Information Processing and Management 47 (2011) 843–855

Contents lists available at ScienceDirect

Information Processing and Management

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

Improving graph-based random walks for complex question answering using syntactic, shallow semantic and extended string subsequence kernels

Yllias Chali [a], Sadid A. Hasan [a,*], Shafiq R. Joty [b]

[a] University of Lethbridge, Lethbridge, AB, Canada
[b] University of British Columbia, Vancouver, BC, Canada

ARTICLE INFO

Article history:
Received 29 March 2009
Received in revised form 27 September 2010
Accepted 3 October 2010
Available online 19 November 2010

Keywords:
Complex question answering
Graph-based method
Syntactic kernel
Shallow semantic kernel
Extended string subsequence kernel

ABSTRACT

The task of answering complex questions requires inferencing and synthesizing information from multiple documents that can be seen as a kind of topic-oriented, informative multi-document summarization. In generic summarization the stochastic, graph-based random walk method to compute the relative importance of textual units (i.e. sentences) is proved to be very successful. However, the major limitation of the TF*IDF approach is that it only retains the frequency of the words and does not take into account the sequence, syntactic and semantic information. This paper presents the impact of syntactic and semantic information in the graph-based random walk method for answering complex questions. Initially, we apply tree kernel functions to perform the similarity measures between sentences in the random walk framework. Then, we extend our work further to incorporate the Extended String Subsequence Kernel (ESSK) to perform the task in a similar manner. Experimental results show the effectiveness of the use of kernels to include the syntactic and semantic information for this task.

1. Introduction

The size of the publicly indexable world-wide-web has probably surpassed several billions of documents and as yet growth shows no sign of leveling off. The demand for access to different types of information have led to a renewed interest in a broad range of Information Retrieval (IR) related areas such as question answering, topic detection and tracking, summarization, multimedia retrieval (e.g., image, video and music), chemical and biological informatics, text structuring, text mining, genomics, etc.

Automated Question Answering (QA) – the ability of a machine to answer questions, simple or complex, posed in ordinary human language is perhaps the most exciting technological development of the past six or seven years. QA research attempts to deal with a wide range of question types including fact, list, definition, how, why, hypothetical, semantically-constrained, and cross-lingual questions. Search collections vary from a small local document collections, to an internal organization documents, to compiled newswire reports, to the World Wide Web.

Some questions are easier to answer which we call *simple questions*. For example, the question, “Who is the president of Bangladesh?” asks for a person name. This type of questions (i.e. factoid) requires small snippets of text as the answers. Again, the question “Which countries has Pope John Paul II visited?” asks for a list of small snippets (i.e. list questions) of text. Finding answers to these questions are easier than questions that have complex information needs. After having made substantial headway in factoid and list questions, researchers have turned their attention to more complex information

[*] Corresponding author.
E-mail addresses: chali@cs.uleth.ca (Y. Chali), hasan@cs.uleth.ca (S.A. Hasan), rjoty@cs.ubc.ca (S.R. Joty).

0306-4573/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2010.10.002

needs that cannot be answered by simply extracting named entities (persons, organization, locations, dates, etc.) from documents. Unlike simple factoid questions, complex questions often seek multiple different types of information simultaneously and do not presuppose that one single answer could meet all of its information needs. For example, with a factoid question like "How accurate are HIV tests?", it can be safely assumed that the submitter of the question is looking for a number or a range of numbers. However, with complex questions like "What are the causes of AIDS?", the wider focus of this question suggests that the submitter may not have a single or well-defined information need and therefore may be amenable to receiving additional supporting information that is relevant to some (as yet) undefined informational goal. This type of questions require inferencing and synthesizing information from multiple documents. This information synthesis in Natural Language Processing (NLP) can be seen as a kind of topic-oriented, informative multi-document summarization, where the goal is to produce a single text as a compressed version of a set of documents with a minimum loss of relevant information (Amigo, Gonzalo, Peinado, Peinado, & Verdejo, 2004).

The graph-based methods (such as LexRank Erkan & Radev, 2004, TextRank Mihalcea & Tarau, 2004) are applied successfully to generic, multi-document summarization. A topic-sensitive LexRank is proposed in Otterbacher, Erkan, and Radev (2005). In this method, a sentence is mapped to a vector in which each element represents the occurrence frequency (TF*IDF) of a word. However, the major limitation of the TF*IDF approach is that it only retains the frequency of the words and does not take into account the sequence, syntactic and semantic structure. Thus, it cannot distinguish between "The hero killed the villain" and "The villain killed the hero". For the task like *answering complex questions* that requires the use of more complex syntactic and semantics, the approaches with only TF*IDF are often inadequate to perform fine-level textual analysis (Chali & Joty, 2008; Chali & Joty, 2008).

In this paper, we extensively study the impact of syntactic and semantic information in measuring similarity between the sentences in the random walk framework for answering complex questions. We apply the tree kernel functions and Extended String Subsequence Kernel (ESSK) to include syntactic and semantic information. We run our experiments on the DUC 2007 data and based on this we argue that for the complex question answering task, similarity measures based on syntactic and semantic information perform better and can be used to characterize the relation between a question and a sentence (answer) in a more effective way than the traditional TF*IDF based similarity measures.

We organize the paper as follows: Section 2 focuses on the related work, Section 3 describes the graph-based model, Section 4 presents the methods to encode syntactic and shallow semantic structures, Section 5 discusses the syntactic and shallow semantic kernels, Section 6 discusses the theory of Extended String Subsequence Kernel, Section 7 describes the redundancy checking and summary generation module whereas Section 8 presents the experimental details with evaluation results and finally, Section 9 concludes the paper by cuing some future directions.

2. Related work

There are approaches in "recognizing textual entailment", "sentence alignment", and "question answering" that use syntactic and/or semantic information in order to measure the similarity between two textual units. Indeed, this motivated us to include syntactic and semantic features to get the structural similarity between sentences. In MacCartney, Grenager, de Marneffe, Cer, and Manning (2006), they use typed dependency graphs (same as dependency trees) to represent the text and the hypothesis. They try to find a good partial alignment between the typed dependency graphs representing the hypothesis (contains n nodes) and the text (graph contains m nodes) in a search space of $O((m+1)n)$. They use an incremental beam search combined with a node ordering heuristic to do approximate global search in the space of possible alignments. A locally decomposable scoring function was chosen such that the score of an alignment is the sum of the local node and edge alignment scores. The scoring measure is designed to favor alignments which align semantically similar subgraphs, irrespective of polarity. For this reason, nodes receive high alignment scores when the words they represent are semantically similar. Synonyms and antonyms receive the highest score and unrelated words receive the lowest. Alignment scores also incorporate local edge scores which are based on the shape of the paths between nodes in the text graph which correspond to adjacent nodes in the hypothesis graph. In the final step they make a decision about whether or not the hypothesis is entailed by the text conditioned on the typed dependency graphs as well as the best alignment between them. To make this decision they use a supervised statistical logistic regression classifier (with a feature space of 28 features) with a Gaussian prior parameter for regularization.

The importance of syntactic and semantic features in finding textual similarity is described by Zhang and Lee (2003), Moschitti, Quarteroni, Basili, and Manandhar (2007) and Moschitti and Basili (2006). An effective way to integrate syntactic and semantic structures in machine learning algorithms is the use of *tree kernel* functions (Collins & Duffy, 2001) which has been successfully applied to question classification (Zhang & Lee, 2003; Moschitti & Basili, 2006). To the best of our knowledge, no study has used tree kernel functions to encode syntactic/semantic information for more complex tasks such as computing the relatedness between the query sentences and the document sentences. Another good way to encode some shallow syntactic information is the use of Basic Elements (BE) (Hovy, Lin, Zhou, & Fukumoto, 2006) which uses dependency relations. Moreover, the study of shallow semantic information such as predicate argument structures annotated in the PropBank (PB) project (Kingsbury & Palmer, 2002) is a promising research direction.

In Hirao, Suzuki, Isozaki, and Maeda (2004), they represent the sentences using Dependency Tree Path (DTP) to incorporate syntactic information. They apply String Subsequence Kernel (SSK) to measure the similarity between the DTPs of two