Information Sciences 406–407 (2017) 12–28

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

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.

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).

http://dx.doi.org/10.1016/j.ins.2017.04.024
0020-0255/

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 (Bi-LDA), 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.