WEB PAGE SUMMARIZATION BY USING CONCEPT HIERARCHIES

Ben Choi and Xiaomei Huang
Computer Science, Louisiana Tech University, U.S.A
pro@benchoi.org

Keywords: Text summarization, Ontology, Word sense disambiguation, Natural language processing, Knowledge base, Web mining, Knowledge discovery.

Abstract: To address the problem of information overload and to make effective use of information contained on the Web, we created a summarization system that can abstract key concepts and can extract key sentences to summarize text documents including Web pages. Our proposed system is the first summarization system that uses a knowledge base to generate new abstract concepts to summarize documents. To generate abstract concepts, our system first maps words contained in a document to concepts contained in the knowledge base called ResearchCyc, which organized concepts into hierarchies forming an ontology in the domain of human consensus reality. Then, it increases the weights of the mapped concepts to determine the importance, and propagates the weights upward in the concept hierarchies, which provides a method for generalization. To extract key sentences, our system weights each sentence in the document based on the concept weights associated with the sentence, and extracts the sentences with some of the highest weights to summarize the document. Moreover, we created a word sense disambiguation method based on the concept hierarchies to select the most appropriate concepts. Test results show that our approach is viable and applicable for knowledge discovery and semantic Web.

1 INTRODUCTION

In this paper, we describe the development of a system to automatically summarize documents. To create a summary of a document is not an easy task for a person or for a machine. For us to be able to summarize a document requires that we can understand the contents of the document. To be able to understand a document requires the ability to process the natural language used in the document. It also requires the background knowledge of the subject matter and the commonsense knowledge of humanity. Despite the active research in Artificial Intelligence in the past half century, currently there is not machine that can understand the contents of a document and then summarizes the document based on its understanding.

Most past researches in automatic document summarization did not attempt to understand the contents of the documents, but instead used the knowledge of writing styles and document structures to find key sentences in the document that captured the main topics of the documents. For instances, knowing that many writers use topic sentences, the first sentence of a paragraph is considered as the key sentence that summarizes the contents of the paragraph. More examples are provided in the related research section.

Our research briefly described in this paper represents a small step toward the use of semantic contents of a document to summarize the document. There is a long way before we can try to use the word “understand” to describe the ability of a machine. Our research is recently made possible by the advance in natural language processing tools and the availability of large databases of human knowledge. For processing natural language, we chose Stanford parser (Maning & Jurafsky, 2008), which can partition an English sentence into words and their part-of-speech. To serve as the background knowledge of the subject matter and the commonsense knowledge of humanity, we chose ResearchCyc (Cycorp, 2008), which currently is the world's largest and most complete general knowledge base and commonsense reasoning engine.

With the help of the natural language processing tool and the largest knowledge base, our system is able to summarize text documents based on the semantic and conceptual contents. Since the natural language tool and the knowledge base only handle English, our system is currently only applicable to English text documents including texts retrieved from Web pages. When those tools are available for other languages, our proposed approaches should be able to be extended to process other languages as well. Since we use a knowledge base that organized concepts into hierarchies forming an ontology in the domain of human consensus reality, our system is one of the first ontology-based summarization system. Our system can (1) abstracts key concepts and (2) extracts key sentences to summarize documents.

Our system is the first summarization system that uses a knowledge base to generate new abstract concepts to summarize documents. To generate abstract concepts, we first extract words or phrases from a document and map them to ResearchCyc concepts and increase the weight of those concepts. In order to create generalized concepts, we propagate the weights of the concepts upward on the ResearchCyc concept hierarchy. Then, we extract those concepts with the highest weights to be the key concepts.

To extract key sentences from the documents, we weight each sentence in the document based on the concept weights associated with the sentence. Then, we extract the sentences with some of the highest weight to summarize the document.

One of the problems of mapping a word into concepts is that a word may have several meanings. To address this problem, we developed new ontology-based word sense disambiguation process, which makes use of the concept hierarchies to select the most appropriate concepts to associate with the words used in the sentences.

Test results show that our proposed system is able to abstract key concepts and able to generalize new concepts. In addition to summarization of documents, the abstracted concepts can be used for Semantic Web applications, information retrieval, and knowledge discovery system to tag documents with their key concepts and to retrieve documents based on concepts.

Test results also show that our proposed system is able to extract key sentences from text documents or texts retrieved from Web pages. The results produced by our system can directly be used for search engines, which can present the key sentences as part of the search results. We are working to expand our information classification (Choi & Yao, 2005; Yao & Choi 2007) and search engine project (Choi, 2006) to include the summarization results.

The rest of this paper is organized as follows. Section 2 describes the related research and provides the backgrounds. Section 3 describes our proposed process for abstracting key concepts. Section 4 outlines the process for extracting key sentences. Section 5 describes our proposed ontology-based word sense disambiguation. Section 6 describes the implement testing. And, Section 7 gives the conclusion and outlines the future research.

2 RELATED RESEARCH

Automatic document summarization is the creation of a condensed version of a document. The contents of the condensed version may be extraction from the original documents or may be newly generated abstract (Hahn & Mani, 2000). With a few exceptions, such as (Mittal & Witbrock, 2003) which uses statistical models to analyze Web pages and generate non-extractive summaries, most prior researches are extraction based, which analyze writing styles and document structures to find key words or key sentences from the documents. For instance, by assuming that the most important concepts are represented by the most frequently occurred words, the sentences with frequently occurred words are considered as key sentences. Knowing that the title conveys the content of the document and section headings convey the content of the section, sentences consisted of the title and section heading words are considered as key sentences (Teufel & Moens, 1997). Knowing that many writers use topic sentences, the first sentence of a paragraph is considered as the key sentence that summarizes the contents of the paragraph. Sentences that contain cue words or phrases, such as “in conclusion”, “significantly”, and “importantly”, are also considered as key sentences (Teufel & Moens, 1997; Kupiec et al., 1995).

Some researches (Doran & Stokes, 2004; Silber & McCoy, 2002) cluster sentences into groups based on hyponymy or synonymy, and then select a sentence as the key sentence to represent a group. Some researches classify sentences into nucleus and satellite according to rhetorical structure (Mann 1988). Nuclei are considered more important than satellite. Some analyze paragraph based on similarity and select the paragraph that has many other similar paragraphs (Salton et al., 1997).

Our research is made possible by the advance in natural language processing tools and the availability of large databases of human knowledge.

We chose Stanford parser (Manning & Jurafsky, 2008) as our natural language processing tool. It can partition an English sentence into words and their part-of-speech. We chose ResearchCyc (Cycorp, 2008) as our knowledge base and inference engine. ResearchCyc contains over 300,000 concepts and nearly 3,000,000 facts and rules. The concepts are organized into hierarchy forming an ontology, in which general concepts are provided on the upper nodes and specific concepts are provided on the lower nodes. The links between notes define the relations between concepts.

Some related researches used WordNet for text summarization (Barzilay & Elhadad, 1997) and for word sense disambiguation (Cañas et al., 2004; Simón-Cuevas1 et al., 2008). In our research, we take advantage of a powerful knowledge processing system: Cyc, which includes knowledge base, inference engine, representation language, and natural language processing. In fact, Cyc includes mappings from WordNet to Cyc concepts.

3 ABSTRACTING KEY CONCEPTS

In this section, we describe our proposed ontology-based process to generate new key concepts to summarize a document. The process is outlined in Figure 1. This process has three major parts: (A) it maps words from a document into Cyc concepts that are contained in ResearchCyc knowledge base. (B) It finds more general concepts by propagate weights of the concepts upward on the concept hierarchy of the Cyc ontology. And (C) it retrieves key concepts from Cyc to summarize the document.

The process to map words into Cyc concepts includes the following steps. (1) It takes each sentence of a document and parses the sentence, by using Stanford parser, to words and their part of speeches. (2) From the parsed results, it extracts words that are Noun (include single-word or multi-word noun), Verb, Adjective, and Adverb. (3) It maps each of the word (or word phrase) and the corresponding part of speech into Cyc concepts by using a Cyc language function called "denotation". And (4), it increases the weight of each of the mapped concepts by one when a word is mapped to the concept. We use a weight to associate with a concept to determine the importance of the concepts.

The process to propagate weights of the concepts upward includes the following steps. (1) It takes each of the non-zero weighted concepts and uses the Cyc function "min-genls" to find its nearest general concepts. (2) It scales the weight by a factor of δ and adds resulting weight to the weight of its nearest general concepts. This process is repeated recursively λ times to propagate the weights upward on the concept hierarchy. This process provides a method for generalization. Two factors are used to adjust the performance of the generalization. The λ factor controls how many levels to propagate the weight of a concept upward. The higher the number will result in the more abstract concepts to be generated. In our experiments, we found that setting λ to three produces results that are not too general. Setting λ higher will result in over generalization. The δ factor controls the reduction of the weight of a concept during the upward propagation. The higher the value of δ will result in fewer concepts are required to produce a general concept. To create a general concept, certain number of supporting concepts is needed to be presented. In our experiments, we found that setting δ to be 5% will prevent over generalization.

Part A: Map words into Cyc concepts

1. Parse each sentence of the document into words and their part of speech
2. Extract words that are Noun (single or multi-words), Verb, Adjective, and Adverb
3. Map a word and its part of speech to Cyc concepts
4. Increase the weight of each of the concept by one

Part B: Propagate weights of the concepts upward

1. Propagate non-zero weighted concepts λ levels upward to their upper concepts
2. Scale the weight by δ for each level upward

Part C: Retrieve key concepts from Cyc

1. Select some of the highest weighted concepts.

Figure 1: Generate key concepts of a document.

Part A: Find the total weight for each sentence in the document

1. For the words in the sentence that are noun, verb, adjective, and adverb, use word sense disambiguation to find the weight of the corresponding Cyc concept
2. Sum all the weights of the corresponding concepts
3. Normalize by dividing the total weight by the number of concepts

Part B: Select top sentences

1. Select some of the highest weighted sentences

Figure 2: Retrieve key sentences of a document.

The process to retrieve key concepts from Cyc is simply to select some (such as 5 to 10) highest weighed concepts from the Cyc knowledge base. Some of those retrieved concepts may be the results of the generalization. The retrieved concepts represent the key concepts of the document.

4 EXTRACTING KEY SENTENCES

In this section, we outline our proposed process to extract key sentences from a document based on concept weights. The process consists of two major parts as outlined in Figure 2. (A) It finds the total concept weight of each sentence in the document. (B) It selects top highest weighted sentences. This process is the preformed after process to generate key concepts from the document. Thus, words in the sentences have been mapped into concepts in Cyc knowledge base.

The process to find the total concept weight of each sentence consists of the following steps. (1) For each of the words in the sentence that is noun, verb, adjective, and adverb, it uses word sense disambiguation process (described in Section 5) to find the weight of the corresponding Cyc concept. (2) It sums all the weights of the corresponding concepts to get the total weight of the sentence. And then, (3) it divides the total weight by the number of concepts in the sentence to get the normalized weight of the sentence. The resultant normalized weight represents the concept weight of the sentence, which determines the importance of the sentence.

The process to select top sentences is simply to retrieve some (such as 3 to 10) of the highest concept-weighted sentences. These sentences represent the key sentences of the document.

5 ONTOLOGY-BASED WORD SENSE DISAMBIGUATION

One of the problems of mapping a word into concepts is that a word may have several meanings. To select the most appropriate concepts to associate with the words used in the sentences, we developed a new ontology-based word sense disambiguation process, which is outlined as shown in Figure 3.

The process performs the following steps for each word in a sentence that is a noun, verb,

Part A: For each word, perform the following disambiguation steps:

1. Find all the Cyc concepts associated with the word
2. Select the concepts with the highest weight
3. If there is only one selected concept, assign it as the concept of the word, done.

Part B: If there are more than one highest weighted concepts, then perform the following steps:

1. For each highest weighted concept, retrieve its upper concepts.
2. Select the upper concepts with the highest weight
3. If there is only one selected upper concept, assign the original concept as the concept of the word, done.

Part C: If there are more than one highest weighted upper concepts, then perform the following steps:

1. For each highest weighted upper concept, retrieve its conceptually related concepts
2. Sum the weights of all those conceptually related concepts
3. Choose one concept that has the highest weight, done.

Figure 3: Ontology-based word sense disambiguation.

adjective, or adverb. (1) It uses the Cyc function “denotation” to find all the Cyc concepts associated with word. (2) It selects the concepts that have the highest concepts weights. And (3), if there is only one selected concept, then it assigns that concept as the most appropriate concepts.

Since several concepts may have the same highest weight, the process performs the following steps if there is more than one concept having the highest weights. (1) For each of the concepts, it uses the Cyc function “min-genls” to find the nearest general upper concepts. (2) It selects those upper concepts having the highest weight. (3) If there is only one such concept, then it assigns the original concept (that is associated with the upper concept) as the most appropriate concepts. In this case, it used the upper concepts to break the tie of the original concepts.

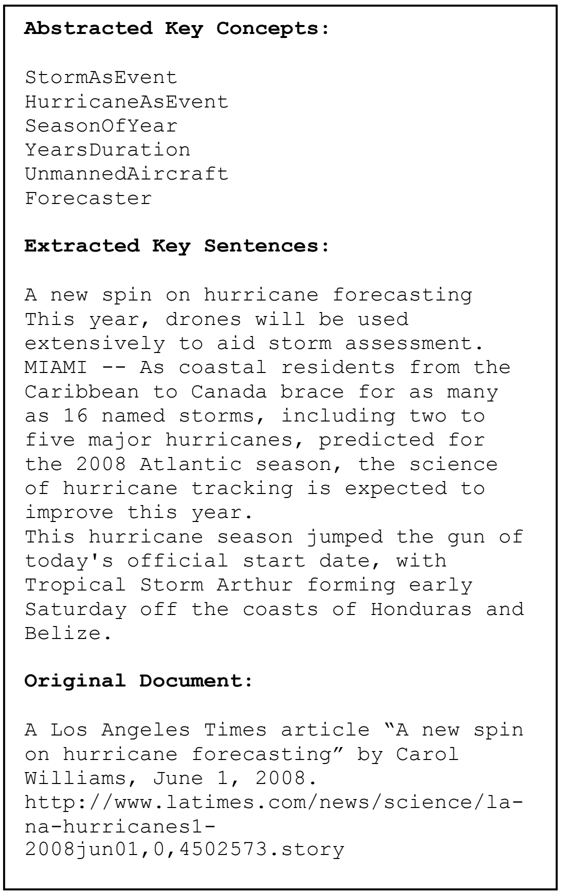
If the above steps still cannot break the tie, the process uses the “conceptuallyRelated” concepts to help break the tie by performing the following steps. (1) For each of the original concepts, it used the Cyc function “conceptuallyRelated” to find the conceptually related concepts. (2) It sums the weights of all those conceptually related concepts and uses the total weight to represent the weight of the original concept. (3) It chooses the original concept that has the highest associated weight. In this case, if there is still more than one concept having the highest weight, it arbitrarily chooses one concept and assigns that concept as the most appropriate concepts.

6 IMPLEMENTATION AND TESTING

Our proposed system has been implemented and tested. For the implementation, we use Java language to interface with Stanford parser and ResearchCyc. We also use a Cyc language called subL to define special functions to interface directly with ResearchCyc knowledge base. One of our subL functions defines a weight to be associated with a concept. Another subL function implements the process for propagating concept weight to the upper more general concepts. Our experience shows that using subL to define such functions is more effective than using Java API to interface with ResearchCyc.

We have tested our system using documents retrieved from varies sources, such as newspapers, Wikipedia, and technical papers. Figure 4 shows a sample of a test result, which provides a summary of a news article called “A new spin on hurricane forecasting”. Our system produced six abstract concepts, such as “StormAsEvent” and “HurricaneAsEvent”. These are the names of the concepts encoded in Cyc knowledge base. These abstract concepts provide insights to the content of the document. Our system produced three sentences as shown in Figure 4. In this case, the natural language processing tool, Stanford parser, attaches the title to the first sentence. At any rate, by reading through the original document, we found that the extracted sentences provide reasonable insights to the central subject of the document.

Evaluating the results of summarization is not an easy task and many researches have been done to provide standard database and evaluation metric (Lin, 2004; NIST 2008). Currently, we evaluate our system by comparing it with AutoSummary from Microsoft. Based on reading through the extracted sentences, we found that the extracted results from both systems are compatible. On the other hand, since our abstracted concepts are unique, no comparison is made on this function of our system.



```
Abstracted Key Concepts:

StormAsEvent
HurricaneAsEvent
SeasonOfYear
YearsDuration
UnmannedAircraft
Forecaster

Extracted Key Sentences:

A new spin on hurricane forecasting
This year, drones will be used
extensively to aid storm assessment.
MIAMI -- As coastal residents from the
Caribbean to Canada brace for as many
as 16 named storms, including two to
five major hurricanes, predicted for
the 2008 Atlantic season, the science
of hurricane tracking is expected to
improve this year.
This hurricane season jumped the gun of
today's official start date, with
Tropical Storm Arthur forming early
Saturday off the coasts of Honduras and
Belize.

Original Document:

A Los Angeles Times article “A new spin
on hurricane forecasting” by Carol
Williams, June 1, 2008.
http://www.latimes.com/news/science/la-
na-hurricanes1-
2008jun01,0,4502573.story
```

Figure 4: Test result of a summary of a news article.

7 CONCLUSIONS AND FUTURE RESEARCH

In this paper we proposed an ontology-base summarization system that can abstract key concepts and can extract key sentences to summarize text documents including Web pages. We introduced unique methods that have two advantages over existing methods. One advantage is the use of multi-level upward propagation to solve word sense disambiguation problem. The other is that the propagation process provides a method for the generalization of concepts. We have implemented and tested the proposed system. Our test results show that the system is able to abstract key concepts, generalize new concepts, and extract key sentences. In addition to summarization of documents, the system can be used for semantic Web, information retrieval, and knowledge discovery applications.

Based on our approaches, there are great potentials for future research. One challenging research is to create new abstract sentences to summarize a document. In this task, we are requiring computers to write meaningful sentences. This is not an easy task. We have been working on this task for years. Now, we are able to create simple sentences. We will report this work after more testing and fine-tuning. We are also working to incorporate automatic Web page summarization with Web page classification (Choi & Yao, 2005) and clustering (Yao & Choi, 2007) to create the next generation of search engine (Choi, 2006). Much research remains to be done to address the problem of information overload and to make effective use of information contained on the Web.

REFERENCES

Barzilay R. and Elhadad M, "Using lexical chains for text summarization," *Proceedings of the ACL Workshop on Intelligent Scalable Text Summarization*, pp. 10-17, 1997.

Cañas A. J., Valerio A, Lalinde-Pulido J., Carvalho M, & Arguedas M., "Using WordNet for Word Sense Disambiguation to Support Concept Map Construction," *Lecture Notes in Computer Science: String Processing and Information Retrieval*, Vol. 2857/2003, pp. 350-359, 2004.

Choi B. & Yao Z., "Web Page Classification," *Foundations and Advances in Data Mining*, Springer-Verag, pp. 221 - 274, 2005.

Choi B., "Method and Apparatus for Individualizing and Updating a Directory of Computer Files," United States Patent # 7,134,082, November 7, 2006.

Cycorp, ResearchCyc, http://research.cyc.com/, http://www.cyc.com/, 2008.

Doran W., Stokes N., Carthy J., & Dunnion J., "Comparing lexical chain-based summarisation approaches using an extrinsic evaluation," *In Global WordNet Conference (GWC)*, 2004.

Hahn U. & Mani I., "The Challenges of Automatic Summarization", IEEE Computer, Vol. 33, Issue 11, pp. 29-36, Nov. 2000.

Kupiec J., Pedersen J., & Chen F., "A Trainable Document Summarizer," *In Proceedings of the Eighteenth Annual International ACM Conference on Research and Development in Information Retrieval (SIGIR)*, 68–73. Seattle, WA, 1995.

Lin C.Y., "ROUGE: A Package for Automatic Evaluation of Summaries," *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, Barcelona, Spain, pp. 74-81, July, 2004.

Mann W.C. & Thompson S.A., "Rhetorical Structure Theory: Toward a Functional Theory of Text Organization," Text 8(3), 243–281. Also available as USC/Information Sciences Institute Research Report RR-87-190, 1988.

Manning C. & Jurafsky D., The Stanford Natural Language Processing Group, The Stanford Parser: A statistical parser, http://nlp.stanford.edu/software/lex-parser.shtml, 2008.

Mittal V.O. & Witbrock M. J., "Language Modeling Experiments in Non-Extractive Summarization," Chapter 10 in Croft, W. Bruce and Lafferty, John, *Language Modeling for Information Retrieval*, Kluwer Academic Publishers, 2003.

NIST, "Text Analysis Conference", http://www.nist.gov/tac/, National Institute of Standards and Technology, 2008.

Salton G., Singhal A., Mitra M., & Buckley C., "Automatic text structuring and summarization," *Information Processing and Management*, 33, 193-20, 1997.

Silber G. & McCoy K., "Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization," *Computational Linguistics*, 2002.

Simón-Cuevas1 A., Ceccaroni L., Rosete-Suárez A., Suárez-Rodríguez A., & Iglesia-Campos, M., "A concept sense disambiguation algorithm for concept maps," *Proc. of the Third Int. Conference on Concept Mapping*, Tallinn, Estonia & Helsinki, Finland 2008.

Teufel S. & Moens M., "Sentence Extraction as a Classification Task," *In Proceedings of the Workshop on Intelligent Scalable Summarization. ACL/EACL Conference*, 58–65. Madrid, Spain, 1997.

Yao Z. & Choi B., "Clustering Web Pages into Hierarchical Categories," *International Journal of Intelligent Information Technologies*, Vol. 3, No. 2, pp.17-35, April-June, 2007.