Information Processing and Management 47 (2011) 692–705

Contents lists available at ScienceDirect

Information Processing and Management

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

Lexical and Syntactic knowledge for Information Retrieval

Antonio Ferrández *

Dept. Languages and Information Systems, Carretera San Vicente S/N, University of Alicante, 03080 Alicante, Spain

ARTICLE INFO

Article history:
Received 14 June 2010
Received in revised form 21 December 2010
Accepted 5 January 2011
Available online 3 February 2011

Keywords:
Information Retrieval
Natural Language Processing
Term Proximity
Question Answering
Lexical and syntactic relationships

ABSTRACT

Traditional Information Retrieval (IR) models assume that the index terms of queries and documents are statistically independent of each other, which is intuitively wrong. This paper proposes the incorporation of the lexical and syntactic knowledge generated by a POS-tagger and a syntactic Chunker into traditional IR similarity measures for including this dependency information between terms. Our proposal is based on theories of discourse structure by means of the segmentation of documents and queries into sentences and entities. Therefore, we measure dependencies between entities instead of between terms. Moreover, we handle discourse references for each entity. It has been evaluated on Spanish and English corpora as well as on Question Answering tasks obtaining significant increases.

1. Introduction

Nowadays, the available information particularly that obtained through the Internet is progressively increasing. The main way to access this information is through Information Retrieval (IR) systems. An IR system takes a user's query as input and returns a set of documents sorted by their relevance to the query. IR systems are usually based on the segmentation of documents and queries into index terms, and their relevance is computed according to the index terms they have in common, as well as according to other information such as the characteristics of the documents (e.g., number of words, hyperlink between papers or bibliographic references) or some probabilistic information (e.g., the estimation of the likelihood that a shared term indicates relevance). Some well-known state-of-the-art models to compute relevance are the cosine measure in its Kaszkiel version (Kaszkiel, Zobel, & Sacks-Davis, 1999), the pivoted cosine (Singhal, Buckley, & Mitra, 1996), the Okapi measure (Roberston, Walker, & Beaulieu, 2000) and the Deviation from Randomness (DFR) measure (Amati, Carpineto, & Romano, 2004). These measures assume that the index terms of queries and documents are statistically independent of each other, thus these are usually called *bag of words* approaches. However, this assumption is intuitively wrong, although this has not so far been proved (Raghavan & Wong, 1986; Wong, Ziarko, & Raghavan, 1987).

In this paper, we present a proposal to include this dependency information between query terms, called LexSIR (Lexical and Syntactic knowledge for IR). It is based on the lexical and syntactic knowledge generated by a POS-tagger and a syntactic Chunker. It has been tested on Spanish and English, proving that it obtains significant improvements in the worst conditions: in an IR system with mean average precision (MAP) of over 0.5, with a more inflected language like Spanish, with different corpora and with different query lengths.

The next section will analyze several proposals that sought to improve the effectiveness of IR systems. Later, we will present our proposal and finally, it will be evaluated.

* Tel.: +34 96 590 3400.
E-mail address: antonio@dlsi.ua.es

0306-4573/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2011.01.003

2. Previous research

Many attempts to incorporate dependency information between index terms have been reported in the past. Some of these attempts use query expansion techniques, which usually improve precision results[1] in IR systems by means of the incorporation of new terms to the query. Our work does not investigate query expansion because we want to isolate the effects of our proposal with just the same set of terms used in traditional IR systems. Therefore, a number of papers are not reported in this section (e.g. Li, 2008 or Peat & Willet, 1991).

Many researchers incorporate the dependency information between index terms by means of the concept of Term Proximity (TP) information. It means the incorporation of the following two intuitions (Vechtomova & Karamuftuoglu, 2008): (1) the closer the terms are in a document, the more likely it is that they are related, and (2) the closer the query terms are in a document, the more likely it is that the document is relevant to the query.

We consider that these previous attempts suffer from the weak points analyzed in the following three subsections.

2.1. The detection of terms that must satisfy proximity restrictions

Researchers usually calculate the TP between pairs of query terms, specifically between all possible combinations of query pairs. For example, in the query *Letter Bomb for Kiesbauer*, TP is calculated for the pairs *"Letter – Bomb, Letter – Kiesbauer, Bomb – Kiesbauer"*. This is a problem in long queries, in which there is not a clear dependency relation between some query terms.[2] Therefore, many researchers evaluate their proposals on short queries[3] in order to reduce the number of possible query pairs.

This problem is also overcome in Mitra, Buckley, Singhal, and Cardie (1997) by means of selecting only those query pairs that co-occur in the corpus at least 25 times. With this method, they obtained an improvement of +3.9% in MAP (from 0.3616 to 0.3758). Their experiments were replicated in Turpin and Moffat (1999), obtaining a maximum improvement of +5.7% (from 0.3373 to 0.3579), and confirming that phrases generated as index terms are not the precision enhancing devices that they should be. Moreover, they observed that using phrases helped at low recall levels, but did not help in the top 20 documents retrieved; and this method works correctly for some query phrases but is quite wrong for others.

Another example is the work by Rasolofo and Savoy (2003) that considers the instance of each pair of terms in the document if they are within a maximal distance of five terms. Nevertheless, as the authors conclude, the weakest point of this work is the restriction of the maximal distance of the pair of terms, because it does not work for every pair of terms: its success depends on the terms, the query and the documents. They obtain an improvement of +2.4% in MAP (from 0.2525 to 0.2586). In the same way, Büttcher, Clarke, and Lushman (2006) reproduce Rasolofo and Savoy's work, and they conclude that TP is more important with longer documents, when the size of the text collection increases and when the queries are stemmed.

Other researchers propose to overcome the problem of the selection of the set of query pairs by means of parsing. For example, Fagan (1989) states that the shortcomings of the nonsyntactic approach can be overcome by incorporating syntactic information into the phrase construction process. However, Mitra et al. (1997) state that there is no significant difference in the benefits obtained from using syntactic vs. statistical phrases. In the same way, in Strzalkowski (1999) several NLP proposals are analyzed, all of them with low benefits. Mittendorfer and Winiwarter (2002) present an approach for exploiting the syntactic structure of a query for an IR system that performs better than a standard vector space model in only some cases. They conclude that a categorization of the queries should be previously performed in order to decide when to use syntactic knowledge. Byung-Kwan, Jee-Hyub, Geunbae, and Jung Yun (2000) index exclusively Korean compound nouns with only a 0.8% of improvement in MAP. With respect to the phrases, they have to devise complex measures of similarity between syntactic trees. Other authors simplify the comparison between syntactic trees by selecting head-modifier[4] pairs (e.g. Alonso, Vilares, & Darriba, 2002; Gonzalez, Strube de Lima, & Valdeni de Lima, 2006; Vilares, Alonso, & Vilares, 2008).

Arampatzis, van der Weide, Koster, and van Bommel (2000) put forward some possible reasons for the lack of success when using Natural Language Processing (NLP) techniques in IR, particularly when using syntactic phrases. They state that, firstly, the currently available NLP techniques suffer from a lack of accuracy and efficiency; and secondly, there are doubts about whether syntactic structure is a good substitute for semantic content. They list the main problems to be solved: morphological, lexical, semantic and syntactic variation, especially for more inflected languages. Consequently, other researchers (e.g. Gonzalez et al., 2006) have also incorporated other lexical variations such as nominalization and transformations of a word into different lexical categories (adjective, verb or adverb).

We propose to overcome the weak points reported in this subsection by adopting the scheme of chunking the query and documents into sets of simple phrases. Thus, our approach is more robust against errors from parsing, it does not require

[1] For example, in DFR measure by Amati et al. (2004), a MAP of 0.5510 is reported with query expansion, whereas 0.4907 is achieved without query expansion.

[2] For example, in the query "*A letter bomb from right-wing radicals sent to the black TV personality Arabella Kiesbauer exploded in a studio of the TV channel PRO7 on June 9th, 1995. An assistant was injured. All reports on the explosion and police inquiries after the event are relevant. Other reports on letter bomb attacks are of no interest*", there is no dependency relation between the pair "letter-interest".

[3] This type of query is usually the *title* version of TREC or CLEF competitions. An example is presented in Fig. 1.

[4] The *head* means the central element of a phrase (main verb or noun), whereas the *modifier* means the remaining modifiers of the head in the phrase.