Using Semantic Information to Answer Complex Questions

Yllias Chali, Sadid A. Hasan, and Kaisar Imam

University of Lethbridge
Lethbridge, AB, Canada
{chali,hasan,imam}@cs.uleth.ca

Abstract. In this paper, we propose the use of semantic information for the task of answering complex questions. We use the Extended String Subsequence Kernel (ESSK) to perform similarity measures between sentences in a graph-based random walk framework where semantic information is incorporated by exploiting the word senses. Experimental results on the DUC benchmark datasets prove the effectiveness of our approach.

Keywords: Complex Question Answering, Graph-based Random Walk Method, Extended String Subsequence Kernel

1 Introduction

Resolving complex information needs is not possible by simply extracting named entities (persons, organizations, locations, dates, etc.) from documents. Complex questions often seek multiple different types of information simultaneously and do not presuppose that one single answer can meet all of its information needs. For example, with a factoid question like: "What is the magnitude of the earthquake in Haiti?", it can be safely assumed that the submitter of the question is looking for a number. However, the wider focus of a complex question like: "How is Haiti affected by the earthquake?" suggests that the user may not have a single or well-defined information need and therefore may be amenable to receiving additional supporting information relevant to some (as yet) undefined informational goal [6]. 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 [1]. So, in this paper, given a complex question and a set of related data, we generate a summary in order to use it as an answer to the complex question. The graph-based methods (such as LexRank [4], TextRank [10]) are applied successfully to generic, multi-document summarization. In topic-sensitive LexRank [11], a sentence is mapped to a vector in which each element represents the occurrence frequency (TF–IDF[1]) of a word. However, for the task like *answering*

[1] The TF–IDF (term frequency-inverse document frequency) is a statistical measure used to evaluate how important a word is to a document in a collection or corpus.

complex questions that requires the use of more complex semantic analysis, the approaches with only TF–IDF are often inadequate to perform fine-level textual analysis. In this paper, we extensively study the impact of using semantic information in the random walk framework for answering complex questions. We apply the Extended String Subsequence Kernel (ESSK) [8] to include semantic information by incorporating disambiguated word senses. We run all experiments on the DUC[2] 2007 data. Evaluation results show the effectiveness of our approach.

2 Background and Proposed Framework

2.1 Graph-based Random Walk

In [4], the concept of graph-based centrality is used to rank a set of sentences, in producing generic multi-document summaries. A similarity graph is produced for the sentences in the document collection. In the graph, each node represents a sentence. The edges between nodes measure the cosine similarity between the respective pair of sentences where each sentence is represented as a vector of term specific weights. The term specific weights in the sentence vectors are products of local and global parameters. The model is known as term frequency-inverse document frequency (TF–IDF) model. To apply the LexRank in a query-focused context, a topic-sensitive version of LexRank is proposed in [11]. The score of a sentence is determined by a mixture model of the relevance of the sentence to the query and the similarity of the sentence to other high-scoring sentences. The relevance of a sentence s to the question q is computed by:

$$rel(s|q) = \sum_{w \in q} log\,(tf_{w,s} + 1) \times log\,(tf_{w,q} + 1) \times idf_w$$

where, $tf_{w,s}\ and\ tf_{w,q}$ are the number of times w appears in s and q, respectively. A sentence that is similar to the high scoring sentences in the cluster should also have a high score. For instance, if a sentence that gets high score based on the question relevance model is likely to contain an answer to the question, then a related sentence, which may not be similar to the question itself, is also likely to contain an answer. This idea is captured by the following mixture model [11]:

$$p(s|q) = d \times \frac{rel(s|q)}{\sum_{z \in C} rel(z|q)} + (1-d) \times \sum_{v \in C} \frac{sim(s,v)}{\sum_{z \in C} sim(z,v)} \times p(v|q) \qquad (1)$$

2.2 Our Approach

We claim that for a complex task like answering complex questions where the relatedness between the query sentences and the document sentences is an important factor,

[2] Document Understanding Conference– http://duc.nist.gov/

the graph-based method of ranking sentences would perform better if we could encode the semantic information instead of just the TF–IDF information in calculating the similarity between sentences. Thus, our mixture model for answering complex questions is:

$$p(s|q) = d \times SEMSIM(s,q) + (1-d) \times \sum_{v \in C} SEMSIM(s,v) \times p(v|q) \qquad (2)$$

where, *SEMSIM(s,q)* is the normalized semantic similarity between the *query (q)* and *the document sentence (s)* and C is the set of all sentences in the collection. In this paper, we encode semantic information using ESSK [7] and calculate the similarity between sentences. We reimplemented ESSK considering each word in a sentence as an "alphabet", and the alternative as its disambiguated sense [3] that we find using our Word Sense Disambiguation (WSD) System [2]. We use a dictionary based disambiguation approach assuming one sense per discourse. We use WordNet [5] to find the semantic relations among the words in a text. We assign weights to the semantic relations. Our WSD technique can be decomposed into two steps: (1) building a representation of all possible senses of the words and (2) disambiguating the words based on the highest score. We use an intermediate representation (disambiguation graph) to perform the WSD. We sum the weight of all edges leaving the nodes under their different senses. The one sense with the highest score is considered the most probable sense. In case of tie between two or more senses, we select the sense that comes first in WordNet, since WordNet orders the senses of a word by decreasing order of their frequency.

3 Evaluation and Analysis

3.1 Task Definition

In this research, we consider the main task of DUC 2007 to run our experiments. The task was: "Given a complex question (topic description) and a collection of relevant documents, the task is to synthesize a fluent, well-organized 250-word summary of the documents that answers the question(s) in the topic". We choose 35 topics randomly from the given dataset and generate summaries for each of them according to the task guidelines.

3.2 Automatic Evaluation

We carried out the automatic evaluation of our summaries using ROUGE [9] toolkit (i.e. ROUGE-1.5.5 in this study). The comparison between the TF–IDF system and the ESSK system is presented in Table 1. To compare our systems' performance with the state-of-the-art systems, we also list the ROUGE scores of the NIST baseline system (defined in DUC-2007) and the best system in DUC-2007. The NIST baseline system generated the summaries by returning all the leading sentences (up to 250 words) in the $\langle TEXT \rangle$ field of the most recent document(s). Analysis of the results show that the ESSK system improves the ROUGE-1 and ROUGE-SU scores over the TF–IDF system by 0.26%, and 1.48%, respectively whereas the ESSK system performs closely to the best system besides beating the baseline system by a considerable margin.

Systems	ROUGE-1	ROUGE-SU
TF–IDF	0.379	0.135
ESSK	0.380	0.137
NIST Baseline	0.334	0.112
Best System	0.438	0.174

Table 1. ROUGE F-scores for all systems

3.3 Manual Evaluation

Even if the ROUGE scores had significant improvement, it is possible to make bad summaries that get state-of-the-art ROUGE scores [12]. So, we conduct an extensive manual evaluation in order to analyze the effectiveness of our approach. Each summary is manually evaluated for a Pyramid-based evaluation of contents and also a user evaluation is conducted to get the assessment of readability (i.e. fluency) and overall responsiveness according to the TAC 2010 summary evaluation guidelines[3].

Content Evaluation In the DUC 2007 main task, 23 topics were selected for the optional community-based pyramid evaluation. Volunteers from 16 different sites created pyramids and annotated the peer summaries for the DUC main task using the given guidelines[4]. 8 sites among them created the pyramids. We used these pyramids to annotate a randomly chosen 5 peer summaries for each of our system to compute the modified pyramid scores. Table 2 shows the modified pyramid scores of all the systems including the NIST baseline system and the best system of DUC-2007. From these results we see that all the systems perform better than the baseline system and ESSK performs the best.

Systems	Modified Pyramid Scores
NIST Baseline	0.139
Best System	0.349
TF–IDF	0.512
ESSK	0.547

Table 2. Modified pyramid scores for all systems

User Evaluation Some university graduate students judged all the system generated summaries (70 summaries in total) for readability (fluency) and overall responsiveness. The readability score reflects the fluency and readability of the summary (independently of whether it contains any relevant information) and is based on factors such as the summary's grammaticality, non-redundancy, referential clarity, focus, and structure

[3] http://www.nist.gov/tac/2010/Summarization/Guided-Summ.2010.guidelines.html

[4] http://www1.cs.columbia.edu/ becky/DUC2006/2006-pyramid-guidelines.html

and coherence. The overall responsiveness score is based on both content (coverage of all required aspects) and readability. The readability and overall responsiveness is each judged on a 5-point scale between 1 (very poor) and 5 (very good). Table 3 presents the average readability and overall responsive scores of all the systems. Again, the NIST–generated baseline system's scores and the best DUC-2007 system's scores are given for meaningful comparison. The results show that the ESSK system improves the readability and overall responsiveness scores over the TF–IDF system by 30.61%, and 42.17%, respectively while it performs closely to the best system's scores besides beating the baseline system's overall responsiveness score by a significant margin.

Systems	Readability	Overall Responsiveness
NIST Baseline	4.24	1.80
Best System	4.11	3.40
TF–IDF	2.45	2.30
ESSK	3.20	3.27

Table 3. Readability and overall responsiveness scores for all systems

4 Conclusion

In this paper, we used semantic information and showed its impact in measuring the similarity between the sentences in the random walk framework for answering complex questions. We used Extended String Subsequence Kernel (ESSK) to include semantic information by applying disambiguated word senses. We evaluated the systems automatically using ROUGE and reported an extensive manual evaluation to further analyze the performance of the systems. Comparisons with the state-of-the-art systems showed effectiveness of our proposed approach.

Acknowledgments

The research reported in this paper was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada – discovery grant and the University of Lethbridge.

References

1. Amigo, E., Gonzalo, J., Peinado, V., Peinado, A., Verdejo, F.: An Empirical Study of Information Synthesis Tasks. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. pp. 207–214. Barcelona, Spain (2004)
2. Chali, Y., Joty, S.R.: Word Sense Disambiguation Using Lexical Cohesion. In: Proceedings of the 4th International Conference on Semantic Evaluations. pp. 476–479. ACL, Prague (2007)

3. Chali, Y., Hasan, S.A., Joty, S.R.: Improving Graph-based Random Walks for Complex Question Answering Using Syntactic, Shallow Semantic and Extended String Subsequence Kernels. Information Processing and Management In Press, Corrected Proof (2010), http://www.sciencedirect.com/science/article/B6VC8-51H5SB4-1/2/4f5355410ba21d61d3ad9f0ec881e740
4. Erkan, G., Radev, D.R.: LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research 22, 457–479 (2004)
5. Fellbaum, C.: WordNet - An Electronic Lexical Database. MIT Press, Cambridge, MA (1998)
6. Harabagiu, S., Lacatusu, F., Hickl, A.: Answering complex questions with random walk models. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 220–227. ACM (2006)
7. Hirao, T., , Suzuki, J., Isozaki, H., Maeda, E.: Dependency-based Sentence Alignment for Multiple Document Summarization. In: Proceedings of COLING 2004. pp. 446–452. COLING, Geneva, Switzerland (2004)
8. Hirao, T., Suzuki, J., Isozaki, H., Maeda, E.: NTT's Multiple Document Summarization System for DUC2003. In: Proceedings of the Document Understanding Conference (2003)
9. Lin, C.Y.: ROUGE: A Package for Automatic Evaluation of Summaries. In: Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of Association for Computational Linguistics. pp. 74–81. Barcelona, Spain (2004)
10. Mihalcea, R., Tarau, P.: TextRank: Bringing Order into Texts. In: Proceedings of the Conference of Empirical Methods in Natural Language Processing. Barcelona, Spain (2004)
11. Otterbacher, J., Erkan, G., Radev, D.R.: Using Random Walks for Question-focused Sentence Retrieval. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. pp. 915–922. Vancouver, Canada (2005)
12. Sjöbergh, J.: Older Versions of the ROUGEeval Summarization Evaluation System Were Easier to Fool. Information Processing and Management 43, 1500–1505 (2007)