Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 39 (2016) 88–107

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Sentence alignment using local and global information☆

Hamed Zamani [a], Heshaam Faili [a,b], Azadeh Shakery [a,b,*]

[a] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

[b] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), P.O. Box 19395-5746, Tehran, Iran

Received 15 June 2015; received in revised form 28 February 2016; accepted 23 March 2016
Available online 1 April 2016

Abstract

Parallel corpora are essential resources for statistical machine translation (SMT) and cross language information retrieval (CLIR) systems. Creating parallel corpora is highly expensive in terms of both time and cost. In this paper, we propose a novel approach to automatically extract parallel sentences from aligned documents. To do so, we first train a Maximum Entropy binary classifier to compute the local similarity between each two sentences in different languages. To consider global information (e.g., the position of sentence pairs in the aligned documents), we define an objective function to penalize the cross alignments and then propose an integer linear programming approach to optimize the objective function. In our experiments, we focus on English and Persian Wikipedia articles. The experimental results on manually aligned test data indicate that the proposed method outperforms the baselines, significantly. Furthermore, the extrinsic evaluations of the corpus extracted from Wikipedia on both SMT and CLIR systems demonstrate the quality of the extracted parallel sentences. In addition, Experiments on the English–German language pair demonstrate that the proposed ILP method is a language-independent sentence alignment approach. The extracted English–Persian parallel corpus is freely available for research purposes.

Keywords: Parallel corpus; Sentence alignment; Bilingual resource; Global information; Integer linear programming

1. Introduction

Statistical translation techniques have been shown to perform promising when they are trained using a large amount of high quality data. One essential category of resources for training statistical machine translation (SMT) and cross language information retrieval (CLIR) systems is large and accurate parallel texts (Brown et al., 1991; Vogel and Tribble, 2002). It is shown that training translation techniques via sentence-level aligned corpora, henceforth called *parallel corpora*, is more effective than using other types of parallel texts (Ma, 2006). Parallel corpora contain several sentence pairs in two (source and target) languages, which are translations of each other (Gale and Church, 1993).

It is known that manual sentence alignment is very expensive in terms of both time and cost. This fact has motivated researchers to automatically extract parallel sentences from the available bilingual data (Gale and Church, 1993; Wu, 1994; Li et al., 2010; Pal et al., 2014). The exponential growth of the Web, and thus the availability of huge amounts

☆ This paper has been recommended for acceptance by Roger K. Moore.

[*] Corresponding author. Tel.: +98 2182089722.
E-mail addresses: h.zamani@ut.ac.ir (H. Zamani), hfaili@ut.ac.ir (H. Faili), shakery@ut.ac.ir (A. Shakery).

http://dx.doi.org/10.1016/j.csl.2016.03.002
0885-2308/© 2016 Elsevier Ltd. All rights reserved.

of textual data in various languages intensifies the importance of automatic sentence alignment (Resnik and Smith, 2003).

A main part of automatic sentence aligners is computing the similarity between the source and the target sentences. Several methods have been so far proposed to compute length-based, lexicon-based, and hybrid similarities between two given sentences (Ma, 2006; Gale and Church, 1993; Wu, 1994). We refer to this kind of information that can be captured from the source and the target sentences as "**local**" information. Although local information can show the similarity between two sentences, it cannot consider the information that comes from the other sentences in the documents. We refer to the information related to the other sentences of the documents as "**global**" information. In this paper, we propose a language-independent method for extracting parallel sentences[1] from the aligned documents by exploiting both local and global information.

We explore a learning approach to compute the similarity between a given sentence pair to consider local information. To this aim, we consider a number of length-based, lexicon-based, alignment-based, and miscellaneous features and train a Maximum Entropy (*MaxEnt*) binary classifier. Since MaxEnt is a probabilistic classifier, it enables us to compute the probability of being parallel for any sentence pair.

A number of previous work (Gale and Church, 1993; Ma, 2006; Baratalipour, 2012) assume that the sentences in the target document are translated from the source sentences exactly in the same order. Therefore, they avoid cross alignments during their aligning process. Although this assumption might be true in many situations, there are also several aligned documents, such as comparable corpora, in which documents are not translation of each other. These aligned documents also may contain several parallel sentences. Hence, to avoid this strict assumption, we propose a method to penalize cross alignments, instead of completely ignoring them. In other words, we develop a method that avoids cross alignments unless the aligned sentences are highly similar. To this aim, we design a bipartite weighted graph for each aligned source and target documents. Each vertex in the graph corresponds to a sentence[2] and the edge weights between the vertices are computed using the aforementioned MaxEnt classifier. Then, we introduce an objective function to maximize the similarity of aligned sentences and also to penalizes the cross alignments. To optimize the objective function, we propose a novel method based on integer linear programming (ILP).

In the experiments, we focus on the English–Persian language pair, since the available parallel corpora in the Persian language (also known as Farsi) are limited in terms of both domain and size. We extract parallel sentences from the Wikipedia[3] articles, since they cover wide variety of domains. We extensively evaluate our method in different scenarios. We first intrinsically evaluate the proposed sentence alignment methods using manually tagged test data extracted from English–Persian Wikipedia articles. The intrinsic evaluation shows that the proposed method significantly outperforms competitive baselines. We further create a parallel corpus using the Wikipedia articles, which is freely available for research purposes.[4] We extrinsically evaluate the proposed method using CLIR and SMT applications. The experimental results show that the quality of the extracted English–Persian parallel corpus is higher than the existing parallel corpora. We also consider the English–German language pair to evaluate the proposed method in an additional language pair. The results demonstrate that the proposed method can perform effectively in other language pairs, such as English–German.

The remainder of this paper is structured as follows: Section 2 reviews related work. We present the characteristics of Wikipedia articles in Section 3. Section 4 introduces the proposed method to exploit both local and global information to extract parallel sentences. The proposed method is evaluated in Section 5. We finally conclude our paper and discuss possible future directions in Section 6.

2. Related work

In this section, we first present related sentence alignment methods. We then introduce the existing English–Persian bilingual resources. We finally review the previous work related to integer linear programming.

[1] Note that although the sentences aligned by the automatic aligners are not always exact translations of each other, they can be considered as parallel sentences for further learning processes (Gupta et al., 2013). That is why several methods (Resnik and Smith, 2003; Patry and Langlais, 2011; Baratalipour, 2012) have been so far proposed to extract parallel sentences from the Web.

[2] We also define super-nodes to be able to extract parallel sentences with more than one sentence in at least one of the source and target languages.

[3] http://wikipedia.org/.

[4] http://ece.ut.ac.ir/en/project/wikipedia-parallel-corpus.