Proceedings of the 10th IASTED International Conference
Artifical Intelligence and Applications (AIA 2010)
February 15 - 17, 2010 Innsbruck, Austria

CREATING NEW SENTENCES TO SUMMARIZE DOCUMENTS

Ben Choi & Xiaomei Huang

Computer Science, Louisiana Tech University, USA
pro@BenChoi.org

ABSTRACT

This paper describes the first summarization system that is able to create new sentences to summarize documents. Creating new sentences to summarize documents is a challenging research and no prior research is able to do so. Most prior researches are extraction based that analyze writing styles and document structures to find key words or key sentences from documents and use those words or sentences as summaries. In this paper, we propose a new method to generate new simple sentences based on the main concepts contained in the documents. Our system starts with creating simple sentences that consists of subject, predicate, and object. It first simplifies each sentence of a document to the format of subject, predicate, and object, when possible. Then, it clusters the sentences into compatible classes that have similar concepts. It then creates a sentence for each of some of the largest compatible classes. Those created sentences serve as the summary of the document. The assumption used here is that the central ideas of a document are those with many supporting concepts. However, this approach does not yet capture the temporal and causal relations between sentences. The system has been implemented and tested. Test results show that our approach is viable for future research and applicable for knowledge discovery and semantic Web.

KEY WORDS

Document summarization, knowledge discovery, natural language processing, data mining, knowledge base, Web mining.

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. 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 instance, 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 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 [1], 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 [2], 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 ontology in the domain of human consensus reality, our system is one of the first ontology-based summarization system.

Our system can (1) abstract key concepts, (2) perform word sense disambiguation, (3) extract key sentences, and (4) create new sentences to summarize documents. This paper focuses on abstracting key concepts and creating new sentences, while extracting key sentences and word sense disambiguation are reported on Choi and Huang [3].

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.

One of the problems of mapping a word into concepts is that a word may have several meanings. To address this problem, we developed a new ontology-based word sense disambiguation process [3], which makes use of the concept hierarchies to select the most appropriate concepts to associate with the words used in the sentences.

Based on the mapped and the generalized concepts, our system attempts to create new sentences to describe the document. It starts with creating simple sentences that consists of subject, predicate, and object. It first simplifies each sentence of a document to the format of subject, predicate, and object, when possible. Then, it clusters the sentences into compatible classes that have similar concepts. It then creates a sentence for each of some of the largest compatible classes. Those created sentences serve as the summary of the document. The assumption used here is that the central ideas of a document are those with many supporting concepts.

Our proposed system has been implemented and tested. Test results show that our proposed system is able to abstract key concepts and able to generalize new concepts. The system is also able to create general sentences to describe concepts contained in the test documents.

In addition to summarization of documents, the abstracted concepts and the created new sentences 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. The results produced by our system can directly be used for search engines, which can present the abstracted concepts and the created new sentences as part of the search results. We are working to expand our information classification [4][5] and search engine project [6] 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 creating new sentences. Section 5 describes the implementation and testing. And, Section 6 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 [7]. With a few exceptions, such as Mittal and Witbrock [8], 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 [9]. 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 [9][10].

Some researches [11][12] 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 [13]. Nuclei are considered more important than satellite. Some analyze paragraph based on similarity and select the paragraph that has many other similar paragraphs [14].

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 [1] as our natural language processing tool. It can partition an English sentence into words and their part-of-speech. We chose ResearchCyc [2] 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 [15] and for word sense disambiguation [16][17]. 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.

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 Creating New Sentences

In this section, we describe our proposed semantic-based process to create new sentences. The proposed process is outlined in Figure 2. This process has four major parts as detailed below.

(A) It simplifies each sentence in the given document into subject, predicate, and object, when possible. To do that, it uses a syntactic parser to parse each sentence into words and word phases and their part-of-speech. From the parsed results, it extracts the subject, predicate, and object to form a simplified sentence.

(B) It clusters the simplified sentences into compatible classes. To do that, it creates an n by n compatible matrix given n simplified sentences. Two simplified sentences are considered to be compatible, if each pairs of the three fields (subject, predicate, and object) are compatible. Two fields are considered to be compatible, if they have the same name, or if they have the same parent in the concept hierarchy, or if they are conceptually related (which is determined by using Cyc inference engine). Then, it uses clustering methods to create larger compatible classes of size more than two members by insuring all the members in the class are pair-wise compatible.

(C) It creates a sentence for each of some of the largest compatible classes. To do that, it selects some of the largest compatible classes to represent the key topics of the document. For each of the selected compatible classes, it creates a sentence to represent the class. To create a new sentence, each field (subject, predicate, and object) of the new sentence is created by checking the corresponding field of all members in the class. If the fields of some of the members have the same parent in the concept hierarchy, used the name of the parent as the filed for the new sentence, otherwise arbitrary uses the name of one of the member as the field for the new sentence.

The generated new sentence may be quantified in the following way. The proposed process quantifies a new sentence by "Some" if some of the subjects have the same parent but they parent has some children that are not presented. It quantifies a new sentence by "All" if the parent has all the children presented in the subject field. If these two cases are not met, no quantifier will be used for the newly created sentence.

(D) Those newly created sentences will serve as summary for the document.

Some additional details are required to clarify the proposed process for creating new sentence as shown in Figure 2. For step (B.3), word sense disambiguation method is used to map word into the most appropriate concept, since a word may be associated with several concepts. We proposed an ontology-based word sense disambiguation process, which is reported on our previous paper [3]. For step (B.4), we use the clustering algorithm called "Finding Maximal Compatibility Classes" [18]. For step (C.1), three to ten classes or more may be selected depending on the number of sentences in the document and user preference. For step (C.4), using an upper (the parent) concept to represent instances of the members (children) concepts is a method of generalization. The difficult question is how many members or what percentage of presented members is sufficient to justify the generalization, to consider as a whole. Too small the percentage will result in over generalization, while too large the percentage may not create condensed summary. This problem turns out to be the same problem in machine learning [18][20]. This problem is partially addressed by using quantifier. For step (C.5), by using quantifier "Some" and "All" to quantify subject of a newly generalized sentence addresses the problem of over generalization.

Process to create new sentences to summarize a document:

(A) Simplify each sentence in the given document to (Subject, Predicate, Object)
(B) Clustering the simplified sentences into compatible classes
(C) Create a sentence for each of some of the largest compatible classes
(D) Use the created sentences as summary for the document

(A) Simplify each sentence in the given document to (Subject, Predicate, Object)
(1) Use syntactic parser to parse a sentence into words and word phases and their part-of-speech
(2) From the parsed results, extract the subject, predicate, and object to form a simplified sentence

(B) Clustering the simplified sentences into compatible classes
(1) Given n simplified sentences, create an n x n compatible matrix
(2) Two simplified sentences are compatible, if each of the three fields (Subject, Predicate, Object) are compatible
(3) Two fields are compatible, if they have the same name or if they have the same parent in the concept hierarchy or if they are conceptually related
(4) Create larger compatible classes of size more than two members by insuring all the members in the class are pair-wise compatible

(C) Create a sentence for each of some of the largest compatible classes
(1) Select some of the largest compatible classes to represent the key topics of the document
(2) For each of the selected compatible classes, create a sentence to represent that class
(3) Each field (Subject, Predicate, Object) of a new sentence is created by checking the corresponding field of all members in the class
(4) If the fields of some of the members have the same parent in the concept hierarchy, use the name of the parent as the field for the new sentence, otherwise arbitrary uses the name of one of the member as the field for the new sentence
(5) Quantify new sentence by "Some" if some of the subjects have the same parent but the parent has some children that are not presented, by "All" if the parent has all the children presented in the subject field, otherwise no quantifier is used.

(D) Use the created sentences as summary for the document
(1) Use the sentences created in the last stage as the summary for the document

Figure 2. Process for Creating New Sentences to Summarize Documents [19].

5 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 3 shows a sample of part of a test result, which provides a list of abstracted key concepts from 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 some insights to the content of the document.

Figure 4 shows test results of some of the newly generated sentences from several documents. Each one of the documents is summarized independently. One of the documents, for example, is taken from Wikipedia.org on the subject “Dog”. Two of documents, for example, are taken from scientific articles about tree tomato and about grapefruit. The first sentence in the figure, “Tigers eat mammal meat.”, for example, shows that the system is able to create an abstract concept “mammal meat” to describe various meats that tigers eat. The sentence, “Some Canis consume fruit.”, for example, shows that the system is able to use the quantifier “Some” to limit the scope of the generated sentence. It also shows that the system is able to use an abstract concept “Canis” to describe various species, including dogs, wolves, and coyotes. The sentence “Fruit be round.”, for example, shows that the system uses the word “be” to denotes the verb-to-be concept. It also shows that the system did not use the quantifier “All” in the sentence, since not all fruits are round. The sentence, “Some external anatomical parts be things.”, shows that the system produces a concept “things” that may be too general. More testing details and results are provided in [21]. Although improvements needed to be made, these test results show that the system is able to create general sentences to describe concepts contained in the test documents.

Abstracted Key Concepts:

StormAsEvent
HurricaneAsEvent
SeasonOfYear
YearsDuration
UnmannedAircraft
Forecaster

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 results of some key concepts generated from a news article.

Newly Generated Sentences:

“Tigers eat mammal meat.”
“Dogs be carnivore.”
“Some Canis consume fruit.”
“Fruit be round.”
“Some external anatomical parts be things.”

Figure 3. Test results of some of the newly generated sentences from several documents.

Currently, we have performed qualitative evaluations by manually reading the generated concepts and sentences to determine whether they capture central ideas of the documents. We found most of them do but some may be too general. Evaluating the results of summarization is not an easy task and many researches have been done to provide standard database and evaluation metrics [22][23]. However, the evaluation metrics are more applicable for extraction based summarization system and not directly applicable for our system. Since the abstracted concepts and created new sentences are unique, no comparison is made on these functions of our system. The functions are parts of a larger system, which also able to extract key sentences, for which we have compared our results with others [3].

6 Conclusion and Future Research

In this paper we proposed an ontology-base summarization system that can abstract key concepts and can create new sentences to summarize text documents including Web pages. We introduced unique methods. Our concept propagation process provides a method for the generalization of concepts. Our process to create new sentences is the first of its kind.

We have implemented and tested the proposed system. Test results show that the system is able to abstract key concepts, and although improvements needed to be made, the test results also show that the system is able to create general sentences to describe concepts contained in the test documents. However, some of the generated concepts and sentences may be too general and they do not capture the temporal and causal relations between sentences.

Based on our approaches, there are great potentials for future research. One challenging research is to improve the process to create new sentences by creating more complex sentences and by allowing sentences to link to each other conceptually. In this task, we are requiring computers to write meaningful and connected sentences. This is not an easy task. The process described in this paper to create simple sentences represents a small step toward this goal. We are working toward this goal and also working to incorporate automatic Web page summarization with Web page classification [4] and clustering [5] to form the next generation of search engine [6]. 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

[1] 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.

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

[3] Choi B. & Huang X., "Web Page Summarization by Using Concept Hierarchies," *International Conference on Agents and Artificial Intelligence* (ICAART), pp.281-286, Proto, Portugal, January, 2009.

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

[5] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] Choi B., "Inductive Inference by Using Information Compression," *Computational Intelligence*, 19 (2), 164-185, 2003.

[19] Choi B., "Method and Apparatus for Creating New Sentences to Summarize Documents," Report of Invention 2008-21, Louisiana Tech University, 2008.

[20] Choi B., "Automata for Learning Sequential Tasks," *New Generation Computing: Computing Paradigms and Computational Intelligence*, Vol. 16 No. 1, pp. 23-54, 1998.

[21] Huang, X., "Text Summarization Using Concept Hierarchy," Dissertation, Louisiana Tech University, May, 2009.

[22] 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.

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