Information Processing and Management 48 (2012) 1203–1229

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

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

Effective sentence retrieval based on query-independent evidence

Ronald T. Fernández *, David E. Losada

Centro de Investigación en Tecnologías de la Información, Universidad de Santiago de Compostela (CITIUS), Campus Vida s/n, 15782 Santiago de Compostela, Spain

ARTICLE INFO

Article history:
Received 3 February 2011
Received in revised form 24 January 2012
Accepted 30 January 2012
Available online 9 March 2012

Keywords:
Sentence retrieval
Opinion mining
Named entities
Sentence length
Query-independent evidence

ABSTRACT

In this paper we propose an effective sentence retrieval method that consists of incorporating query-independent features into standard sentence retrieval models. To meet this aim, we apply a formal methodology and consider different query-independent features. In particular, we show that opinion-based features are promising. Opinion mining is an increasingly important research topic but little is known about how to improve retrieval algorithms with opinion-based components. In this respect, we consider here different kinds of opinion-based features to act as query-independent evidence and study whether this incorporation improves retrieval performance. On the other hand, information needs are usually related to people, locations or organizations. We hypothesize here that using these named entities as query-independent features may also improve the sentence relevance estimation. Finally, the length of the retrieval unit has been shown to be an important component in different retrieval scenarios. We therefore include length-based features in our study.

Our evaluation demonstrates that, either in isolation or in combination, these query-independent features help to improve substantially the performance of state-of-the-art sentence retrieval methods.

1. Introduction

Retrieving sentences that are relevant to a given user need is a problem that has been addressed in the literature from different perspectives. However, an effective solution is still to be found. Given a set of documents and a user need expressed as a textual query, sentence retrieval (SR) consists of supplying the user with a ranked set of sentences that satisfy his/her user need. Effective solutions to this problem are potentially beneficial in different information retrieval areas, such as question answering, novelty detection and text summarization (Murdock, 2006).

Many of the approaches proposed in the SR literature are direct adaptations of document retrieval methods. These methods are usually based on matching query and sentence terms. Nevertheless, sentences are very short pieces of text and, therefore, there are usually very few matching terms. Some researchers tried to alleviate this problem by applying query expansion. However, we take here an alternative approach focused on combining query-independent evidence (related to the sentences) with sentence retrieval scores, leading to effective estimations of the relevance of sentences. More specifically, we consider opinion-related information, named entities and the sentence length as query-independent features. Our intuition is that, in many situations, users are mostly interested in subjective material. This is usually the case with news articles about controversial topics. In these cases, the subjective pieces of information (people's opinions, politician's views, etc.) are likely more important than objective statements related to the topic. For instance, given a query "*partial birth abortion ban*", an opinionated sentence such as "*Eventually, he'd like all abortions to be banned because he believes they are*

* Corresponding author. Tel.: +34 881 813 569; fax: +34 981 528 012.
E-mail addresses: ronald.teijeira@usc.es (R.T. Fernández), david.losada@usc.es (D.E. Losada).

0306-4573/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2012.01.007

murder" is likely more important than another sentence such as "*We are performing so-called partial-birth abortions as defined by Kansas law*". Similarly, many information needs are related to a person, a location or an organization. Therefore, sentences that contain named entities may be highly relevant. For instance, given the query "*U.S. Embassy bombings in Africa, 1998*", the sentence "*Suspected bombs exploded outside the U.S. embassies in the Kenyan and Tanzanian capitals Friday [...]*", which contains explicit references to certain African locations, might be more important than another sentence such as "*At least 82 were killed and more than 1,700 injured, officials said as dawn broke Saturday*". Finally, the use of sentence length as a query-independent feature may also be helpful because long sentences usually provide more information than short ones and, therefore, they are more likely relevant (some short sentences act solely as connectors between the pieces of the discourse).

Summing up, the set of features considered in our study are:

(a) *opinion-based features*, including the subjectivity nature of sentences (a sentence may be objective or subjective) and the polarity of the sentence terms (the number of positive terms in a sentence, the number of negative terms, and the number of opinionated terms). Observe that sentence retrieval is an appropriate scenario to study these issues because sentences are compact pieces of information and their subjective or objective nature can be reasonably estimated (with coarse-grained chunks such as documents or paragraphs, this opinion-based classification is more problematic because it is hard to classify a document as subjective or objective).
(b) *named entities features*, i.e. names of persons, locations, organizations, etc.
(c) *sentence length*, i.e. the number of terms in a sentence, ignoring stopwords.

The features described above are considered in isolation or in combination. This helps to understand the configuration of query-independent features that performs the best. In order to incorporate these sentence features as query-independent evidence into SR models, we follow a formal methodology based on kernel density estimation (Craswell, Robertson, Zaragoza, & Taylor, 2005). We show that the combination of these query-independent features with state of the art SR scores yields to important improvements in performance with negligible computational costs at retrieval time.

The rest of the paper is organized as follows. Section 2 comments on some related work. The methodology followed to combine sentence retrieval scores with opinion-based scores is explained in Section 3. Section 4 presents the query-independent features and the software utilized to estimate them. Section 5 reports the experiments and analyzes their outcomes. The paper ends with Section 6, where we expose the conclusions of our study.

2. Related work

In the sentence retrieval literature, most of the proposals consist of addressing the SR problem by adapting document retrieval methods with little change. We believe that this is problematic because the peculiarities of the task are largely ignored. Sentences are short pieces of information. Most sentence retrieval methods are based on a regular matching between query and sentences. However, sentences that do not contain query terms may be relevant for a query. Query expansion is a mechanism that tries to address this vocabulary mismatch problem, which is rather severe in sentence retrieval. The study conducted by Losada in Losada (2010) analyzes carefully different query expansion methods applied to sentence retrieval. This included well-known term selection techniques, such as those based on regular pseudo-relevance feedback and Local Context Analysis (Xu & Croft, 1996, 2000), and two different expansion configurations: before and after sentence retrieval. The paper concludes that the ideal expansion configuration depends strongly on the quality of the initial query. Evolved expansion methods, based on selective feedback, were studied by Jaleel et al. in Abdul-Jaleel et al. (2004). They are more stable than standard feedback methods but require training data. On the other hand, other authors resort to lexical expansion, i.e. they utilize query-related terms (i.e. synonyms or related terms from a lexical resource) to expand the query. This approach may not be appropriate because noisy terms are likely introduced into the expanded query (Voorhees, 1993) and, moreover, a large terminological resource is not always available.

Given the inconsistent effects on performance and the time requirements involved at query time, query expansion is problematic for sentence retrieval. We therefore take here a different avenue to address retrieval problems at sentence level. We claim that the estimation of relevance could be more accurate by using query-independent information. In the literature, there is not much evidence about the combination of query-dependent and query-independent information to estimate relevance for SR problems. We consider some opinion-based features and study whether or not they help to improve sentence retrieval performance. We also include other features, such as name entities and sentence length, in our study. Additionally, we analyze whether the combination of features of the same or different nature improves performance over individual incorporations.

The use of opinions for sentence retrieval was also explored by Kim et al. (2004). For the TREC 2003 and 2004 opinion topics, relevant opinion sentences were recognized using opinion-bearing word lists. However, the authors assumed that opinion-based methods are only effective for opinion topics. We demonstrate here otherwise. Furthermore, the performance achieved by the methods described in Kim et al. (2004) was not higher than the performance of state of the art methods. Unfortunately, the experiments reported in Kim et al. (2004) cannot be replicated here because they are based on collecting manually opinion-bearing words from resources such as WordNet. Since the manual lists are not publicly available we