Information Processing and Management 52 (2016) 299–318

Contents lists available at ScienceDirect

Information Processing and Management

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

Extracting translations from comparable corpora for Cross-Language Information Retrieval using the language modeling framework

Razieh Rahimi[a], Azadeh Shakery[a,b,*], Irwin King[c]

[a] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, P.O. Box 14395-515 Tehran, Iran
[b] School of Computer Science, Institute for Research in Fundamental Sciences (IPM), P.O. Box 19395-5746 Tehran, Iran
[c] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong

ARTICLE INFO

Article history:
Received 13 November 2014
Revised 10 August 2015
Accepted 17 August 2015
Available online 10 November 2015

Keywords:
Translation model
Bilingual lexicon
Comparable corpora
Cross-Language Information Retrieval
Language modeling framework

ABSTRACT

A main challenge in Cross-Language Information Retrieval (CLIR) is to estimate a proper translation model from available translation resources, since translation quality directly affects the retrieval performance. Among different translation resources, we focus on obtaining translation models from comparable corpora, because they provide appropriate translations for both languages and domains with limited linguistic resources. In this paper, we employ a two-step approach to build an effective translation model from comparable corpora, without requiring any additional linguistic resources, for the CLIR task. In the first step, translations are extracted by deriving correlations between source–target word pairs. These correlations are used to estimate word translation probabilities in the second step. We propose a language modeling approach for the first step, where modeling based on probability distribution provides two key advantages. First, our approach can be tuned easier in comparison with heuristically adjusted previous work. Second, it provides a principled basis for integrating additional lexical and translational relations to improve the accuracy of translations from comparable corpora. As an indication, we integrate monolingual relations of word co-occurrences into the process of translation extraction, which helps to extract more reliable translations for low-frequency words in a comparable corpus. Experimental results on an English–Persian comparable corpus show that our method outperforms the previous approaches in terms of both translation quality and the performance of CLIR. Indeed, the proposed method is naturally applicable to any comparable corpus, regardless of its languages. In addition, we demonstrate the significant impact of word translation probabilities, estimated in the second step of our approach, on the performance of CLIR.

1. Introduction

Cross-Language Information Retrieval (CLIR) refers to the retrieval process where documents and queries are in different languages. Some sort of processing is thus needed to match query and document representations. The most general and promising

[*] Corresponding author at: School of Electrical and Computer Engineering, College of Engineering, University of Tehran, P.O. Box 14395-515 Tehran, Iran. Tel.: +982182089722.

E-mail addresses: r.rahimi@ece.ut.ac.ir (R. Rahimi), shakery@ut.ac.ir (A. Shakery), king@cse.cuhk.edu.hk (I. King).

http://dx.doi.org/10.1016/j.ipm.2015.08.001

approach for the CLIR task is to use translation resources. Bilingual corpora are widely used in CLIR to obtain statistical translation knowledge which addresses some translation issues such as out of vocabulary, neologism, and ambiguous words. Bilingual corpora are of two types: *parallel* and *comparable*. Aligned documents in a parallel corpus are literal translations of each other, while the ones in a comparable corpus cover the same or similar topics. Parallel corpora are generally better translation resources than comparable corpora because of more precise alignments. However, they are not widely available due to the higher cost of creation.

Comparable corpora are important translation resources for both languages and domains with limited linguistic resources. For language pairs with limited translation resources, on one hand, comparable corpora can be built at lower costs from more abundant sources of monolingual texts, such as articles from news agencies, compared to parallel corpora. Comparable corpora for such language pairs help to obtain initial translation knowledge which, in turn, facilitates building more exact and complete translation resources. For a domain with limited translation resources, on the other hand, an in-domain comparable corpus may provide more accurate translations compared to a general parallel corpus (Talvensaari et al., 2008). This is because, the domain of the constituting documents of bilingual corpora impacts the sense of extracted translations for words. Lower cost of building comparable corpora facilitates building several domain-specific corpora. Therefore, comparable corpora are also valuable resources for resource poor domains.

For the purpose of this paper, we focus on extracting translation knowledge from comparable corpora without employing additional linguistic resources. In this paper, we derive high-quality translation models from comparable corpora for CLIR through the following contributions:

1. We propose a language modeling approach to extract correlations between each pair of bilingual words from comparable corpora. The intuition behind the proposed method is that words that are translations of each other have similar contributions in generating language models of aligned documents. Our method improves the accuracy of extracted translations and their related words over the previous approaches. Indeed, the proposed approach can be optimized straightforwardly.
2. We show that integrating monolingual relations of word co-occurrences into word models helps to improve the accuracy of translations by providing more exact estimates of word statistics. In addition, our method provides a principled basis for integrating other sources of lexical and translational relations in order to improve the accuracy of extracted translations from comparable corpora.
3. We compare different estimations of translation probabilities from word correlations and show the significant impact of this estimation on the performance of CLIR. This reveals a new quality criterion for translation models to be suitable for Cross-Language Information Retrieval.

We evaluated our proposed approach on an English–Persian comparable corpus. Assessment of extracted translations using a machine-readable bilingual dictionary demonstrates that our approach obtains meaningful correlations between words. We further adopted the obtained translation models for English–Persian CLIR. Using translations extracted from only the comparable corpus, we achieve performance of CLIR between 36% and 58% of that of monolingual IR over three standard CLEF datasets.

A preliminary version of our work was presented in Rahimi and Shakery (2013). In the current paper, in addition to a detailed description of the proposed method, we extend our previous work through (1) defining a similarity function to derive bidirectional correlations between source and target word pairs, (2) proposing an approach to better estimate the models of low-frequency words in a comparable corpus, (3) providing a detailed analysis of the effectiveness of our method to demonstrate words for which the proposed method can produce reliable translations, and (4) comparing different functions to estimate word translation probabilities based on word correlations and show how much it impacts the performance of CLIR by empirical evaluation.

The remainder of this paper is organized as follows. In Section 2, we review previous work. Section 3 gives a description of the translation extraction problem. Then, we present our proposed language modeling method for translation extraction in Section 4. Following, experimental design and the results are reported in Sections 5 and 6, respectively. Finally, the paper is concluded in Section 7.

2. Related work

The goal of CLIR is to score documents with respect to a query in another language than that of the documents. Due to the different languages of queries and documents, some sort of processing is needed to match document terms with query terms. Cross-lingual retrieval between similar languages (such as Italian–French and Chinese–Japanese (Savoy, 2005)) can be performed without any translation (Buckley et al., 2000; He et al., 2003; Mcnamee & Mayfield, 2004). However, the most general approach for this task is to use translation resources.

Among translation resources, we focus on comparable corpora and discuss the existing methods for obtaining translation knowledge from these corpora. A majority of approaches in this domain extract translations based on the assumption that there is a correlation between collocates of a word and those of its translation (Fung, 1998; Rapp, 1999). Based on this assumption, a context vector is built for each word representing all its collocates in a window-based context. Source-language word vector is transferred into a target-language vector, relying on an existing dictionary. Target words are then ranked based on the similarity of their vectors to the vector of the source word. Chiao and Zweigenbaum (2002) adapt this strategy for extracting translations from a comparable corpus on a specific medical domain. Sadat et al. (2003) build a translation model by merging translation