Artificial Intelligence in Medicine xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Leveraging Wikipedia knowledge to classify multilingual biomedical documents

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

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

ARTICLE INFO

Article history:
Received 11 August 2017
Received in revised form 6 April 2018
Accepted 23 April 2018

Keywords:
Biomedical document classification
Hybrid word-concept document representation
Multilingual text classification
Wikipedia-based bag of concepts document representation
Wikipedia Miner semantic annotator

ABSTRACT

This article presents a classifier that leverages Wikipedia knowledge to represent documents as vectors of concepts weights, and analyses its suitability for classifying biomedical documents written in any language when it is trained only with English documents. We propose the cross-language concept matching technique, which relies on Wikipedia interlanguage links to convert concept vectors between languages. The performance of the classifier is compared to a classifier based on machine translation, and two classifiers based on MetaMap. To perform the experiments, we created two multilingual corpus. The first one, Multi-Lingual UVigoMED (ML-UVigoMED) is composed of 23,647 Wikipedia documents about biomedical topics written in English, German, French, Spanish, Italian, Galician, Romanian, and Icelandic. The second one, English-French-Spanish-German UVigoMED (EFSG-UVigoMED) is composed of 19,210 biomedical abstract extracted from MEDLINE written in English, French, Spanish, and German. The performance of the approach proposed is superior to any of the state-of-the art classifier in the benchmark. We conclude that leveraging Wikipedia knowledge is of great advantage in tasks of multilingual classification of biomedical documents.

1. Introduction

The biomedical domain stands out as a relevant area where to apply automatic document classification techniques, given the huge amount of biomedical information readily available on the Internet. Biomedical documents may come from multiple sources and they may be written in different languages [1]. Even though English is the principal language for publishing biomedical research, research conducted in languages other than English cannot be underestimated [2]. In addition, information managed by clinical and research centers (e.g. medical records) it is usually written in the official language of each country or region.

Document classification is commonly modeled as a supervised machine learning problem: a classification algorithm is trained with a labeled set of documents, and it is later applied to a set of unlabeled documents [3]. Algorithmic classification may perform poorly when there are not enough documents in a particular language to train the classifier [4], and it is in this scenario when cross-language text classification (CLTC) becomes relevant. It consists in training a classifier with a labeled set of documents written in a language L_1 – where large training sequences are available – to classify a set of unlabeled documents written in a different language L_2, where there is not a set of documents large enough to train the classifier.

Text documents have to be represented in a way that classifiers can understand them. The most used representation is the bag of words (BoW) model [5], 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 documents has traditionally been approached by using the BoW representation together with machine translation (MT) techniques, either translating the documents before extracting the set of features [6,7], or translating the features themselves [8,9]. Both approaches suffer from a series of drawbacks related to both the BoW representation and MT techniques. On the one hand, the BoW representation is suboptimal, since it only accounts for word frequency in text, which involves the emergence of two language-related problems that affect the classification performance: redundancy (synonymy problem) and ambiguity (polysemy problem) [10]. On the other hand, MT techniques have a major drawback: the ambiguity [9,11], which negatively affects the quality of translations. Ambiguity is in turn divided into (i) lexical ambiguity, that is to say, a word in one

* Corresponding author at: Telecommunication Engineering School, Department of Telematics Engineering, University of Vigo, Campus Lagoas-Marcosende, 36310 Vigo, Spain.
E-mail addresses: marcos@gist.uvigo.es (M. Antonio Mouriño García), roberto.perez@gist.uvigo.es (R. Pérez Rodríguez), lanido@gist.uvigo.es (L. Anido Rifón).

https://doi.org/10.1016/j.artmed.2018.04.007
0933-3657/

language can have multiple translations in another language, with several different meanings; and (ii) structural ambiguity, when there is more than one way of analyzing the underlying structure of a sentence according to the grammar used. Thus, if an incorrect translation is selected, it can distort the precision of the classifier due to the introduction of erroneous features. Therefore, when the BoW representation is combined with MT techniques, the drawbacks of each one add up, which leads to an increased error probability.

Aiming at overcoming the problems associated with BoW representations, several authors explored a different approach: the bag of concepts (BoC) representation, where a concept is a 'unit of meaning' [12]. Documents are thus represented as vectors of weighted concepts, in accordance with their relevance in the texts. Approaches to mapping text documents to concepts include those that make use only of information in the documents used for training the algorithm [13,14], and those that use external knowledge sources [15,16].

In our view, there exists a research gap in the application of concept-based representations that leverage Wikipedia knowledge to create multilingual classifiers of biomedical text documents. This paper aims at bridging that gap by describing the foundations and reporting the evaluation results of a multilingual biomedical document classifier that leverages Wikipedia knowledge to represent documents as vectors of concepts weights (concretely by using the Wikipedia Miner semantic annotator), and to convert concept vectors between languages. To that end, we propose the cross-language concept matching technique (CLCM), which converts the BoC representation of a document between languages, by leveraging Wikipedia interlanguage links. We call the proposed classifier WikiBoC-CLCM. Furthermore, we also propose a hybrid model (Hybrid-WikiBoC) that combines the BoW-MT and WikiBoC-CLCM approaches.

To evaluate the classifiers proposed we performed benchmarking by comparing their performance to three state-of-the-art approaches: a classifier based on the BoW representation along with MT techniques (BoW-MT), and two classifiers that uses the domain-specific semantic annotator MetaMap. Since there is not available a standard corpus for evaluating multilingual classification of biomedical documents [9], we expressly create two multilingual corpora. The first one, Multi-Lingual UVigoMED (ML-UVigoMED),[1] is composed of Wikipedia articles about biomedical topics. The second one, English-French-Spanish-German UVigoMED (EFSG-UVigoMED)[2] is composed of biomedical abstracts extracted from MEDLINE.

We consider the main contributions of this work are the following:

- The WikiBoC-CLCM and Hybrid-WikiBoC approaches for classifying biomedical text documents.
- The benchmarking of the state-of-the-art classification approaches.
- The ML-UVigoMED and EFSG-UVigoMED corpora.

The remainder of this article is organized as follows: Section 2 reviews the most relevant state-of-the-art approaches to perform multilingual classification of biomedical documents. Section 3 presents the Support Vector Machines algorithm, the Wikipedia Miner semantic annotator, the cross-language concept matching technique, and the description and the generation process of the ML-UVigoMED and EFSG-UVigoMED corpora. Section 4 exposes the two approaches proposed: WikiBoC-CLCM and Hybrid-WikiBoC. 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 and proposals for future work.

2. Literature review

The number of works about multilingual classification of biomedical documents is much lower than the amount of work about text classification in general [7]. Therefore, we include works about bilingual or cross-lingual classification as a particular case of multilingual classification. In this section, we briefly review (i) published studies grouped in accordance with the method followed to represent documents – bag of words and bag of concepts – and (ii) studies about multilingual classification of biomedical documents.

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. The main approach to cross-language text classification is Cross-Lingual Training. Cross-lingual training classifiers are trained on a corpus of translated documents [6], leveraging machine translation [11] tools such as Google Translate. A variation of this approach consists in translating the features extracted from documents instead [8,9], either during the training or the classification phase.

2.2. Classifiers based on bag of concepts representations

This type of classifiers represents documents as weighted concept vectors. The main proposals existing in the literature are Latent Semantic Analysis (LSA) [13], Latent Dirichlet Allocation (LDA) [14], Explicit Semantic Analysis (ESA) [15], Word Embeddings (WE) [17,18] and semantic annotators [16].

The Latent Semantic Analysis (LSA) technique is based on the distributional hypothesis [19], which exposes that the words that occur in similar contexts tend to have similar meanings. In the LSA model, the meaning of a word is captured as the vector of occurrences in different contexts, being a context a text document.

The Latent Dirichlet Allocation (LDA) model assumes that each document within a collection comprises a small number of topics, each of them "generating" words. Thus, LDA automatically finds topics in texts by "going back" from the document and finds the set of topics that may have generated it.

The Explicit Semantic Analysis (ESA) technique leverages external knowledge sources such as Wikipedia to generate features from text documents. Unlike LSA and LDA, ESA performs textual analysis identifying topics that are explicitly present in background knowledge bases instead of latent topics. In other words, ESA analyze the text to index it with concepts from the knowledge base used.

Word embeddings, also known as distributed representations of words, are a set of language modeling and feature learning techniques where words or phrases from a vocabulary are mapped to dense real-valued vectors, serving as rich and coherent word representations. Conceptually, it involves a mathematical embedding from a space with one dimension per word to a continuous vector space with much lower dimension.

Several authors, such as [14,20–22], report positive results when using the aforementioned concept-based approaches to perform monolingual classification of biomedical documents.

The approach we propose is based on semantic annotators, software elements responsible for creating the bag of concepts

[1] The corpus is freely available at http://www.itec-sde.net/ML-UVigoMED.zip (accessed 6.04.18).
[2] The corpus is freely available at http://www.itec-sde.net/EFSG-UVigoMED.zip (accessed 6.04.18).

Please cite this article in press as: Antonio Mouriño García M, et al. Leveraging Wikipedia knowledge to classify multilingual biomedical documents. Artif Intell Med (2018), https://doi.org/10.1016/j.artmed.2018.04.007