Ontology-based Text Summarization for Business News Articles

Chia-Wei Wu and Chao-Lin Liu
Department of Computer Science
National Chengchi University
Taipei 11605, Taiwan
{g9010,chaolin}@cs.nccu.edu.tw

Abstract[†]

In this paper, we compare two methods for article summarization. The first method is mainly based on term-frequency, while the second method is based on ontology. We build an ontology database for analyzing the main topics of the article. After identifying the main topics and determining their relative significance, we rank the paragraphs based on the relevance between main topics and each individual paragraph. Depending on the ranks, we choose desired proportion of paragraphs as summary. Experimental results indicate that both methods offer similar accuracy in their selections of the paragraphs.

Keywords：Information Retrieval, Natural Language Processing, Ontology Application, Text Summarization

1. Introduction

In the past decade, the explosively growing number of online articles has made efficient information gathering a challenging necessity. Instead of requiring readers to go through all articles, providing summaries of articles is one way to save people time. However, not every article includes a summary, due to the high costs of summarizing articles with human power. Therefore, there is an increasing need to build an automatic summarization system.

The literature has seen two main approaches to the summarization task: linguistic approaches, statistical approaches, and their combinations [6]. There are also two closely related formats of “summary”: Abstraction and summarization. Abstraction is the process of understanding, interpreting, and paraphrasing a new brief for the original article [11]. Unfortunately, it is still very difficult for computers to do such a human task. Therefore, many researchers turn to seek a more viable way for article summarization: Ranking and showing the most relevant, original sentences in the given article. The main research issue of this summarization task is then determining the criteria for ranking the relevance of each sentence.

In this paper, we study and implement an ontology-based method for the ranking task. For comparison, we also implement another system that is based on the term-frequency techniques. We collect articles from the New York Times and the Wall Street Journal, and obtain summaries of the articles from the ProQuest database. Performance evaluation of the summarization methods is conducted by comparing outputs of the algorithms with summaries specified in the ProQuest database. We compare the implemented systems by standard performance measures, including recall, precision and F-measure.

Section 2 reviews some previously proposed summarization techniques. Section 3 briefly introduces ontology. Section 4 explains our data source. Section 5 describes details of our approaches. Section 6 present the results of experiments designed to evaluate the performance of the two methods, and discussion.

2. Related Work

Meown and Radev propose an approach for abstraction. They build a natural language processing system for analyzing text documents. They design many templates for the structure of abstracting [9].

Kupiec et al. developed a summarizer that selects sentences using such features as average-term frequency, title words, and sentence locations. Their approach performs well despite its simplicity [11], and their method’s performance becomes the baseline in this field [4]. The drawback of their approach is not using structural information in the text. Allan et al. generate summary of news by adding “time” factor into their summarization system to tract news and generate summary [1].

Many other summarization systems consider the users’ queries, the contents of articles, and their relevance to generate summary. See [8] for a recent survey of this field.

3. Ontology

Applications of ontology-related techniques have become increasingly popular in recent years [3]. Nevertheless, there is no unique definition of ontology in literature yet. We use Gruber’s definition of ontology [7]: “An ontology is an explicit specification of some topics. It is a formal and declarative representation, which includes the vocabulary (or names) for referring to the terms in a specific subject area and the logical statements that describe what terms are, how they are related to each other.”

Essentially, the ontology decomposes the world into several objects for describing them. The determination of the way we describe objects and the formalism of representation depend on individual applications. In this paper, the ontology is designed for analyzing and gath-

[†] This paper appeared in the *Proceedings of the ISCA Eighteenth International Conference on Computers and Their Applications*, 389-392. Honolulu, Hawaii, USA, 26-28 March 2003.

ering the semantic information of a class of article. Assuming every article contains several subtopics, we use the ontology for identifying subtopics of articles, and encode each of these possible subtopics by a non-overlapping portion of the ontology.

4. Data Source

We collect 51 sample articles from the ProQuest database by entering the search keyword “SONY.” These articles appeared in the New York Times and the Wall Street Journal, and contained 882 paragraphs in total. Usually, one paragraph has one or two sentences. The ProQuest database also provides summaries for articles. Among the 882 paragraphs, 133 paragraphs were selected as summarizing paragraphs. Because articles have already been separated into several paragraphs in the ProQuest database, we do not have to repeat this task.

5. Methods

Basically, summarization system will give each paragraph a relevance score and rank them by the scores. Higher scores imply that the paragraphs are more possible to be selected into the summary. In the end, we extract a desired portion of paragraphs as the summary.

5.1 Method 1：Non-ontology-based

The first method employs several features proposed in previous work on summarization in [4], [6], and [11]. We implement four features in our system.

I. Term-frequency: Count the frequency of each word, and the select most frequent N words for scoring. Next, each paragraph is scored based on the appearance of these N words. We set N to 5 in our experiments.

II. Sentence length: Given a threshold for all paragraphs, we ignore the paragraphs that do not have sufficient number of words.

III. Bonus words: If one paragraph contains bonus words, then the probability of the paragraph being chosen into the summary is higher too. We use fifteen articles as the training corpus to select the bonus words.

IV. Proper nouns : The significance of a paragraph is related to the number of occurrence of proper nouns. For counting proper nouns, we simply count the number of words with leading upper-case letters in each paragraph.

After getting values of these features, we score each paragraph with the following formula.

$$G_j = L(w_1 f_{j1} + w_2 f_{j2} + \cdots + w_n f_{jn})$$

G_j is the grades of the jth paragraph; f_{ji} is the value of the ith feature of the jth paragraph; w_i is the weight of ith feature. L is 1 if the paragraph has sufficient number of words, otherwise 0.

5.2 Method 2：Ontology-based

Using semantic information encoded in the ontology, our system determines which topics are useful for extracting paragraphs. Designing and constructing the ontology are the first two steps for building the summarization system.

Until now there is no standard for designing and constructing the ontology. After consulting methods in [10], [12],and [13] we follow the following steps for this task.

I. Define the purpose of using ontology: The purpose of using ontology is to analyze articles and acquire semantic information.

II. Determine the domain of ontology: In this paper, the domain is about the SONY corporation, which includes productions, related financial information, competitors, and so on [14].

III. Construct the ontology: First, we collect vocabularies and synonyms. Next, we put those words by the Data model of ontology. The following figure shows part of our ontology. Ontology includes 142 words.

The first step of our method is to determine the main subtopics of the article of interest. This is achieved by comparing the words of articles with terms in the ontology. If the word does not exist in the ontology, we ignore it. Otherwise, we record the number of times the word appears in the ontology.

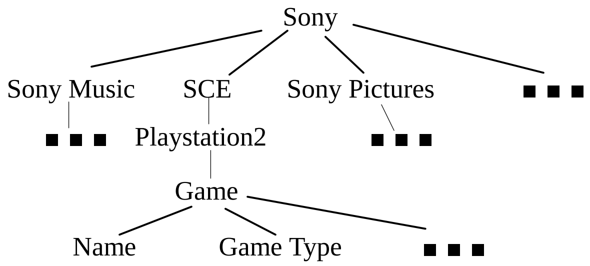

Figure 1: Top levels of our ontology

We encode the ontology with a tree structure, and each node includes the concepts represented by the node’s children. When the count of any node increases, the counts associated with their ancestors will also increase. We use this principle to score paragraphs. For example, ”Spider-man” is a child node of node “movie.” If one paragraph contains a word ”Spider-man”, then counts of both “Spider-man” and “movie” are simultaneously increased. By this design, the root of the ontology will always get the highest grade, while nodes in the second level, which represent subtopics, will get different scores. After marking the counts of the nodes in the ontology, we select second-level nodes that have higher counts as the main subtopics of the article. Generally speaking, one article is composed of several subtopics, so our system will select multiple subtopics.

There are limited topics an article can contain, and a reasonable summary probably should include fewer. Therefore, we only choose a limited number of subtopics and ignore others. We choose to ignore the subtopic if its count is less than 10. In addition, we choose only top three subtopics.

After obtaining the subtopics, our system will use them for selecting paragraphs as the summary. We rank the paragraphs based on their " closeness" to the selected subtopics. The selection procedure follows.

1. Compute relevance between paragraphs and the selected subtopics. We compare the words of each selected subtopic with words in each paragraph, and associate with each paragraph the counts of common words that appear in the paragraph and the selected subtopics. Assuming there are *n* selected subtopics, there will also be *n* scores associated with each paragraph, and these *n* scores represent the relevance of the paragraph with each selected subtopic.

2. Compute the score for each paragraph. The score of each paragraph is the sum of its weighted relevance with subtopics. The weights are determined dynamically based on counts that we used to selected main subtopics. The weight of each topic has an intuitive explanation: Primary topic is more representative than other topics, so the weight should be higher than others.

We give a simple example of the process. Following is a segment of news quoted from the *Wall Street Journal July* 26, 2002, and the title is "*Movie Helps Sony Post Profit.*"

> *Sony Chief **Financial** Officer Teruhisa Tokunaka said box-office receipts of the film "**Spider-Man**" have reached $675 million, making it the fifth-largest-**grossing movie** ever (unadjusted for inflation) and boosting sales at Sony's **movie** business to 173.6 billion yen, a 28% increase from a year earlier. Mr. Tokunaka said operating **profit** at Sony's **electronics business**, which accounts for 70% of the company's annual sales, rose to 49 billion yen from 1.5 billion yen a year earlier.*

Assume that we got three topics at the first step: *movie* with 20 counts, *electronic business* with 10 counts, and *financial* with 15 counts, after processing the whole article (not just the segment above). A paragraph will get 20 points for each movie-related word that is contained in its body at the second step. Similarly, when a word related with electronic business appears, the paragraph will get 15 points for the topic of electronic business. The paragraph's scores are the weighted sum of these counts.

3. Rank paragraphs, and select a desired proportion of the paragraphs as the summary.

In summary, we use the following formula for scoring paragraphs:

$$P_j = w_1 o_{j1} + w_2 o_{j2} + \cdots + w_n o_{jn},$$

where P_j is the score of the j^{th} paragraph, o_{ji} is the score of the i^{th} topic of P_j, w_i is the weight of the i^{th} topic.

6. Experimental Results

6.1 Experiments

In our experiments, we input articles to our summarizing systems. Next, we compare their outputs with the summaries specified in the ProQuest database. We evaluate the results by precision, recall, and the F-measure. For obtaining the baseline performance, we randomly select paragraphs as the summary. We observe the quality of our summarization by letting them choose 1, 2, … , 10 top paragraphs in separate experiments. These selected paragraphs are then compared with the summaries provided by the ProQuest database to compute the performance measures. In these experiments, we use 51 different articles.

The performance measures are defined below [6].

$$\text{Precision}=\frac{J}{K}$$

$$\text{Recall}=\frac{J}{\min(M,K)}$$

$$\text{F}=\frac{2}{\dfrac{1}{precision}+\dfrac{1}{recall}}$$

Let J be the number of paragraphs that are selected correctly. K is the total number of selected paragraphs. M is the number of paragraphs in the summary specified in the ProQuest database. J, M, and K are the sum of results of each article's experiment. In order to account for the fact that a compressed summary does not have the opportunity to return the full set of relevant sentences, we use a normalized version of recall and a normalized version of F-measure [6].

6.2 Results

Table 1 shows the results of the ontology-based method. Table 2 shows the results of the random-selection method. The performance of ontology-based method is much better than random-selection method. This result demonstrates that ontology-based method could catch some useful information for the summarization task.

Table 1: Ontology-based method

	10	9	8	7	6	5	4	3	2	1
Precision	0.24	0.27	0.29	0.32	0.36	0.42	0.46	0.52	0.54	0.70
Recall	0.94	0.93	0.91	0.87	0.84	0.81	0.70	0.60	0.54	0.70

10~1is the number of paragraphs that are selected

Table 2:Random-selection method

	10	9	8	7	6	5	4	3	2	1
Precision	0.13	0.14	0.14	0.13	0.14	0.16	0.18	0.17	0.16	0.19
Recall	0.22	0.21	0.21	0.19	0.20	0.21	0.21	0.18	0.16	0.19

Table 3 compares performance of all three implemented methods using the F-measure.

Table 3: F-measure

	10	9	8	7	6	5	4	3	2	1
Ontology	0.39	0.41	0.45	0.47	0.51	0.56	0.55	0.55	0.54	0.70
Non-ontology	0.38	0.41	0.44	0.46	0.49	0.53	0.55	0.55	0.51	0.52
Random	0.22	0.22	0.21	0.19	0.20	0.21	0.21	0.19	0.17	0.19

Although the first and the second methods are based on different concepts, the experimental results indicate that both methods offer similar accuracy in their selections of the paragraphs, indicating that both methods catch part of the key to the summarization task. We believe that both methods could obtain some features useful for summarization. Term-frequency methods are able to identify frequent content words that are not included in the ontology. If the ontology does not include the content words at the design time, the ontology-based method will not work.

On the other hand, ontology could find the subtopics more precisely even when major content words do not appear many times in the paragraphs. Ontology-based method identifies their occurrences without relying on frequencies. Ontology-based method could also perform well as we need high compression ratio rate of summary, which is demonstrated by better performance when the system selects only one paragraph. This is because the most relevant paragraph in our sampled articles typically contains words that are directly related to the main topic. These keywords are easily identified by the ontology-based method.

7. Discussion

The experimental results demonstrate the value of ontology-based method for summarization. Unfortunately, designing, constructing, and maintaining the ontology even only in one specific domain is still costly. If we can construct ontology fully automatically or at least partially automated, we can than apply ontology to realistic applications. In fact, we are applying latent semantics[5] analysis for assisting construction of ontology.

There is an obvious drawback of our ontology-based method, however. If the ontology does not include the right words at design time, our current method would collapse completely.[*] This occurs even if synonyms of the words used in the ontology appear in the articles. We could improve performance of our system by including a synonym dictionary.

In the future work, we plan to design a method that can determine how many topics should be chosen for summary automatically. Finally, since the strength of the term-frequency based and ontology based methods compensate with each other, we believe that combining the ontology-based and non-ontology-based systems will be positive for the quality of summary.

Acknowledgements

This research was support in part by Grant NSC-91-2213-E-004-013 from the National Science Council of Taiwan.

References

[1] J. Allan, R. Gupta, and V. Khandelwal, Temporal Summaries of News Topics, *Proc. of the 24th annual Int'l ACM SIGIR Conf. on Research and development in information retrieval, 10-18*, 2001.

[2] J. Goldstein, V. Mittal, J. Carbonell, and J. Callan, Evaluating Multi-Document Sentence Extract Summaries, *Proc. of the 9th Int'l Conf. on Information and Knowledge Management*, 165-172, 2000

[3] B. Chandrasekaran, J. R. Josephson, and V. R. Benjamins, What Are Ontologies, and Why Do We Need Them?, *IEEE Intelligent Systems*, 14(1), 20-26, 1999.

[4] W. T. Chuang and J. Yang, Extracting Sentence Segment for Text Summarization：A Machine Learning Approach, *ACM SIGIR 2000*, 152-159,2000.

[5] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, Indexing by Latent Semantic Analysis, *J. of the American Society of Information Science*, 41, 391-407, 1990.

[6] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell, Summarizing Text Documents: Sentence Selection and Evaluation Metrics, *ACM SIGIR 1999*, 121-128, 1999.

[7] T. Gruber, Ontology Definition, www-ksl.Stanford.edu/kst/what-is-an-ontology.htm

[8] U. Hahn, I. Mani. The challenges of automatic summarization, *IEEE Computer*, 33(11), 29–35, 2000.

[9] K.n Meown and D. R.Radev, Generating Summaries of Multiple News Articles, *ACM SIGIR* 1995, 74-82, 1995.

[10] D. D. McDonald, The View from the Trenches: Issues in the Ontology of Restricted Domains, *Proc. of the Int'l conf. on Formal Ontology in Information Systems 2001*, 22-33, 2001.

[11] J. Kupiec, J. Pederseon, and F. Chen, A Trainable Document Summarizer. *ACM SIGIR 95*, 68-73, 1995.

[12] J. F. Sowa. Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks/Cole Pub Co, 1999.

[13] H. Snoussi, L. Magnin, and J.-Y. Nie, Heterogeneous Web Data Extraction using Ontology, *Proc. of the 3rd Int'l Bi-Conf. Workshop Agent-Oriented Information System*, 99-110 2001.

[14] Data source:http://www.sony.com

[*] In our experiments we only filter out very short articles that contained only few sentences. We did not exclude articles that use words not in our ontology.