Information Processing and Management 52 (2016) 646–657

Contents lists available at ScienceDirect

Information Processing and Management

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

Studying the effect and treatment of misspelled queries in Cross-Language Information Retrieval

Jesús Vilares[a,*], Miguel A. Alonso[a], Yerai Doval[a,b], Manuel Vilares[b]

[a] Grupo LYS, Departamento de Computación, Facultade de Informática, Universidade da Coruña, Campus de Elviña, 15071 – A Coruña, Spain

[b] Grupo COLE, Departamento de Informática, E.S. de Enxeñaría Informática, Universidade de Vigo, Campus As Lagoas, 32004 – Ourense, Spain

ARTICLE INFO

Article history:
Received 14 October 2015
Revised 16 December 2015
Accepted 17 December 2015
Available online 12 January 2016

Keywords:
Misspelled queries
Cross-Language Information Retrieval
Machine translation
Spelling correction
Character *n*-grams

ABSTRACT

In contrast with their monolingual counterparts, little attention has been paid to the effects that misspelled queries have on the performance of Cross-Language Information Retrieval (CLIR) systems. The present work makes a first attempt to fill this gap by extending our previous work on monolingual retrieval in order to study the impact that the progressive addition of misspellings to input queries has, this time, on the output of CLIR systems. Two approaches for dealing with this problem are analyzed in this paper. Firstly, the use of automatic spelling correction techniques for which, in turn, we consider two algorithms: the first one for the correction of isolated words and the second one for a correction based on the linguistic context of the misspelled word. The second approach to be studied is the use of character *n*-grams both as index terms and translation units, seeking to take advantage of their inherent robustness and language-independence. All these approaches have been tested on a from-Spanish-to-English CLIR system, that is, Spanish queries on English documents. Real, user-generated spelling errors have been used under a methodology that allows us to study the effectiveness of the different approaches to be tested and their behavior when confronted with different error rates. The results obtained show the great sensitiveness of classic word-based approaches to misspelled queries, although spelling correction techniques can mitigate such negative effects. On the other hand, the use of character *n*-grams provides great robustness against misspellings.

1. Introduction

When facing the current context of globalization of the use of Internet and the Web, classic Information Retrieval (IR) systems (Manning, Raghavan, & Schütze, 2008) have to deal with the fact that in many cases the information available is written in a language different from that of its potential customers, the language they most probably use when submitting a query. In response to this issue, the field of Cross-Language Information Retrieval (CLIR) has arisen within the IR community.

In brief, CLIR is a particular case of IR where queries and documents are written in different languages (Grefenstette, 1998; Kim, Ko, & Oard, 2015; Nie, 2010; Peters, Braschler, & Clough, 2012): given a query in one language (called *source language*), the system provides the user with the means and skills needed to find relevant documents written in another

* Corresponding author. Tel.: +34981167000.
E-mail addresses: jesus.vilares@udc.es (J. Vilares), miguel.alonso@udc.es (M.A. Alonso), yerai.doval@udc.es, yerai.doval@uvigo.es (Y. Doval), vilares@uvigo.es (M. Vilares).

http://dx.doi.org/10.1016/j.ipm.2015.12.010
0306-4573/

language (called *target language*). Most CLIR systems apply some kind of intermediate *translation stage* in order to convert the cross-language configuration to a classic monolingual configuration (i.e., with queries and documents written in the same language) that can be managed by classic IR systems. Within this framework, we refer to query-translation based CLIR, document-translation based CLIR and interlingua-based CLIR when queries, documents or both queries and documents are translated, respectively (Wu, He, Ji, & Grishman, 2008). Due to the fact that the translation of large document collections has serious practical limitations, works in this field have mostly focused on query translation (Nie, 2010).

In parallel, and also as a result of this phenomenon of globalization of access to information, it becomes increasingly necessary to have systems capable of operating on texts with misspelling errors, particularly in the case of queries (Guo, Xu, Li, & Cheng, 2008). In fact, nowadays it is common to assume that some text-cleaning processing stage is needed in order to extract useful information from user-generated content, such as product reviews (Vilares, Alonso, & Gómez-Rodríguez, 2015a) or microblog entries (Vilares, Alonso, & Gómez-Rodríguez, 2015b; 2015c). In this work, we consider as *misspelling errors* those corresponding to typographical errors during writing, those due to the ignorance of the actual spelling of a word and those arising from the presence of noise in the generation process, e.g. OCR or speech recognition (Kukich, 1992). Since formal IR models were originally designed to work on texts without errors, their presence can substantially reduce system performance. To deal with this issue, another field has arisen within the IR community: *Tolerant Information Retrieval* (TIR) (Manning et al., 2008, chap. 3).

This article deals with the analysis of the impact of misspelled queries on CLIR systems and the design of Tolerant CLIR systems that are able to operate with such queries. Our practical experience suggests that the inability to deal with misspelled words is a major source of translation errors for the machine translation engines used in CLIR systems for query translation. In order to do this, we will take advantage of our previous experience both in the study of the impact of misspelled queries on monolingual IR (Vilares, Vilares, & Otero, 2011) and in character n-gram based CLIR (Vilares, Vilares, Alonso, & Oakes, 2016). Thus, we will make a comparative analysis of the effectiveness of two possible strategies which are reflected in three specific techniques with different levels of integration of knowledge and linguistic resources. These strategies are in line with the two generic state-of-the-art approaches to the problem of TIR: firstly, the use of words as working units and, secondly, the use of sub-words as working units. The proposed solutions have been subjected to different experiments in a from-Spanish-to-English retrieval context (i.e., queries submitted in Spanish over a collection of English texts). The methodology applied for this purpose allows us to test real human errors rather than artificially-generated ones, giving us a wide range of options.

To the best of our knowledge there are no similar works with this level of detail in a cross-language context. The work of Darwish and Magdy (2007), for example, although distantly-related to ours, differs significantly since it is focused on monolingual retrieval of scanned documents containing OCR errors, instead of multilingual retrieval with misspelling errors present in the queries, as is our case.

The structure of the rest of this paper is as follows. Section 2 describes our proposals for the treatment of queries with errors. Next, Section 3 discusses in detail our proposal based on the use of words as working units in conjunction with the use of spelling correction techniques, while Section 4 presents our proposal based on the use of character *n*-grams as working units both in the translation and retrieval stages. In Section 5, the methodology employed for testing is explained, and the experimental results obtained by means of it are analyzed in Section 6. Finally, Sections 7 and 8 present, respectively, our conclusions and proposals for future work.

2. Processing misspelled queries

Treatment of misspelled queries is usually based on replacing the original search algorithm for exact matches by a more flexible method allowing approximate ones. Having analyzed the state of the art, we consider here two different strategies for dealing with misspelled queries (Manning et al., 2008; Vilares et al., 2011): one that operates at word level and another one that operates at subword level.

As has been said before, the first of these strategies employs the word as working unit. This strategy relies on the use of dictionary-based Natural Language Processing (NLP) techniques in order to implement a query pre-processing stage for detecting and correcting the spelling mistakes that it may contain (Vilares et al., 2011). Once pre-processed, the query is translated and submitted to the system so the search process can be performed by a traditional IR engine. Our spelling correction solutions for this strategy will be described in Section 3.

At this point, we draw attention to the differences between IR and other areas of application of this type of automatic correction such as, for example, word processors. In this latter area, the usual solution consists of performing an ineffective first guess from which the system interacts with the user by showing several candidate corrections, asking the user to choose the right one. However, in the case of IR systems this type of approach is impractical, thus we require more complex, fully automatic error handling approaches with no further user intervention after entering the initial query.

On the other hand, the second strategy operates at sub-word level and consists of using character *n*-grams as processing units (Vilares et al., 2016). This kind of approach, to be described in greater detail in Section 4, can tackle the problem in a simpler way, independently of the degree of knowledge and linguistic resources available.