Information Processing and Management 54 (2018) 291–302

Contents lists available at ScienceDirect

Information Processing and Management

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

The Dilution/Concentration conditions for cross-language information retrieval models

Bo Li[*,a], Eric Gaussier[b], Dan Yang[c]

[a] School of computer science, Central China Normal University, Wuhan, China
[b] CNRS-LIG/AMA, Université Grenoble Alpes, Grenoble, France
[c] China Electric Power Research Institute, Wuhan, China

ARTICLE INFO

Keywords:
Cross-language information retrieval
D/C condition
Information retrieval heuristic

ABSTRACT

Experimental results of cross-language information retrieval (CLIR) do not indicate why a model fails or how a model could be improved. One basic research question is thus whether it is possible to provide conditions by which one can evaluate any existing or new CLIR strategy *analytically* and one can improve the design of CLIR models. Inspired by the heuristics in monolingual IR, we introduce in this paper Dilution/Concentration (D/C) conditions to characterize good CLIR models based on direct intuitions under artificial settings. The conditions, derived from first principles in CLIR, generalize the idea of query structuring approach. Empirical results with state-of-the-art CLIR models show that when a condition is not satisfied, it often indicates non-optimality of the method. In general, we find that the empirical performance of a retrieval formula is tightly related to how well it satisfies the conditions. Lastly, we propose, by following the D/C conditions, several novel CLIR models based on the information-based models, which again shows that the D/C conditions are efficient to feature good CLIR models.

1. Introduction

Cross-Language Information Retrieval (CLIR) is concerned with the problem of finding documents written in a language different from that of the query. If attempts to model multilinguality in information retrieval date back from the early seventies (Salton, 1969), a renewed interest was brought to the field by the rise of the web in the mid-nineties, as pages written in many different languages were available suddenly. International organizations, governments of multi-lingual countries, to name the most important, have been traditional users of CLIR systems, but the need for such systems in everyday life, even though less ascertained, becomes more and more clear, with the development of travels, tourism and multilinguality, at all levels (the book by J.-Y. Nie on cross-language information retrieval (Nie, 2010) exposes in detail the need for cross-language and multilingual IR).

There are several ways to cross the language barrier in CLIR models: through mapping the document representation into the query representation space (an approach known as document translation), through mapping the query representation into the document representation (an approach known as query translation) or through mapping both representations into a third space (interlingua approach). As for implementation, existing CLIR models fall into two categories: model-independent approaches and model-dependent approaches. Model-independent approaches treat translation and retrieval as two separate processes. The queries or the documents are first translated into the corresponding language of the documents or the queries. Monolingual IR models are then applied directly. A typical and also broadly used approach of this kind is the machine translation (MT) approach (e.g. Braschler,

[*] Corresponding author.
E-mail addresses: libo@mail.ccnu.edu.cn (B. Li), eric.gaussier@imag.fr (E. Gaussier), yangdan3@epri.sgcc.com.cn (D. Yang).

https://doi.org/10.1016/j.ipm.2017.11.008
Received 5 March 2017; Received in revised form 17 November 2017; Accepted 26 November 2017

2004; Kraaij, Nie, & Simard, 2003) which employs MT systems to translate the queries or documents before the monolingual retrieval process. Model-dependent methods integrate the translation and retrieval processes in a uniform framework. These methods are mostly developed based on language modeling strategies (Federico & Bertoldi, 2002; Kraaij et al., 2003) and have the advantage of accounting better for the uncertainty of translation during retrieval.

Recent efforts to enhance CLIR performance mostly try to use better language resources or more accurate translation models. Wikipedia contains rich multilingual knowledge and it has been investigated in CLIR by several researchers (Kim, Ko, & Oard, 2015; Sharma & Mittal, 2016; Ye, Huang, He, & Lin, 2012). In terms of low-resource languages, parallel resources are rare, comparable corpora have thus been used instead (Hashemi & Shakery, 2014). Translation is a crucial problem in CLIR and researchers have developed better translation models, mostly inspired by statistical machine translation.The work Seo, Kim, Rim, and Myaeng (2005) deals with translation ambiguities in query translation making using of context information in the translated queries. Both Wang and Oard (2012) and Ture and Lin (2014) aim to find more accurate estimation for the translation probabilities. Some studies pay attention to specific problems in CLIR, for instance the misspelling query problem (Vilares, Alonso, Doval, & Vilares, 2016). Word embeddings approaches have attracted a lot of attention in natural language processing, which have been set as standards in several semantic evaluation tasks (Camacho-Collados, Pilehvar, Collier, & Navigli, 2017). In this direction, such studies as Bhattacharya, Goyal, and Sarkar (2016); Gupta, Banchs, and Rosso (2017); Vulić and Moens (2015) have tried to use cross-lingual word embeddings to aid the translation process in CLIR and the results obtained seem to be promising.

The language barrier makes it difficult to establish direct intuitions such as TF-IDF in the monolingual case. Moreover, experimental results of CLIR do not indicate why a model fails or how a model could be improved. One basic research question is thus whether it is possible to provide conditions by which one can evaluate any existing or new CLIR strategy *analytically* and one can improve the design of CLIR models. There have been only a few pioneering studies in this direction (for instance, the one in Kishida, 2008), which, unfortunately, still relies on complicated experimental evaluation. In this paper, the intuition for building the conditions is partially inspired by the query structuring strategy. Among all existing CLIR methods, query structuring is model-independent and has been investigated broadly in CLIR (Ballesteros & Sanderson, 2003; Pirkola, 1998; Pirkola, Hedlund, Keskustalo, & Järvelin, 2001; Sperer & Oard, 2000). Referring to the simple nature behind query structuring technique, we propose in this paper a family of Dilution/Concentration (D/C) conditions featuring a good CLIR model. In the following sections, we will show the efficiency of the conditions through a series of experiments with both classic CLIR models and a new one based on the information-based models (Clinchant & Gaussier, 2010).

The remainder of the paper is organized as follows. In Section 2, we first introduce a family of conditions to assess the validity of CLIR models, prior to reviewing several strategies used in standard CLIR systems. We then validate the theoretical findings in CLIR experiments. In Section 3, directed by D/C conditions, several CLIR models are proposed based on the information-based models. The reliability of the D/C conditions is again validated by the experiments with information-based models. We discuss some related works and conclude the paper in Section 4.

2. A family of Dilution / Concentration conditions

In this section, we will abstract a family of Dilution/Concentration (D/C) conditions by considering an “ideal” setting and the behavior good CLIR strategies should have in this ideal setting. This setting provides a theoretical tool which allows one to assess the validity of CLIR strategies. The conditions are developed inspired by but not to fit the query structuring strategy itself. To be exact, the D/C conditions tell a more general principle than the query structuring technique which can be treated as one possible implementation of the D/C conditions. The proposed conditions are then validated in the experiments by considering various CLIR models against the conditions.

The notations we use throughout the paper are summarized in Table 1 (*w* represents a term).

Table 1
Notations used in the paper.

Notation	Description
x_w^q	Number of occurrences of *w* in query *q*
x_w^d	Number of occurrences of *w* in document *d*
t_w^d	Normalized version of x_w^d
l_q	Length of query *q*
l_d	Length of document *d*
l_m	Average document length
L	Length of document collection
N	Number of documents in the collection
N_w	Number of documents containing *w*
$TS(w)$	Set of translations of *w*
$DS(w)$	Set of documents containing *w*
RSV(q, d)	Retrieval status value of doc. *d* for query *q*