unlabelled collection of Spanish documents are represented under the same bilingual feature space. Finally, the SVM classification algorithm is trained on the labelled collection of English documents to classify the collection of Spanish unlabelled documents.

⁶ The original URL of the corpora (https://people.cs.kuleuven.be/~ivan.vulic/software/Wiki_Spanish_English_Subset.zip) is currently offline. The corpora are available at http://www.itec-sde.net/Wiki_Spanish_English_Subset_Ivan_Vulic_Mirror.zip (Accessed 25 January 2017).

Table 1

Example lists of top ten semantically similar words from combined Spanish and English languages.

Coche	Coches, car, driver, drivers, accidente, cars, bicicleta, llegar, pista, antes
Calor	Temperatura, heat, gas, cooling, temperature, gases,vapor, calienta, liquid, gaseous
Imprimir	Tinta,tintas,printing, impresora, print, impresoras, plancha, cyan,printer, ink

**Fig. 5.** Evolution of the F1-score value when varying the length of the training sequence in the Wikipedia Corpus.**Table 2**

Evolution of the F1-score value when varying the length of the training sequence in the Wikipedia Corpus.

	5	10	20	50	100	200	500	1000	2000	2398
WikiBoC-CLCM	0.216	0.310	0.461	0.752	0.795	0.856	0.878	0.871	0.891	0.889
BoW-MT	0.351	0.308	0.415	0.674	0.726	0.776	0.850	0.876	0.900	0.910
ESABoC-CLCM	0.165	0.165	0.165	0.165	0.165	0.165	0.167	0.169	0.207	0.370
Bi-LDA	0.165	0.165	0.343	0.694	0.766	0.757	0.834	0.840	0.853	0.854
BWEs	0.213	0.216	0.256	0.390	0.319	0.406	0.372	0.328	0.410	0.407
Hybrid model	0.352	0.309	0.420	0.708	0.774	0.818	0.871	0.883	0.921	0.932

5.2. Results

This section shows the results of the experiments. They are presented in terms of F_1 – score, the harmonic mean of classical metrics Precision (P) and Recall (R) [35,36].

5.2.1. Wikipedia Corpus

Fig. 5 and Table 2 show the evolution of the F1-score value when varying the length of the training sequence in the Wikipedia Corpus for the six approaches evaluated in this paper: WikiBoC-CLCM, the hybrid model, ESABoC-CLCM, BoW-MT, Bi-LDA, and BWEs. The WikiBoC-CLCM approach clearly outperforms the traditional BoW-MT approach when the length of the training sequence is not too high, achieving F1-score increases up to 11.57%. It also outperforms the Bi-LDA and BWEs approaches in the complete range of training sequences, achieving increases up to 87.87% and 165.55%, respectively. Furthermore, this approach is the one that offers the highest performance when training sequences are not too long. The hybrid approach outperforms the traditional BoW-MT, Bi-LDA and BWEs approaches in the whole range of training sequence

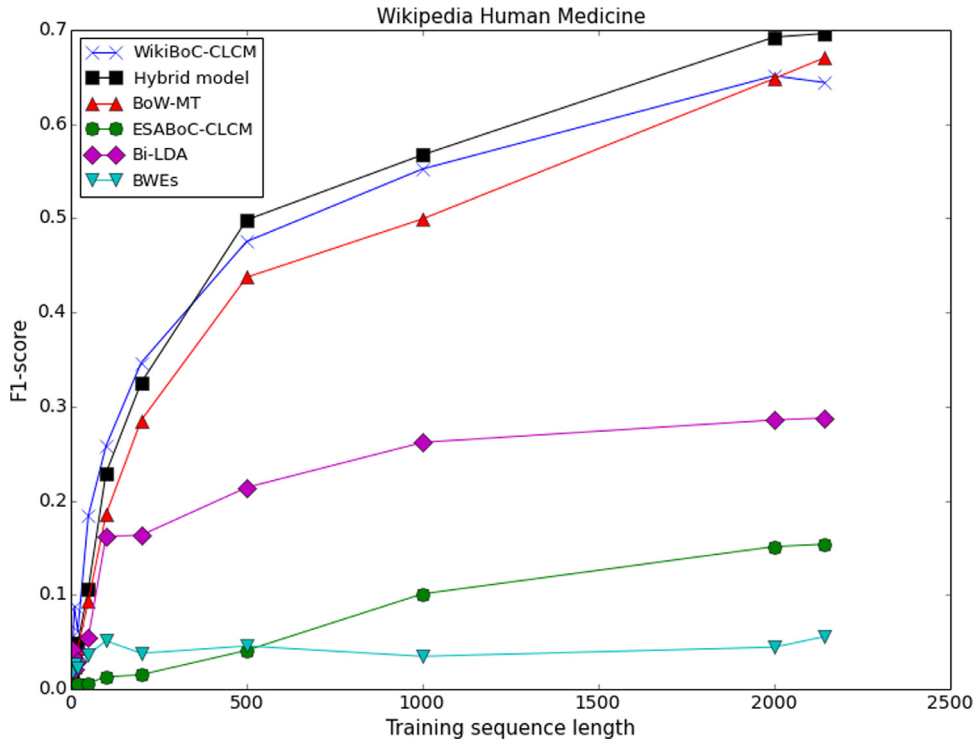


Fig. 6. Evolution of the F1-score value when varying the length of the training sequence in the Wikipedia Human Medicine corpus.

Table 3

Evolution of the F1-score value when varying the length of the training sequence in the Wikipedia Human Medicine corpus.

	5	10	20	50	100	200	500	1000	2000	2143
WikiBoC-CLCM	0.037	0.088	0.061	0.183	0.257	0.346	0.475	0.552	0.651	0.644
BoW-MT	0.016	0.024	0.043	0.093	0.185	0.285	0.437	0.499	0.648	0.670
ESABoC-CLCM	0.006	0.006	0.006	0.106	0.012	0.015	0.041	0.109	0.151	0.154
Bi-LDA	0.037	0.021	0.029	0.055	0.162	0.163	0.214	0.262	0.286	0.288
BWEs	0.018	0.026	0.022	0.036	0.051	0.038	0.046	0.035	0.044	0.055
Hybrid model	0.017	0.030	0.049	0.106	0.229	0.325	0.498	0.567	0.692	0.696

lengths, achieving F1-score increases up to 6.61%, 113.33% and 169.21%, respectively. Finally, the ESABoC-CLCM approach is the one that offers the lowest performance.

5.2.2. Wikipedia Human Medicine corpus

Fig. 6 and Table 3 show the evolution of the F1-score value when varying the length of the training sequence in the Wikipedia Human Medicine corpus for the six approaches evaluated. The WikiBoC-CLCM approach outperforms the traditional BoW-MT approach in almost the whole range of training sequence lengths, achieving F1-score increases up to 266.66%. The WikiBoC-CLCM performance is also higher than the Bi-LDA one in the entire range of training sequence lengths, reaching increases up to 319.04%. Further, this approach is the one that offers the highest performance when training sequences are short. The hybrid approach outperforms the BoW-MT and Bi-LDA approaches in the whole range of training sequence lengths, achieving F1-score increases up to 23.78% and 141.95%, respectively. Finally, the BWEs and ESABoC-CLCM approaches are those that offer the lowest performance.

5.2.3. JRC-Acquis corpus

Fig. 7 and Table 4 show the evolution of the F1-score value when varying the length of the training sequence in the JRC-Acquis corpus. The ESABoC-CLCM approach is the one that offers the lowest performance, being zero or close to zero in the whole range of training sequence lengths. The performance of the WikiBoC-CLCM approach is also lower than the performance offered by the traditional BoW-MT approach in the whole range of training sequence lengths. However, the performance of the hybrid approach is higher – or at least equal – than the performance offered by the BoW-MT approach for almost every training sequence length, achieving performance increases up to 0.92%. Regarding the Bi-LDA model, the WikiBoC-CLCM and hybrid approaches outperform it in the complete range of training sequences, reaching performance

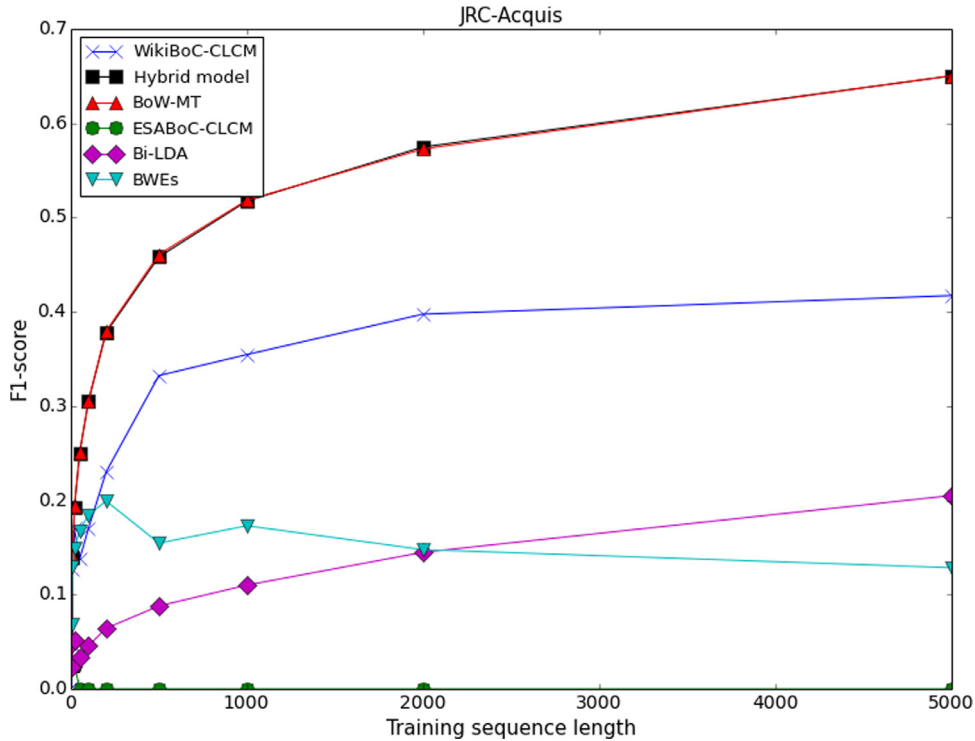


Fig. 7. Evolution of the F1-score value when varying the length of the training sequence in the JRC-Acquis corpus.

Table 4

Evolution of the F1-score value when varying the length of the training sequence in the JRC-Acquis corpus.

	5	10	20	50	100	200	500	1000	2000	5000
WikiBoC-CLCM	0.024	0.127	0.169	0.138	0.170	0.230	0.332	0.354	0.397	0.417
BoW-MT	0.025	0.140	0.192	0.250	0.305	0.378	0.458	0.517	0.575	0.650
ESABoC-CLCM	0.026	0.026	0.026	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Bi-LDA	0.024	0.048	0.051	0.034	0.046	0.064	0.088	0.110	0.145	0.205
BWES	0.069	0.130	0.149	0.168	0.184	0.199	0.155	0.173	0.148	0.128
Hybrid model	0.025	0.140	0.194	0.250	0.306	0.379	0.461	0.518	0.572	0.650

increases up to 304.93% and 636.70%, respectively. Finally, the performance of the WikiBoC-CLCM and hybrid approaches is also higher than the one offered by the BWES approach, achieving increases up to 225.73% and 407.89%, respectively.

6. Discussion

The discussion section is divided into three sections, one for each corpus.

6.1. Wikipedia Corpus

In the Wikipedia Corpus, when the training sequences are short, the WikiBoC-CLCM approach is the one that offers the highest performance. Nevertheless, as we increase the length of the training sequence, the difference in performance between the WikiBoC-CLCM and BoW-MT approaches is reduced, and when the training sequence is very large, the BoW-MT is the most advantageous. Gao et al. [13] and Yang and Pedersen [49] state that the major difficulties of text classification task are the data sparsity and the high dimensionality of the feature space. Several authors affirm that the reduction of the dimensionality alleviates the influence of the data sparsity problem and improves classification performance [21], especially when the amount of training data is small [39]. In line with previous research, we consider that the high performance of the WikiBoC-CLCM approach when the training sequence is short is because the use of concepts reduces the influence of data sparsity. When creating the WikiBoC representations from text documents, we are implicitly doing a dimensionality reduction task, because the number of different dimensions – concepts – in the WikiBoC-CLCM approach is much lower than the number of dimensions – words – in the BoW-MT approach.

Table 5

Concepts extracted by both WikiBoC and ESABoC approaches from the same text document (outliers in bold).

A tarn (or corrie loch) is a mountain lake or pool, formed in a cirque excavated by a glacier. It is formed when either rain or river water fills the cirque. A moraine may form a natural dam below a tarn. The word is derived from the Old Norse word tjrn meaning pond. Its more specific use as a mountain lake emerges as it is the commonly used term for all ponds in the upland areas of Northern England. Here, it retains a broader use, referring to any small lake or pond, regardless of its location and origin (e.g., Talkin Tarn). In Scandinavian languages, a tjern or tjrn, trn or tjrn is a small natural lake, often in a forest or with vegetation closely surrounding it or actually growing into the tarn. Lake Tear of the Clouds (tarn) in the Adirondack Mountains, New York. Velke Hincovo pleso tarn, the biggest and deepest tarn in Slovakia. Lousy Lake (tarn) in N. Cascades National Park, Pickett Range, Washington.		
WikiBoC		ESABoC
Cirque	Forest	Pond
Adirondack Mountains	Loch	Yeadon, West Yorkshire
Tarn (lake)	England	Kosciuszko National Park
Moraine	Rain	Cumbria
Glacier	North Germanic languages	81 (number)
Old Norse	National park	High Street (Lake District)
Lake	Highland	Jean Jaurès
Mountain	Washington (state)	Moraine
River	New York	Sca Fell
Pond	Lake Tear of the Clouds	Montauban
Vegetation	Landslide dam	Castres
Dam	Excavation (archaeology)	Millau Viaduct
Northern England	Swimming pool	Lake District
Water	Northern Europe	Lake
Cascade Range		Robert Stephens
		Hellenistic civilization

The hybrid approach offers a higher performance than BoW-MT in the whole range of training sequences. In line with Huang et al. [18] and Sahlgrén and Cöster [35], the features – in this case concepts – which enrich the BoW representation add useful information for the classifier, thus improving its performance. Evidence suggests that this added information compensates the possible pernicious effect of the increase in dimensionality produced when adding concepts. Even though the hybrid approach outperforms BoW-MT, the performance of the WikiBoC-CLCM approach is still higher when the training sequence is short. This may be due to the extraordinarily negative effect of the high dimensionality problem together with the data sparsity problem – in our case, short training sequences.

The Bi-LDA model offers the second best performance among the state-of-the-art approaches, and although it achieves high F1-score values, the proposed WikiBoC-CLCM and hybrid methods outperform it in the whole range of training sequences. As we said in Section 5.1, the Bi-LDA and BWEs approaches rely on comparable corpora to learn the bilingual features. It seems clear that these corpora have a crucial role in the model. The one used to implement the Bi-LDA and BWEs approaches is composed of a subset of English and Spanish Wikipedia articles. Therefore, it is possible that some topics are not present in the corpora, so they will not be extracted. In order to verify if the comparable corpora have influence on performance, we repeated the Bi-LDA and BWEs experiments using larger comparable corpora, also composed of English and Spanish Wikipedia articles.⁷ Fig. 8(a) and (d) show the performance of the Bi-LDA and BWEs approaches using the comparable corpora proposed by Vulić et al. [42,43] (blue circles) and the performance using larger comparable corpora (red squares) when classifying Wikipedia Corpus. The approach we propose does not use comparable corpora and makes use of the entire Wikipedia as background knowledge, so that all included topics may be potentially extracted.

The worst results are obtained with the ESA concept-based approach (ESABoC-CLCM). The authors of ESA [12] believe that considering the document as a single unit can often be misleading, because its text might be too diverse to be readily mapped to the right set of concepts, while notions mentioned only briefly may be overlooked. Then, they propose to partition the document into a series of non-overlapping segments at different levels (word, phrase, sentence, paragraph, and the entire document), generate features (concepts) for each segment, and finally perform feature selection in order to obtain the best features. This is a costly, laborious and time-consuming process. The approach we propose uses a semantic annotator that is able to extract high-quality concepts from an entire document, without having to perform any process to the text, thus saving time and resources. In order to perform the experiments under the same conditions, we use the entire documents without performing any kind of pre-processing to create both WikiBoC and ESABoC representations of documents. Besides, ESA has tendency to generate outliers [9] which hampers its usefulness in classification. Table 5 shows an example of the concepts extracted by both ESA and the Wikipedia Miner semantic annotator from a text document randomly selected amongst the corpora used. The number of concepts extracted by the WikiBoC approach is bigger than the number of concepts extracted by ESA, and this behaviour is further accentuated the higher the document's length. We can also see that ESA generates a lot of outliers, such as the number 81, the politician *Jean Jaurès*, or the English actor *Robert Stephens*,

⁷ The comparable corpora used are available at <http://linguatoools.org/tools/corpora/wikipedia-comparable-corpora/> (Accessed 25 January 2017).

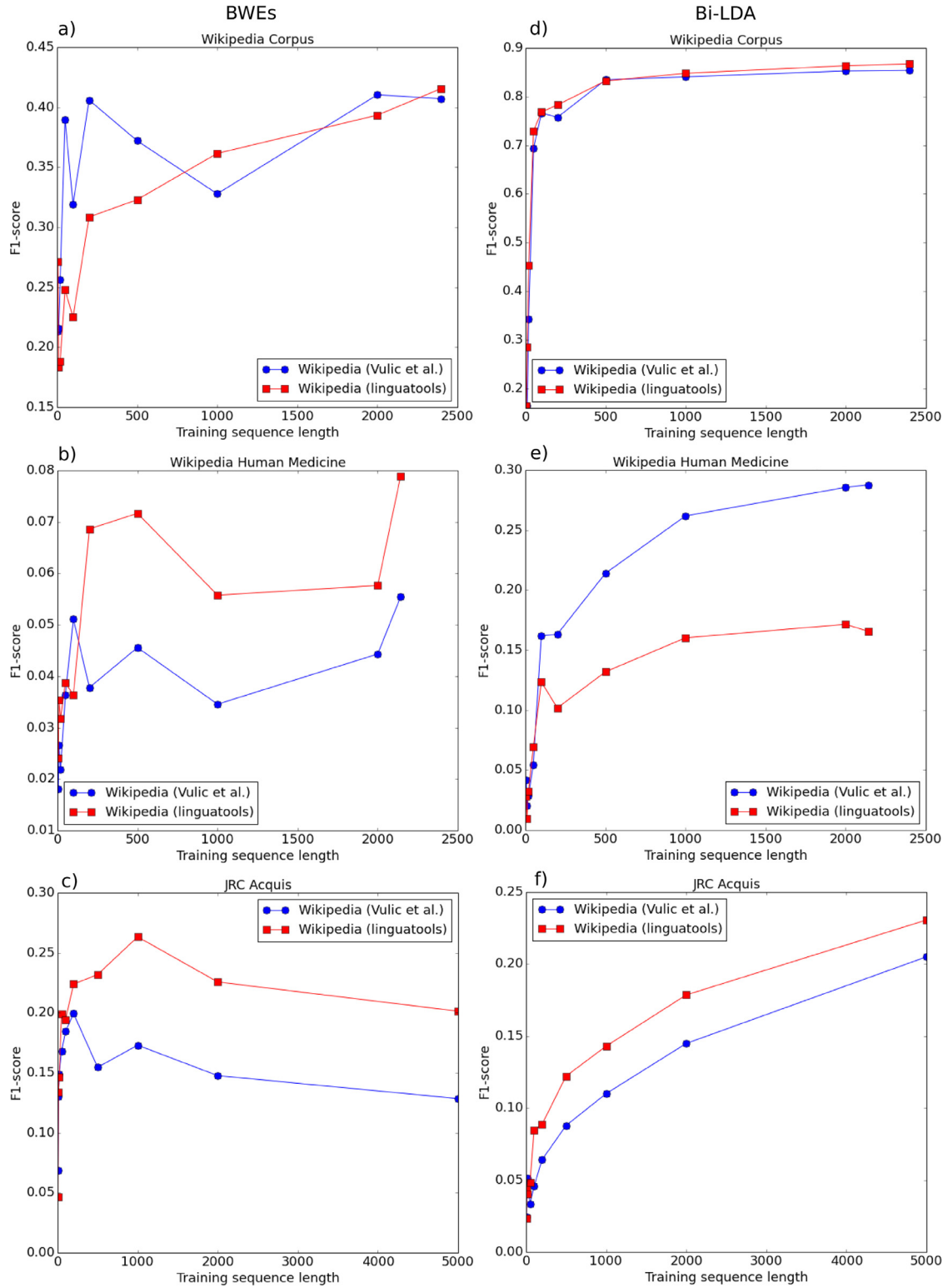


Fig. 8. Comparable corpora influence on BWEs and Bi-LDA approaches. (For interpretation of the references to colour in this figure, the reader is referred to the web version of this article.)

among others. On the contrary, the WikiBoC approach only extracted one outlier: *swimming pool*. Finally, whereas Wikipedia Miner performs entity recognition, and extracts concepts that are explicitly present in text, ESA extraction process is based on surface overlap with Wikipedia articles. This causes WikiBoC concepts to be more related to the text than those extracted by ESA. Therefore, we can clearly see that the quality of concepts extracted by the WikiBoC approach is far superior than the quality of those extracted through ESA, which has a high impact on the performance of the classifier.

6.2. Wikipedia Human Medicine corpus

Regarding the Wikipedia Human Medicine corpus, when training sequences are short, the WikiBoC-CLCM approach is the one that offers the highest performance. In line with the aforementioned, the use of concepts reduces the dimensionality, which alleviates the influence of the data sparsity problem and improves the classification performance. Also, the performance of the WikiBoC-CLCM approach is higher than the offered by the BoW-MT approach in almost the whole range of training sequences, which is not the case for the Wikipedia Corpus, except when training sequences are short.

The hybrid model outperforms the BoW-MT approach in the whole range of training sequences. Besides, the average performance increase of the hybrid approach over BoW-MT is 15,03%, being the average performance increase in the Wikipedia Corpus 2.99%. These results show that there is some type of difference between the two corpora. Several authors demonstrate that the biomedical documents are excellent candidates for the BoC representation [4,29], since they have some distinctive characteristics. Yetisgen-Yildiz and Pratt [50] spotted that medical texts present a high prevalence of long phrases, and medical phrases are prone to synonymy.

The Bi-LDA model offers the second highest performance among the state-of-the art approaches, but unlike in the case of Wikipedia Corpus, the F1-score values are very low and they are much further away from those offered by the methods we propose. The WikiBoC-CLCM and hybrid models outperform Bi-LDA and the BWEs models in the complete range of training sequences. As we previously stated, comparable corpora are key to Bi-LDA and BWEs models. It seems the corpora used are not able to provide the granularity and specificity required to adequately extract domain-specific topics. In the Wikipedia Corpus case, Bi-LDA performs relatively well since the classification problem is easier and the documents are about more general topics. Nevertheless, it does not perform so well in classification tasks that involve documents about more specific topics. The same occurs with the BWEs approach, whose performance is much lower than the one offered when classifying the Wikipedia Corpus. Fig. 8(b) shows the variation in performance of the BWEs approach depending on the comparable corpora used. The evidence suggests that corpora used have an influence in the performance of the model, larger corpora offering higher performance. Fig. 8(e) shows the difference in performance depending on the corpora used in the Bi-LDA model. In this case the use of larger corpora provides lower performance. This kind of situations do not happen with the approach proposed, since it uses all Wikipedia knowledge to represent the documents, which includes from the most general to the most specific topics.

Again, in the same way aforementioned, the weak results obtained by the ESABoC-CLCM approach are due to its poor performance when extracting concepts from entire documents.

6.3. JRC-Acquis

Among the three corpora proposed, the task of classifying the JRC-Acquis is undoubtedly the most difficult, since each document can be classified into one or several categories.

The performance offered by the WikiBoC-CLCM approach is lower than the performance offered by the traditional BoW-MT approach in the whole range of training sequence lengths. However, the performance of the hybrid model is higher – or at least equal – than the performance offered by the BoW-MT approach for almost every training sequence length. It is worth noting that the performance increase achieved is much lower than that reached with the other two corpora. This is in line with the aforementioned. The major difficulties of text classification task are the data sparsity and the high dimensionality of the feature space. Although when creating concept representations from text documents we are implicitly doing a dimensionality reduction task, it is not sufficient for this purpose, because the documents in the JRC-Acquis corpus are much larger than the documents in Wikipedia corpora, which directly involves a larger number of concepts. Anyway, the concepts which enrich the BoW representation add valuable information for the classifier.

As well as in the Wikipedia Human Medicine corpus, the F1-score values obtained by Bi-LDA and BWEs approaches are very low, and the two approaches we propose outperform them. It seems the reason of the bad performance of the Bi-LDA and BWEs approaches is the same as in the Wikipedia Human Medicine corpus case: the comparable corpora used are not able to provide the sufficient granularity to extract domain-specific topics to adequately represent the documents, all of them of legal nature. Fig. 8(c) and (f) show that the use of larger comparable corpora increases the performance of the Bi-LDA and BWEs approaches.

Once again, the ESABoC-CLCM approach offers the worst results, providing performance values close to zero, since ESA fails when extracting concepts from entire documents. Besides, the more extensive is the document, the lower is the performance offered by ESA, which is consistent with the results obtained with the JRC-Acquis corpus, the one that contains the most extensive documents.

Table 6

First ten Wikipedias ordered by number of articles.

#	Language	#articles	#	Language	#articles
1	English	5,330,410	6	French	1,834,986
2	Cebuano	3,830,240	7	Russian	1,367,908
3	Swedish	3,782,236	8	Italian	1,328,862
4	German	2,023,412	9	Spanish	1,309,956
5	Dutch	1,891,282	10	Waray	1,262,078

7. Limitations

A clear drawback of our approach is the possible loss of information during the *cross-language concept matching* (CLCM) process. This is because not all Wikipedia articles written in a particular language have their correspondence in another different language. Table 6 shows the first ten Wikipedias ordered by number of articles⁸, where it can clearly be seen that there are large differences between in the amount of articles of each edition of Wikipedia. Even though in our approach the CLCM process converts BoC representations from a “small” feature space – 1,309,956 Spanish Wikipedia articles – to a larger feature space – 5,330,410 English Wikipedia articles – there are still many Spanish articles that not have correspondence in English, such as local people or small places.

8. Conclusions

This article has presented the application of the Wikipedia Miner bag of concepts document representation (WikiBoC) to cross-language text classification (CLTC). To accomplish this, we propose a technique, called cross-language concept matching (CLCM), which converts the concept-based representation of documents from one language to another without performing any kind of translation. The results of the experiments conducted allow us to show the benefits of using a Wikipedia Miner concept-based representation along with the CLCM technique proposed. On the one hand, the proposal 1, purely based-on-concepts, is the one that offers the highest performance with short training sequences, outperforming state-of-the-art approaches, both word-based and concept-based. On the other hand, the proposal 2, based on a combination of words and concepts, outperforms the state-of-the-art approaches regardless of the length of the training sequence. Then, this advantage can be leveraged by using only a purely concept-based document representation in applications where training data is limited, or using the hybrid approach when the training data is ample. In any case, the use of the WikiBoC-CLCM approach proposed is advantageous, since the concepts extracted through the semantic annotator add valuable information which is useful for the classification algorithm, thus improving its performance, especially with documents in the biomedical field.

Although this study was conducted using English and Spanish documents, it is easily extensible to any language, simply by using a different language version of Wikipedia as the background knowledge base for the Wikipedia Miner semantic annotator. Therefore, this study could be extended by classifying documents in other languages, using as training sequence English documents or documents written in another language. For example, it could be interesting to verify the performance of the approach proposed when classifying documents written in minority languages, which generally keep smaller Wikipedias. Finally, it is worth noting that several previous studies demonstrate that the bag of concepts document representation provides good results in information retrieval tasks, which makes the approach proposed in this paper promising for performing cross-language information retrieval.

References

- [1] M. Amini, N. Usunier, C. Goutte, Learning from multiple partially observed views—an application to multilingual text categorization, in: *Proceedings of the Advances in Neural Information Processing Systems*, 2009, pp. 28–36.
- [2] Y. Bengio, R. Ducharme, P. Vincent, C. Jauvin, A neural probabilistic language model, *J. Mach. Learn. Res.* 3 (Feb) (2003) 1137–1155.
- [3] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, *J. Mach. Learn. Res.* 3 (2003) 993–1022.
- [4] S. Bloehdorn, A. Hotho, Boosting for text classification with semantic features, in: *Proceedings of the Sixth International Conference on Knowledge Discovery on the Web*, Springer, 2004, pp. 149–166.
- [5] F. Colace, M. De Santo, L. Greco, P. Napoletano, Improving relevance feedback-based query expansion by the use of a weighted word pairs approach, *J. Assoc. Inf. Sci. Technol.* 66 (11) (2015) 2223–2234.
- [6] W. De Smet, J. Tang, M.-F. Moens, Knowledge transfer across multilingual corpora via latent topics, *Lect. Notes Comput. Sci.* 6634 (2011) 549–560.
- [7] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, *J. Am. Soc. Inf. Sci.* 41 (6) (1990) 391.
- [8] K. Duh, A. Fujino, M. Nagata, Is machine translation ripe for cross-lingual sentiment classification? in: *Proceedings of the Forty-Ninth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers*, 2, Association for Computational Linguistics, 2011, pp. 429–433.
- [9] O. Egozi, S. Markovitch, E. Gabrilovich, Concept-based information retrieval using explicit semantic analysis, *ACM Trans. Inf. Syst. (TOIS)* 29 (2) (2011) 8.
- [10] European Union, EuroVoc: Multilingual Thesaurus of the European Union (version 4.4), Technical Report, 2015. Publications Office of the European Union. ISSN 1830-6500.

⁸ https://en.wikipedia.org/wiki/List_of_Wikipedias (Accessed 25 January 2017).

- [11] B. Fortuna, J. Shawe-Taylor, The use of machine translation tools for cross-lingual text mining, in: *Proceedings of the ICML Workshop on Learning with Multiple Views*, 2005.
- [12] E. Gabrilovich, S. Markovitch, Wikipedia-based semantic interpretation for natural language processing, *J. Artif. Intell. Res.* (2009) 443–498.
- [13] L. Gao, S. Zhou, J. Guan, Effectively classifying short texts by structured sparse representation with dictionary filtering, *Inf. Sci. (NY)* 323 (2015) 130–142.
- [14] A. Gliozzo, C. Strapparava, Cross language text categorization by acquiring multilingual domain models from comparable corpora, in: *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Association for Computational Linguistics, 2005, pp. 9–16.
- [15] Y. Guo, M. Xiao, Cross language text classification via subspace co-regularized multi-view learning, in: *Proceedings of the 29th International Conference on Machine Learning (ICML12)*, 2012, pp. 1615–1622.
- [16] M.S. Hajmohammadi, R. Ibrahim, A. Selamat, H. Fujita, Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples, *Inf. Sci. (NY)* 317 (2015) 67–77.
- [17] M.A. Hearst, S.T. Dumais, E. Osman, J. Platt, B. Scholkopf, Support vector machines, *IEEE Intell. Syst. Appl.* 13 (4) (1998) 18–28.
- [18] A. Huang, D. Milne, E. Frank, I.H. Witten, Clustering documents using a Wikipedia-based concept representation, in: *Advances in Knowledge Discovery and Data Mining*, Springer, 2009, pp. 628–636.
- [19] L. Huang, D. Milne, E. Frank, I.H. Witten, Learning a concept-based document similarity measure, *J. Am. Soc. Inf. Sci. Technol.* 63 (8) (2012) 1593–1608.
- [20] W.J. Hutchins, H.L. Somers, *An Introduction to Machine Translation*, 362, Academic Press London, 1992.
- [21] H. Kim, P. Howland, H. Park, Dimension reduction in text classification with support vector machines, *J. Mach. Learn. Res.* (2005) 37–53.
- [22] A. Klementiev, I. Titov, B. Bhattarai, Inducing crosslingual distributed representations of words, in: *Proceedings of COLING 2012: Technical Papers*, 2012, pp. 1459–1474.
- [23] X. Ling, G.-R. Xue, W. Dai, Y. Jiang, Q. Yang, Y. Yu, Can Chinese web pages be classified with English data source? in: *Proceedings of the Seventeenth International Conference on World Wide Web*, ACM, 2008, pp. 969–978.
- [24] B. Lu, C. Tan, C. Cardie, B.K. Tsou, Joint bilingual sentiment classification with unlabeled parallel corpora, in: *Proceedings of the Forty-Ninth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1, Association for Computational Linguistics, 2011, pp. 320–330.
- [25] M. Maleshkova, L. Zilka, P. Knöth, C. Pedrinaci, Cross-lingual web API classification and annotation, in: *Proceedings of the Second Workshop on the Multilingual Semantic Web*, 2011, p. 1.
- [26] D. Milne, I.H. Witten, An open-source toolkit for mining Wikipedia, *Artif. Intell.* 194 (2013) 222–239.
- [27] Z.-Y. Ming, T.S. Chua, Resolving polysemy and pseudonymity in entity linking with comprehensive name and context modeling, *Inf. Sci. (NY)* 307 (2015) 18–38.
- [28] S. Montalvo, R. Martínez, A. Casillas, V. Fresno, Multilingual news clustering: feature translation vs. identification of cognate named entities, *Pattern Recognit. Lett.* 28 (16) (2007) 2305–2311.
- [29] M.A. Mouriño García, R. Pérez Rodríguez, L.E. Anido Rifón, Biomedical literature classification using encyclopedic knowledge: a Wikipedia-based bag-of-concepts approach, *PeerJ*, 3 (2015) e1279.
- [30] X. Ni, J.-T. Sun, J. Hu, Z. Chen, Cross lingual text classification by mining multilingual topics from Wikipedia, in: *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ACM, 2011, pp. 375–384.
- [31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: machine learning in Python, *J. Mach. Learn. Res.* 12 (2011) 2825–2830.
- [32] M.F. Porter, An algorithm for suffix stripping, *Program* 14 (3) (1980) 130–137.
- [33] L. Rigutini, M. Maggini, B. Liu, An em based training algorithm for cross-language text categorization, in: *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, IEEE, 2005, pp. 529–535.
- [34] M. Sahlgren, The distributional hypothesis, *Ital. J. Linguist.* 20 (1) (2008) 33–54.
- [35] M. Sahlgren, R. Cöster, Using bag-of-concepts to improve the performance of support vector machines in text categorization, in: *Proceedings of the Twentieth International Conference on Computational Linguistics*, Association for Computational Linguistics, 2004, p. 487.
- [36] F. Sebastiani, Machine learning in automated text categorization, *ACM Comput. Sur. (CSUR)* 34 (1) (2002) 1–47.
- [37] L. Shi, R. Mihalcea, M. Tian, Cross language text classification by model translation and semi-supervised learning, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2010, pp. 1057–1067.
- [38] R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, D. Varga, The JRC-Acquis: a multilingual aligned parallel corpus with 20+ languages, in: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006)*, 2006, pp. 2142–2147.
- [39] K. Toutanova, F. Chen, K. Papat, T. Hofmann, Text classification in a hierarchical mixture model for small training sets, in: *Proceedings of the Tenth International Conference on Information and Knowledge Management*, ACM, 2001, pp. 105–113.
- [40] C.-T. Tsai, D. Roth, Cross-lingual wikification using multilingual embeddings, in: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, San Diego, California, 2016, pp. 589–598.
- [41] S. Upadhyay, M. Faruqi, C. Dyer, D. Roth, Cross-lingual models of word embeddings: an empirical comparison, in: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2016, pp. 1661–1670.
- [42] I. Vulić, W. De Smet, J. Tang, M.-F. Moens, Probabilistic topic modeling in multilingual settings: an overview of its methodology and applications, *Inf. Process. Manag.* 51 (1) (2015) 111–147.
- [43] I. Vulić, M.-F. Moens, Bilingual distributed word representations from document-aligned comparable data, *J. Artif. Intell. Res.* 55 (2016) 953–994.
- [44] C. Wan, R. Pan, J. Li, Bi-weighting domain adaptation for cross-language text classification, in: *Proceedings of the International Joint Conference on Artificial Intelligence*, 22, 2011, p. 1535.
- [45] X. Wan, Co-training for cross-lingual sentiment classification, in: *Proceedings of the Joint Conference of the Forty-Seventh Annual Meeting of the ACL and the Fourth International Joint Conference on Natural Language Processing of the AFNLP*, 1, Association for Computational Linguistics, 2009, pp. 235–243.
- [46] P. Wang, J. Hu, H.-J. Zeng, Z. Chen, Using Wikipedia knowledge to improve text classification, *Knowl. Inf. Syst.* 19 (3) (2009) 265–281.
- [47] B. Wei, C. Pal, Cross lingual adaptation: an experiment on sentiment classifications, in: *Proceedings of the ACL Conference Short Papers*, Association for Computational Linguistics, 2010, pp. 258–262.
- [48] L. Wittgenstein, *Philosophical Investigations*, Blackwell Publishing, 1953.
- [49] Y. Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, in: *Proceedings of the International Conference on Machine Learning*, 97, 1997, pp. 412–420.
- [50] M. Yetisgen-Yildiz, W. Pratt, The effect of feature representation on medline document classification, in: *Proceedings of the AMIA Annual Symposium*, 2005, American Medical Informatics Association, 2005, p. 849.