Information Processing and Management 50 (2014) 384–398

Contents lists available at ScienceDirect

Information Processing and Management

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

Mining a Persian–English comparable corpus for cross-language information retrieval ☆,☆☆

Homa B. Hashemi [a], Azadeh Shakery [a,b,*]

[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

ARTICLE INFO

Article history:
Received 16 June 2011
Received in revised form 17 October 2013
Accepted 17 October 2013
Available online 11 November 2013

Keywords:
Comparable corpora
Cross-language information retrieval
Term association network
Translation validity check

ABSTRACT

Knowledge acquisition and bilingual terminology extraction from multilingual corpora are challenging tasks for cross-language information retrieval. In this study, we propose a novel method for mining high quality translation knowledge from our constructed Persian–English comparable corpus, University of Tehran Persian–English Comparable Corpus (UTPECC). We extract translation knowledge based on Term Association Network (TAN) constructed from term co-occurrences in same language as well as term associations in different languages. We further propose a post-processing step to do term translation validity check by detecting the mistranslated terms as outliers. Evaluation results on two different data sets show that translating queries using UTPECC and using the proposed methods significantly outperform simple dictionary-based methods. Moreover, the experimental results show that our methods are especially effective in translating Out-Of-Vocabulary terms and also expanding query words based on their associated terms.

1. Introduction

The researches on Cross Language Information Retrieval (CLIR) have recently received much attention, due to the fast growth of the World Wide Web and the availability of information in different languages on the Web. One of the main issues in CLIR is where to obtain the translation knowledge (Oard & Diekema, 1998). Multilingual corpora are widely used for this purpose which are actually available in many language pairs and extracting translation knowledge from multilingual corpora has been extensively studied using various statistical methods. They can be either in the form of parallel or comparable corpora. However, there are limitations in obtaining parallel corpora in all domains and languages while comparable corpora are much easier resources to obtain. Thus, recently, there has been considerable interest in using comparable corpora as translation resources (e.g. Fung & Yee, 1998; Sadat, 2010b; Talvensaari, Laurikkala, Järvelin, Juhola, & Keskustalo, 2007; Tao & Zhai, 2005). In this paper we use a Persian–English comparable corpus, University of Tehran Persian–English Comparable Corpus (UTPECC) (Hashemi, Shakery, & Faili, 2010), to do English–Persian cross language information retrieval by extracting term associations from the comparable corpus. These association terms my contain both translations of a term, and terms that are actually related to its (correct) translations. In this paper, all selected translations and related terms for a term are referred to as its "translations". As a basic method, we obtain term associations based on co-occurrence of terms in the alignments. We further propose a novel way of extracting translations in different languages based on Terms Association

☆ This research was in part supported by a grant from IPM. (No. CS1389-4-05).
☆☆ The first author is currently attending University of Pittsburgh. This work was done while she was at University of Tehran.
* Corresponding author.
E-mail addresses: H.B.Hashemi@ece.ut.ac.ir (H.B. Hashemi), Shakery@ut.ac.ir (A. Shakery).

0306-4573/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.10.002

Network (TAN) which exploits term associations in monolingual data as well as bilingual term associations to better detect translation knowledge. TAN method uses a network of terms with implicit mutual information links between terms in the same language and term association links between terms in different languages. The main contribution of this paper lies in combining these term association links as a network, which in turn improves CLIR effectiveness. Basically, we use the neighborhoods of terms in this network to re-score the translation alternatives. Two terms are translations of each other if their neighborhoods in the same language are strongly connected and vice versa they are not likely to be translations if their neighborhoods are not strongly connected. Also, in order to discard misleading translation candidates, we do translation validity check using cross-outlier detection method. Intuitively, if the distribution of the weights of a term's translations is different from that of its neighbors' translations, the term is considered as an outlier.

We evaluated our methods on Hamshahri and INFILE data sets by doing cross-language information retrieval using the cross-lingual term associations extracted from the comparable corpus. We use the extracted translation knowledge to construct a query language model in the target language corresponding to each query in the source language and rank the documents based on the KL-divergence between the query language model and document language models. Experiments show promising results for extracting translation knowledge from the UTPECC by (1) translating Out Of Vocabulary (OOV) terms, such as proper nouns, which are not in our dictionaries, (2) expanding query words with their related terms and also (3) using probability scores of extracted translations. Also, using comparable corpora helps to complement dictionaries by translating OOV terms and finding related terms to expand query words.

The rest of the paper is organized as follows. We first present some previous work done on exploiting comparable corpora in Section 2. We then introduce our translation extraction and query translation methods in Sections 3 and 4. We present the experiment results in Section 5 and finally bring the conclusions and future work of our study in Section 6.

2. Previous work

Using comparable corpora as a language resource has been studied extensively in the existing literature, in fields such as cross-language information retrieval (Picchi & Peters, 1996; Sadat, 2010a, 2010b; Sadat, Yoshikawa, & Uemura, 2003; Talvensaari et al., 2007), cross-lingual document association (Tao & Zhai, 2005; Vu, Aw, & Zhang, 2009), in extracting parallel sentences (Abdul-Rauf & Schwenk, 2009; Munteanu & Marcu, 2005) and extracting word translations (Fung, 1995; Fung & Yee, 1998; Hassan, Fahmy, & Hassan, 2007; Laroche & Langlais, 2010; Morin, Daille, Takeuchi, & Kageura, 2007; Morin, Daille, Takeuchi, & Kageura, 2010; Otero & Campos, 2010; Rapp, 1995; Rapp & Zock, 2010; Tanaka & Iwasaki, 1996; Tao & Zhai, 2005; Udupa, Saravanan, Kumaran, & Jagarlamudi, 2008; Yu & Tsujii, 2009). Most of the early work in extracting word translations employ an initial lexicon of seed words (e.g. Franz, McCarley, & Roukos, 1999; Fung & Yee, 1998; Picchi & Peters, 1996; Rapp & Zock, 2010; Sadat et al., 2003) and some of them are based on linguistic knowledge such as language morphologies (e.g. Sadat, 2010b; Yu & Tsujii, 2009). We follow the research on extracting word translations without requiring any additional linguistic resources (e.g. Talvensaari et al., 2007; Tao & Zhai, 2005) and use the extracted knowledge to translate queries for cross-language information retrieval. In order to do CLIR, we construct query language models based on scores of extracted related terms from the comparable corpus. Trieschnigg, Hiemstra, de Jong, and Kraaij (2010) use a similar approach to train their translation model in CLIR using a corpus.

The problem of exploring a comparable corpus and its combination with a dictionary to do CLIR has been studied before in Talvensaari et al. (2007) and Sadat (2010b). However, our methods in both extracting translation knowledge and constructing query translations are different from theirs. Talvensaari et al. (2007) have built a comparable corpus query translation program (Cocot) which we use as our baseline. They also combine the comparable corpus results with dictionary-based query translation (UTACLIR) and construct queries in the InQuery format, while we use a simple dictionary and construct query language models. Sadat (2010b) presents a two-stage corpus-based translation model which aims to find translations of a source word in the target language corpus and also translations of the target words in the source language corpus. The two stages contain bi-directional extraction of bilingual terminology from comparable corpora and selection of best translation alternatives based on a morphological analyzer. She has also exploited the linear combination of comparable corpus results with bilingual dictionaries. In both works, the impact of comparable corpora on cross-language information retrieval especially combining the extracted translations with dictionaries has been shown to be effectively positive.

3. Mining term associations

In this section, we propose a process for learning cross-lingual term associations from comparable corpora. As a first step, we extract translation knowledge using term co-occurrences in the comparable corpus alignments. In the second step, we propose a method based on term association network which exploits term associations in monolingual data as well as bilingual term associations to better extract translation knowledge. We further propose to use cross-outlier detection to filter out misleading translation candidates which are detected as outliers. We will present each step in more detail in the rest of this section.

3.1. Cocot: basic translation extraction method

As the basic method, in order to extract term associations from the comparable corpus, we use the method used in Cocot, the comparable corpus query translation program which is proposed in Talvensaari, Pirkola, Järvelin, Juhola, and