Information Processing and Management 53 (2017) 371–387

Contents lists available at ScienceDirect

Information Processing and Management

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

An expectation-maximization algorithm for query translation based on pseudo-relevant documents

Javid Dadashkarimi[a], Azadeh Shakery[a,b,*], Heshaam Faili[a,b], Hamed Zamani[c,1]

[a] School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
[b] Institute for Research in Fundamental Sciences (IPM), Tehran, Iran
[c] Center for Intelligent Information Retrieval, University of Massachusetts Amherst, MA 01003, USA

ARTICLE INFO

Article history:
Received 10 January 2016
Revised 23 November 2016
Accepted 25 November 2016

MSC:
00-01
99-00

Keywords:
Dictionary-based cross-language information retrieval
Query translation
Expectation maximization
Pseudo-relevant documents

ABSTRACT

Query translation in cross-language information retrieval (CLIR) can be done by employing dictionaries, aligned corpora, or machine translators. Scarcity of aligned corpora for various domains in many language pairs intensifies the importance of dictionary-based CLIR which motivates us to use only a bilingual dictionary and two independent collections in source and target languages for query translation. We exploit pseudo-relevant documents for a given query in the source language and pseudo-relevant documents for a translation of the query in the target language with a proposed expectation-maximization algorithm for improving query translation. The proposed method (called *EM4QT*) assumes that each target term either is translated from the source pseudo-relevant documents or has come from a noisy collection. Since EM4QT does not directly consider term coherency, which is defined as fluency of the target translation, we investigate a crucial question: can EM4QT be improved using either coherency-based methods or token-to-token translation ones? To address this question, we combine different translation models via simple linear interpolation and a proposed divergence minimization method. Evaluations over four CLEF collections in Persian, French, Spanish, and German indicate that EM4QT significantly outperforms competitive baselines in all the collections. Our experiments also reveal that since EM4QT indirectly considers term coherency, combining the method with coherency-based models cannot significantly improve the retrieval performance. On the other hand, investigating the query-by-query results supports the view that EM4QT usually gives a relatively high weight to one translation and its combination with the proposed token-to-token translation model, which is obtained by running EM4QT for each query term separately, soothes the effect and reaches better results for many queries. Comparing the method with a competitive word-embedding baseline reveals the superiority of the proposed model.

1. Introduction

Exponential growth of the Internet and availability of documents in different languages have turned the World Wide Web into a huge multilingual environment. Retrieval systems are obliged to retrieve documents in a language other than

* Corresponding author.
E-mail addresses: dadashkarimi@ut.ac.ir (J. Dadashkarimi), shakery@ut.ac.ir (A. Shakery), hfaili@ut.ac.ir (H. Faili), zamani@cs.umass.edu (H. Zamani).

[1] A part of this work was done while Hamed Zamani was a student in University of Tehran.

http://dx.doi.org/10.1016/j.ipm.2016.11.007
0306-4573/

the users' native language since the users intend to find all relevant information available independent of the language. In these circumstances, it is easier for the users to formulate the queries in their native language (Nie, 2010). Cross-language information retrieval (CLIR) tends to bridge the gap between the languages. To this end, several shared-tasks have been also focused on CLIR and related tasks, including the TREC and CLEF shared-tasks. The following techniques are proposed for CLIR: (1) translating queries to the target language, (2) translating documents to the source language, (3) translating queries and documents to a third language (Nie, 2010), (4) mapping queries and documents to a shared low-dimensional representation space, and (5) using cross-lingual semantic/concept networks (Franco-Salvador, Rosso, & Montes-y Gómez, 2016; Franco-Salvador, Rosso, & Navigli, 2014; Gouws, Bengio, & Corrado, 2014; Gupta, Bali, Banchs, Choudhury, & Rosso, 2014; Platt, Toutanova, & Yih, 2010; Vulic & Moens, 2015; Yih, Toutanova, Platt, & Meek, 2011). Although it is shown that translating documents can outperform the query translation approach in some languages, document translation is a time-consuming approach and demands re-indexing of the entire collection for each language (Chen & Gey, 2003). That is why query translation is the most common technique for CLIR.

Queries can be translated using machine translation systems or various translation resources, such as dictionaries, comparable corpora, and parallel corpora. It is well known that building parallel corpora is highly expensive in terms of both time and cost. Moreover, current translation extraction methods are not able to purify noisy translation candidates from comparable corpora completely and this is why many language pairs are suffering from lack of these linguistic resources. In addition, these resources are usually domain-specific and employing them in domains other than the domain of the corpus can lead to low performance (Nie, 2010). Furthermore, extracting reliable translation knowledge from comparable corpora heavily depends on the size of the collection in terms of the number of alignments (Talvensaari, Laurikkala, Järvelin, Juhola, & Keskustalo, 2007). On the other hand, bilingual machine readable dictionaries are known as available resources with high translation coverage in many language pairs for general domains (Dadashkarimi, Shakery, & Faili, 2014; Gearailt, 2003; Yih et al., 2011). All these facts intensify the importance of studying dictionary-based CLIR.

Dictionaries provide an unweighted list of target terms for each term of the source language. There is an important challenge in the dictionary-based CLIR: ambiguity in translation and swamping effect as a result[1] (Gearailt, 2003). Indeed, in most cases each term in the source language has more than one translation candidate in the target language and thus detecting the correct translation for each term could be a big issue here. Several methods have so far been proposed to address these problems, such as structured query (Pirkola, 1998; Pirkola, Hedlund, Keskustalo, & Järvelin, 2001), iterative translation disambiguation (ITD) (Monz & Dorr, 2005), and maximum coherence model (Liu, Jin, & Chai, 2005). In structured queries all the translations of a word are dealt as members of a synonym set and the number of occurrences of the source word equals the sum of the number of the occurrences of the members. But the probabilistic approaches score documents based on translation probabilities (Dadashkarimi et al., 2014; Ganguly, Leveling, & Jones, 2012; Liu et al., 2005; Monz & Dorr, 2005). Many of these methods aim at disambiguating the query using global mutual information of the translations in the target language. In this paper, our contribution is to use a couple of local in-the-context collections, one from the source language and the other from the target language, to compute a query-dependent translation model for each query. Pseudo-relevant documents in response to the query in both source and target languages comprise these collections.

Pseudo-relevant documents are a number of top-ranked documents in response to the query of the user and are expected to be relevant to the query and thus can potentially be suitable resources for extracting translation knowledge. The proposed method runs an on-line disambiguation process by incorporating two collections, one in the source language, and another in the target language. The method retrieves the source documents for the initial query and then, translates the query using a simple translation technique (e.g., uniform weighting of translations); at the next step target documents are retrieved for the translated query. We expect that distribution of the context terms in the source collection to be similar to the distribution of their translations in the target collection, accepting a small amount of noise from the background collection. In more details, it is expected that each word in the target pseudo-relevant collection either is translated from the source pseudo-relevant collection or has come from a noisy background collection. Based on this expectation, we propose an expectation-maximization (EM) algorithm, an iterative hill-climbing algorithm, to extract query-dependent translation knowledge for each query. This method is called expectation-maximization for query translation (EM4QT). We prove that the proposed method converges to a global optimum solution (see Appendix B).

Although the methods based on term coherency perform promising in dictionary-based CLIR, EM4QT does not directly consider the coherency between target terms. Therefore, in this paper we also investigate a crucial research question: can the performance of EM4QT be improved using the off-line coherency-based CLIR methods? To answer this question, we consider the simple linear interpolation method and also propose a statistical divergence minimization method to combine more than one translation model.

Since the extracted translation model from the proposed EM4QT method usually drifts to the translations which are more coherent in the pseudo-relevant documents and more discriminative through the collection, we investigate another research question: can the performance of EM4QT be improved using a token-to-token translation model? To this aim, we employ a token-to-token translation technique in which each term of the query is posed to the EM-based system individually and the obtained model is combined with that of EM4QT.

[1] retrieving irrelevant documents which is caused by translating the query to non-relevant candidates.