Information Processing and Management 51 (2015) 595–615

Contents lists available at ScienceDirect

Information Processing and Management

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

Bridging the vocabulary gap between questions and answer sentences

Saeedeh Momtazi [*], Dietrich Klakow

Spoken Language Systems, Saarland University, Saarbrücken, Germany

ARTICLE INFO

Article history:
Received 16 July 2013
Received in revised form 14 April 2015
Accepted 22 April 2015
Available online 12 June 2015

Keywords:
Sentence retrieval
Language modeling
Word Clustering
Triggering
Question answering

ABSTRACT

We propose two novel language models to improve the performance of sentence retrieval in Question Answering (QA): *class-based language model* and *trained trigger language model*. As the search in sentence retrieval is conducted over smaller segments of text than in document retrieval, the problems of data sparsity and exact matching become more critical. Different techniques such as the translation model are also proposed to overcome the word mismatch problem. Our class-based and trained trigger language models, however, use different approaches to this aim and are shown to outperform the exiting models. The class model uses word clustering algorithm to capture term relationships. In this model, we assume a relation between the terms that belong to the same clusters; as a result, they can be substituted when searching for relevant sentences. The trigger model captures pairs of trigger and target words while training on a large corpus. The model considers a relation between a question and a sentence, if a trigger word appears in the question and the sentence contains the corresponding target word. For both proposed models, we introduce different notions of co-occurrence to find word relations. In addition, we study the impact of corpus size and domain on the models. Our experiments on TREC QA collection verify that the proposed model significantly improves the sentence retrieval performance compared to the state-of-the-art translation model. While the translation model based on mutual information (Karimzadehgan and Zhai, 2010) has 0.3927 Mean Average Precision (MAP), the class model achieves 0.4174 MAP and the trigger model enhances the performance to 0.4381.

1. Introduction

Sentence retrieval plays an important role in QA systems. It aims at finding small segments of text that contain an exact answer to the users' questions rather than overwhelm them with a large number of retrieved documents, which they must sort through to find the desired answer. The retrieved sentences are then further processed using a variety of techniques to extract the final answers.

Different techniques used for document retrieval have been also applied to sentence-level retrieval. Among them, language model-based information retrieval proposed by Ponte and Croft (1998) has been proven to outperform traditional methods like *TF-IDF* and *Okapi* (Merkel & Klakow, 2007). However, the performance of this model in sentence retrieval is

[*] Corresponding author at: Ubiquitous Knowledge Processing Lab (UKP-DIPF), German Institute for Educational Research and Educational Information, Schlossstrasse 29, 60486 Frankfurt/Main, Germany. Tel.: +49 69 24708866.

E-mail address: momtazi@ukp.informatik.tu-darmstadt.de (S. Momtazi).

http://dx.doi.org/10.1016/j.ipm.2015.04.005
0306-4573/© 2015 Elsevier Ltd. All rights reserved.

worse than the task of retrieving documents, which indicates that current document retrieval techniques are not sufficient for finding relevant sentences. Because there are major differences between document retrieval and sentence retrieval in terms of their size. As a result, many of the assumptions made for document retrieval do not hold for sentence retrieval. One important reason is the brevity of sentences. The brevity of sentences exacerbates the usual term mismatch problem which should be taken into consideration while designing a sentence retrieval engine. This problem is not pronounce in document level retrieval, since documents are large and they normally contain many terms which express the same concept in multiple ways. Such a property leads to there being different terms which talk about the same concept. As a result, even having one of these terms in the query is enough to find the relevant document. In addition, the frequency of individual words in a document is normally greater than one, especially for words that talk about the main topic of the document. Thus the frequency of words in documents helps to find the documents which talk about the user's information need. In sentence retrieval, contrary to document retrieval, we face a very short text which briefly talks about a topic, and it is very unlikely to find sentences which contain the same terms as query terms. Even if the query words appear in a sentence, the frequency of those words is not higher than one or two.

The word unigram model is the most common approach used in the majority of information retrieval literature. When estimating word unigrams, the model only considers the exact literal words present in the query. Since no relationship between words is considered by this model, it will fail to retrieve other relevant information. Even well-known approaches for overcoming term mismatch problem like automatic query expansion and pseudo-relevance feedback, which have been great success stories for document retrieval, had rather mixed success when they were applied to sentence retrieval (Murdock, 2006). This problem motivated us to find a more sophisticated model to search for query words rather than just their distributions in the sentences to be retrieved.

One of the useful approaches is capturing word relationship to reduce the negative effects of the term mismatch problem in sentence retrieval. This goal can be achieved using different approaches such as the class-based language model and the trained trigger language model. The class-based model addresses the sentence brevity problem and bridges the semantic vocabulary gap between the question and the sentences. In this method, we apply a word clustering algorithm to capture word relationships. In other words, instead of building a fine-grained model based on vocabulary terms, a coarse-grained model based on word clusters is built. As a result, the class-based model is able to retrieve sentences which have no or few shared terms with input question, but their words belong to the same clusters as the question terms. The trained trigger model captures pairs of trigger and target words based on their co-occurrences in the training corpus. The model uses unsupervised approaches when training on a large corpus of raw text and apply supervised approaches when training a corpus of question and answer sentence pairs. Having pairs of trigger and target words; in the retrieval step, if the trigger word appears in an input question and the target word appears in a sentence, then we can say there is a relation between the question and the sentence even though they share no or few terms.

The structure of the paper is as follows. In Section 2, we review related work for both general language modeling approaches and term relationship approaches in information retrieval. In Section 3, the class-based model and the word clustering algorithm are introduced. Different notions of word co-occurrence applied in our class-based model are described in Section 4. Section 5 introduces the proposed trained trigger language model. In this section, we also show how the proposed model can be integrated with the exact matching methods to improve the system performance. Notions of word co-occurrence for triggering model are described in Section 6. In Section 7, the experimental results are presented. Finally, Section 8 concludes the paper.

2. Related work

2.1. Language models for information retrieval

Statistical language modeling has been successfully used in speech recognition (Jelinek, 1998) and many natural language processing tasks, including part of speech tagging, syntactic parsing (Charniak, 1993), and machine translation (Brown et al., 1990). Language modeling for information retrieval has received researchers' attention in recent years. The efficiency of this approach, its simplicity, the state-of-the-art performance, and clear probabilistic meaning are the most important factors which contribute to its popularity (Ponte & Croft, 1998; Zhai & Lafferty, 2001).

The idea of using language modeling techniques for information retrieval was first introduced by Ponte and Croft (1998). In the proposed method, called query likelihood, a language model is inferred for each document and then the likelihood of the query according to the estimated language model is calculated. Thereafter, documents are ranked based on their query likelihood scores. Using the query likelihood model, the similarity between a document and the input query is defined as the probability of generating the query Q given the document model D which can be estimated by the multiple Bernoulli model or the multinomial model.

In the original method proposed by Ponte and Croft (1998), the multiple Bernoulli model was used for estimating document model D. In this model, each query is considered a set of unique terms, and two different probabilities are calculated: the probability of producing the query terms, and the probability of not producing other terms. Then, the product of these two factors is used as the model. The multiple Bernoulli model is defined as follows: