Knowledge-Based Systems 81 (2015) 65–75

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Choosing the best dictionary for Cross-Lingual Word Sense Disambiguation

Andres Duque *, Juan Martinez-Romo, Lourdes Araujo

NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Madrid 28040, Spain

ARTICLE INFO

Article history:
Received 23 July 2014
Received in revised form 6 February 2015
Accepted 7 February 2015
Available online 14 February 2015

Keywords:
Word Sense Disambiguation
Bilingual dictionary
Natural language processing
Unsupervised systems
Parallel corpus

ABSTRACT

The choice of the dictionary that provides the possible translations a system has to choose when performing Cross-Lingual Word Sense Disambiguation (CLWSD) is one of the most important steps in such a task. In this work, we present a comparison between different dictionaries, in two different frameworks. First of all, a technique for analysing the potential results of an ideal system using those dictionaries is developed. The second framework considers the particular unsupervised CLWSD system CO-Graph, and analyses the results obtained when using different bilingual dictionaries providing the potential translations. Two different CLWSD tasks from the 2010 and 2013 SemEval competitions are used for evaluation, and statistics from the words in the test datasets of those competitions are studied. The conclusions of the analysis of dictionaries on a particular system lead us to a proposal that substantially improves the results obtained in that framework. In this proposal a hybrid system is developed, by combining the results provided by a probabilistic dictionary, and those obtained with a Most Frequent Sense (MFS) approach. The hybrid approach also outperforms the results obtained by other unsupervised systems in the considered competitions.

1. Introduction

Cross-Lingual Word Sense Disambiguation (CLWSD) can be defined as the task of automatically determining the contextually appropriate translation for a given word, from a source language to a target one. This is a particular case of the Word Sense Disambiguation (WSD) problem, which has been widely studied in the NLP community [11]. WSD is an essential and necessary step for many processes, such as automatic summarisation, information retrieval, topic detection, and in general, any NLP process in which the semantic level of the words is important. WSD has been frequently treated as a supervised learning problem [19,22], based on techniques that depend on semantically tagged corpora or lexical databases like Wordnet [8]. On the other hand, unsupervised techniques, also known as Word Sense Induction (WSI) techniques, do not require those kinds of resources. Their objective is to induce the different senses of a specific word in a given text by selecting groups of words related to a particular sense of the word. The motivation of the CLWSD task comes from the scarcity of sense inventories and sense-tagged corpora, and the need to evaluate the performance of WSD systems in real problems [14].

A Cross-Lingual Word Sense Disambiguation task proposes a set of instances in which a target word can be found. This target word needs to be disambiguated, from an original language (typically English) to a final one. Fig. 1 illustrates this task with an example. The bilingual dictionary that provides translations, both for words surrounding the target word (context) and for the target word itself, is a key part of the disambiguation process. This dictionary offers the potential translations of the target word, and any system which performs the disambiguation has to choose, among the translations, those which are considered most suitable for the particular sentence. This selection is then matched against an expected output or gold standard to determine a score for that specific test instance. In this example, the context taken into account for performing the disambiguation is only composed by nouns, although any other word (e.g. verbs, adjectives) can also be considered.

Many issues arise along the disambiguation process, the choice of an adequate bilingual dictionary being one of the most important for ensuring the good performance of a system. We compare the use of bilingual dictionaries of different nature: manually created by experts, semi-automatic, i.e. extracted with automatic tool but with human supervision or intervention, collaboratively edited by different authors, and statistical dictionaries. This last type of dictionaries, automatically created without human supervision, provide a much larger number of translations, at the price of introducing noise. However, apart from their size and the coverage they can

* Corresponding author.
E-mail addresses: aduque@lsi.uned.es (A. Duque), juaner@lsi.uned.es (J. Martinez-Romo), lurdes@lsi.uned.es (L. Araujo).

http://dx.doi.org/10.1016/j.knosys.2015.02.007
0950-7051/© 2015 Elsevier B.V. All rights reserved.

present (denoted by the number of different translations for each word), this kind of dictionaries provide information about the translation probabilities, since their construction is based on statistical characteristics. The other dictionaries do not usually present this kind of information. Considering that CLWSD tasks are based on translations of words used in general sentences, we can expect that information about the most frequent translations would be useful.

In this work, we analyse different dictionaries that provide the candidate translations, and compare the results obtained using them, both in ideal conditions, and inside a particular unsupervised CLWSD system [7]. These results show the potential variations of the effectiveness of the CLWSD system according to the choice of the bilingual dictionary.

1.1. Background work

For the purposes of this work, we have selected some evaluation tasks related to Cross-Lingual Word Sense Disambiguation, as a framework in which the effect of the selected dictionary can be tested. Specifically, we have selected task 3 of 2010 SemEval competition [14] and task 10 of 2013 SemEval competition [15], both of them based on the Europarl parallel corpus [12]. Many different systems were proposed for these two tasks, and the use of bilingual dictionaries is a common practice inside the proposed algorithms, both for supervised and unsupervised systems. The OWNS system [18] is a supervised system which participated in the 2010 SemEval competition. It uses nearest neighbours classifiers based on pairwise similarity measures. Most of its lexical information is extracted from WordNet [8], although it uses a noisy statistical dictionary learnt from the Europarl corpus for proposing possible translations. Other supervised methods also participated in the 2010 competition: UvT-WSD [32], applying the K-NN algorithm, and FCC [34], using a Naive Bayes classifier. In those cases, the tool used for extracting bilingual dictionaries was GIZA++ [26], which has proven to be the preferred tool for aligning the corpus at word level and extracting translations. Regarding unsupervised systems participating in the 2010 competition, in [30], a co-occurrence graph based on the aligned contexts of the target word is built for performing the disambiguation. This graph aggregates words from different languages and the disambiguation is made through the extraction of the minimum spanning tree. In this work, multilingual dictionaries such as EuroWordNet [35], and PanDictionary [21] are proposed for extracting translations, frequencies and characteristics. The other unsupervised system of the 2010 competition, T3-COLEUR [10] is based on probability tables extracted from the Europarl corpus, and also uses a GIZA-based bilingual dictionary. In this competition, the best results for the Spanish language were obtained by the supervised system UvT-WSD, while the best unsupervised system was T3-COLEUR.

In regard to the 2013 competition, the only system that did not make use of the GIZA++ tool was the supervised system HLDTI [28]. It used maximum entropy classifiers, trained on local context features, to perform the disambiguation, and the aligning tool selected for extracting translations was the Berkeley Aligner [6]. The other systems of this competition used GIZA-based dictionaries, independently of the final languages of the translations. In this group, we can find supervised systems such as WSD2 [33], the new version of the UvT-WSD also based on a K-NN classifier. Unsupervised systems also used this resource: LIMSI [2] addressed the problem by using vectors of features extracted from the corpus. XLING [31] generated topic models from the source corpus using Latent Dirichlet Allocation (LDA) [4]. The main hypothesis is that the different senses of a target word will be classified into different topics by the LDA algorithm. The NRC-SMT system [5] uses a statistical machine translation approach, extracting knowledge only from the Europarl corpus in its first run, and adding information from news data in a second run of the system. In the 2013 competition also a supervised system, HLDTI, obtained the best results. The best unsupervised system was LIMSI.

Finally, we can find other systems that did not participate in any of the competitions, although they present results for some of the proposed datasets: the ParaSense system [16] is a supervised, memory-based algorithm that builds different classifiers using both local context features and binary bag-of-words features. Unsupervised systems as the multilingual system described in [25] also addressed the problem without participating in the competitions. This system exploits the multilingual knowledge base BabelNet [24], for performing WSD and CLWSD, obtaining very competitive results. Both works make use of the GIZA++ tool, the first one as a main aligner for extracting a bilingual dictionary, and the second one for proposing the most frequent sense translations when no sense assignment is attempted.

1.2. Main objectives

In this work we analyse the effect of bilingual dictionaries, both inside an ideal system, and a particular CLWSD system, named CO-Graph. This system is based on an unsupervised algorithm for extracting co-occurrence graphs from text documents [20]. In this case, we focus on the English–Spanish cross-lingual disambiguation, and on the out-of-five evaluation proposed in both SemEval tasks already mentioned. This evaluation scheme requires the systems to provide up to five guesses for each target word in each context, without penalising them due to the number of guesses.

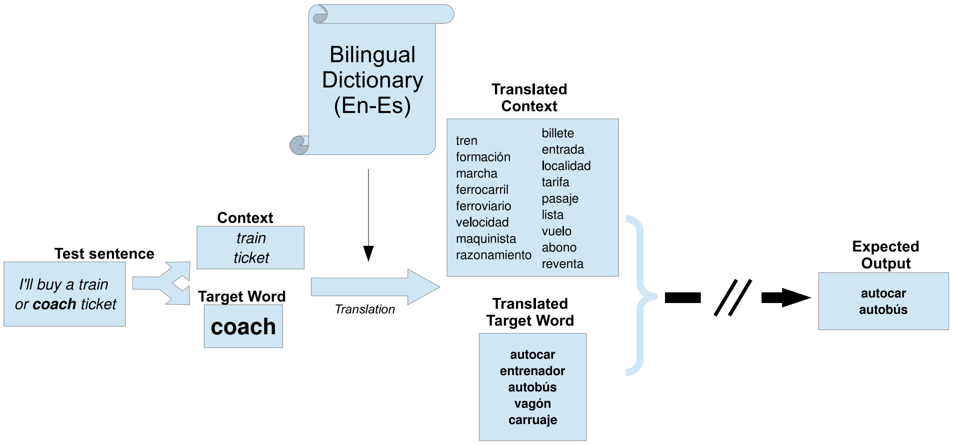

Fig. 1. Example of a general disambiguation process of a sentence containing the target word coach, with Spanish as target language.