Semantic Multi-document Update Summarization Techniques

Rakesh Verma
Computer Science Department
University of Houston
4800 Calhoun Rd., Houston, TX 77204
rverma@uh.edu

David Kent
Computer Science Department
Case Western Reserve University
Cleveland, OH 44106

Ping Chen
Computer and Mathematical Sciences Department
University of Houston-Downtown
1 Main St., Houston, TX 77002
chenp@uhd.edu

January 7, 2009

Abstract

As huge amounts of knowledge are created rapidly, effective information access becomes an important issue. Especially for critical domains, such as medical and financial areas, efficient retrieval of concise and relevant information is highly desired. In this paper we propose two new multi-document summarization techniques that make use of WordNet, a general knowledge source from Princeton University. We participated in the Text Analysis Conference 2008 update summarization task and ranked in the middle tier of about 70 systems.

KEYWORDS: WordNet, Multi-document Summarization, Update Summarization

1 Introduction

The explosive growth of the world wide web and the increase in web "authors" have led to a growing need for people to deal with an overwhelming number of text documents on a daily basis. Examples of text documents include newspaper articles, discussion forums posts, research papers, etc. Thus automatic text summarization will increasingly be an indispensable aid in the future. The goal of automatic text summarization, which is a very challenging task, is to condense text documents into their essence, and display the result to the user. The consequence of achieving this goal would be the ability for the users to effectively manage more textual data in less or equal time.

Automatic text summarization methods are of two types: abstractive and extractive. Abstractive methods may construct a summary using sentences that are not necessarily in the original document or some other abstract representation of the document, e.g., [3]. Extractive methods on the other hand rely only on sentences in the original document. In this paper, the focus is on extractive methods so hence-

forth we will not refer to any abstractive methods from the literature and whenever we speak of a summarization method the term extractive is implicit.

1.1 Related Work

Many automatic text summarization systems employ a vast array of statistical methods, e.g., see [1, 4, 5] and the references cited therein. These methods usually treat text documents as a bag of words with no order, or meaning. Using this idea, many systems were developed to be modestly successful. However, a sentence is more than just a collection of unordered words. Each sentence carries meaning, and a truly good summary can be constructed only if meaning is incorporated into the system. Recently, some research groups have started experimenting with incorporating some semantics into their systems, e.g., see the proceedings of the Document Understanding Conference for the last three years [6, 7, 8] at `http://www-nlpir.nist.gov/projects/duc/pubs.html`. For instance, WordNet has been used to build lexical chains of word synonyms for sentence filtering. We use WordNet in a novel way for sentence filtering in our framework called Document Map.

In this paper, we discuss background information about WordNet and the medical ontology knowledge sources in Section 2. Our summarization system architecture and algorithm are presented in Section 3. We present the evaluation results from the DUC 2007 main task in Section 4. In Section 5 we conclude and discuss our future work.

2 Ontology knowledge

We use one ontology knowledge source in our summarization system, WordNet, for which we give a brief overview here.

2.1 WordNet

WordNet is utilized to decide which sentences are more general, with the assumption that more general sentences are more likely to be thematic sentences. According to Fellbaum [2] "WordNet is neither a traditional dictionary nor a traditional thesaurus but combines features of both types. It resembles a thesaurus in that its building block is a synset consisting of all the words that express a given concept." According to Miller [2], "The basic semantic relation in WordNet is synonymy. Sets of synonyms (synsets) form the basic building blocks. Although synonymy is a semantic relation between word forms, the semantic relation that is the most important in organizing nouns is a relation between lexicalized concepts. It is relation of subordination (or class inclusion or subsumption), which is called hyponymy." Miller writes in [2] that (page 26): "Since a noun usually has a single hypernym, lexicographers include it in the definition." The key point to be noted is that although the hypernymy relation is defined on synsets in WordNet, and hence it could happen that a synset can have more than one hypernym, this situation is not frequent.[1] The reason is that a synset is designed to refer to a single concept and hence we need to disambiguate words in the document to find the correct synset for a noun. For instance the word plant could mean a factory in one context and could mean a tree in another context. Hence the word plant would be found in two different synsets in this case. The relation between nouns to other nouns, and verbs to other verbs is used by WN-SUM.

We use the hypernymy relation between nouns, which is defined as follows: A is a *hypernym* of B if the meaning of A encompasses the meaning of B (B is called the *hyponym*). For example, *animal* is a hypernym of *dog*, and a *dog* is a hyponym of *animal*.

[1] We do take care of the situation in which there are multiple hypernyms as explained in the WordNet score subsection.

All nouns in WordNet are stored in a graph (that is close to a tree) that represents the hypernymy hierarchy. The word *entity* is the root of the tree, because it is believed to encompass the meaning of all other nouns. Traversing down the tree manifests more specific nouns. Figure 1 shows a very small portion of the hypernymy tree.

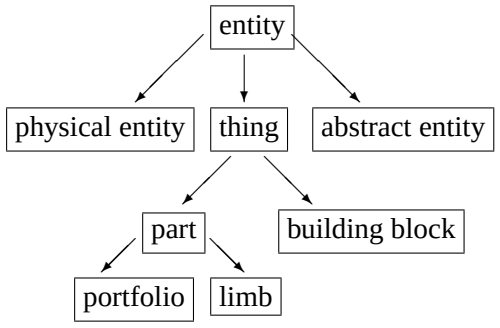

Figure 1: Sample of WordNet hypernymy relation.

From observing the tree, it can be seen that more specific nouns are closer to leaves, and more general nouns are closer to the root. It is assumed that sentences that have nouns closer to the root are more likely to be thematic sentences.

WordNet also stores hypernymy relationships between verbs. The graph for verbs is not an almost tree. Within the graph there might be multiple root nodes. The assumption is, again, that a verb closer to any of these roots is more general as compared to a verb closer to any of the leaves. Therefore, sentences that have verbs closer to a root are considered more likely to be thematic sentences.

3 Summarization System Architecture

Two systems were devised for multi-document update summarization. Both systems used WordNet distance of words in the sentence from words in the topic title, a position score for the sentence in its document and a topic score. For the topic score we used the TextRank algorithm to extract topic words from the documents. These scores were combined linearly to get one score for each sentence.

The main difference between the two systems is in the choice of sentences that are scored. For the first system we scored only the first sentences in each document. For the second system we scored all documents for a topic. Since we used all sentences in System 2 we also added a redundancy module to it. Sentences were dynamically scored for similarity to the sentences already selected for the partial summary and sentences with high similarity values were rejected.

4 Evaluation Results

In this section we discuss the TAC 2008 evaluation results. This is the first year we participated in the Document Understanding Conference update summarization task, and our systems are ranked in the middle tier of overall participating systems (shown in the Table 1).

After analyzing the evaluation results on each news summary, we found that System 1 performs better than System 2, which suggests that the first sentence in each document is quite useful for multi-document summarization.

Topics	Score	Rank
Linguistic Quality B	2.396	27 out of 58
Basic Elements B	0.044	42 out of 72

Table 1: Average Evaluation Results for System 1

Topics	Score	Rank
Linguistic Quality B	2.354	31 out of 58
Basic Elements	0.036	56 out of 72

Table 2: Average Evaluation Results for System 2

5 Conclusion and future work

In this paper we presented our on-going work on multi-document update summarization and our experience of participating in the TAC 2008. Ontology knowledge is proven to be an effective way to go beyond the mere keyword-based information retrieval methods. With our experiment, we feel that ontology knowledge can be further utilized in other fields of broad information management and knowledge discovery process. Our future work includes:

1. Experimenting with different WordNet distance functions.

2. Improving the redundancy module.

6 Acknowledgments

This work was supported in part by NSF grants CCF-0306475, DUE-0313880 and OCI-0453498. David Kent's work was performed while participating in the Summer 2008 REU program at the University of Houston under the guidance of the first author.

References

[1] Ping Chen and Rakesh M. Verma. A query-based medical information summarization system using ontology knowledge. In *CBMS: IEEE Symp. on Computer-based Medical Systems*, pages 37–42, 2006.

[2] Christiane Fellbaum, editor. *WordNet An Electronic Lexical Database*. MIT Press, 1998.

[3] M. Fiszman, T. Rindflesch, and H. Kilicoglu. Abstraction summarization for managing the biomedical research literature. In *HLT-NAACL 2004: computational lexical semantic workshop*, 2004.

[4] Udo Hahn and Inderjeet Mani. The challenges of automatic summarization. *IEEE Computer*, 33(11):29–36, 2000.

[5] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. In *EMNLP 2004: Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2004.

[6] NIST, editor. *DUC '05: Proceedings of the Document Understanding Conference*, 2005.

[7] NIST, editor. *DUC '06: Proceedings of the Document Understanding Conference*, 2006.

[8] NIST, editor. *DUC '07: Proceedings of the Document Understanding Conference*, 2007.