論文の内容の要旨

論文題目

RESEARCH ON INFORMATION AGGREGATION AND INTEGRATION FOR MULTI-DOCUMENT SUMMARIZATION

（複数文書自動要約における情報の集約と統合に関する研究）

37406　岡崎　直観

Numerous computerized documents are accessible on-line. In November 2006, Netcraft Web Server Survey reported that more than 100 million web sites with domain names were found on the Internet. The Internet has doubled in size since May 2004, when the survey hit 50 million. The grand total number of pages on the Web is estimated to be much larger than this figure. These facts suggest *information explosion*, tremendous increase of the amount of published documents. Meanwhile, search engines on the Web achieve a moderate success in keeping up with the growth of computerized documents. Major search engines (e.g. Google and Yahoo!) claim to have indexed more than several billion pages on the Web. Using a search engine on the Web becomes a pervading idea to obtain information at a reasonable cost and time.

Notwithstanding, we are often disappointed with the quantity of retrieved documents despite having narrowed the range of documents of interest. For instance, Google retrieves as many as 10,300 documents with a query “hijacking all nippon airways 61” (as of November 2006); and 144 documents are found, with the same query, in the Mainichi newspaper articles published in 1999. The situation having too much information to utilize is called information overload, and has been regarded a major problem

The information explosion also results in a big shift of structuring information. The more information retrieval systems play an important role in information society, the more often we deal with documents gathered dynamically without inter-document structures, e.g., search results. Unlike human-made document collections such as books and journals, we must sort out the structure of retrieved documents, i.e., what is common in the document set, what is the different part of a document from others, what is the optimal order to read through the documents, etc. In addition, more and more

information is being published in unstructured styles. We cannot expect a coherent story in a series of blog entries as a blogger might ramble about various subjects such as products, news events, or local events. Thus, a mechanism to aggregate information published in different documents is key to the solution of the information overload problem.

Automatic text summarization is a challenge to the information overload problem, allowing users to control the amount of text to be read. The goal of automatic summarization is, given an information source, to present the most important content to the user in a condensed form and in a manner sensitive to the user's need. Multi-Document Summarization (MDS), which is a summarization task specialized in dealing with related documents (e.g. a collection of news stories on the same topic), has attracted much attention in recent years. In addition to the conventional process of information retrieval, a user forwards retrieved documents to the MDS system and obtain a summary. If the summary could satisfy user's information demand directly, the user would save time and effort to read the retrieved documents. The summary could also help the user determine whether if they need an intensive reading for some of the documents.

Each component in an MDS system also presents research challenges in text mining. This thesis addresses methodologies for aggregating information and knowledge across documents, focusing on three research topics essential to an MDS system: *sentence extraction*, *sentence ordering*, and *acronym recognition*. This thesis consists of seven chapters. The first chapter addresses the background, motivation, and goal of this study. The subsequent chapter (Chapter 2) provides a review for automatic text summarization. Chapter 3 presents the task definition and evaluation methodology of the 3rd Text Summarization Challenge (TSC), as this study makes use of its evaluation corpus. Chapter 4 describes a method for sentence extraction in an MDS system, which is formalized as a combinational optimization problem that determines a set of sentences containing as much important information as possible. Chapter 5 addresses two approaches to text structuring for extracts from multi-documents: a novel method to refine the conventional method for arranging sentences; and a machine learning approach to aggregate the multiple criteria for further improvement. Chapter 6 presents a methodology for building an abbreviation dictionary from a large corpus. Chapter 7 remarks future directions of this work and concludes the thesis.

Chapter 4 of this thesis presents a methodology for sentence extraction. Passage extraction is the most basic technology for building an automatic summarization system.

Excluding a few exceptions, most summarization systems employ some kind of extractive techniques. Passage extraction finds important passages in source documents to produce a summary. In general, the extraction problem is formalized as the computer-friendly task of assigning relevance scores to textual passages in source documents. Therefore, passage extraction is obviously the low-cost but main solution to the summarization research.

The current NLP technology cannot deal with the meaning of a text as flexibly and efficiently as humans do. This study assumes that: a human reader breaks a sentence into a set of *information fragments* to which the sentence is referring; an information fragment is independent from each other; and an information fragment has its importance score. Thus, a sentence is approximated by a set of information fragments each of which conveys atomic information in a sentence. Among various sentence representations such as bag-of-words, bi-gram, tri-gram, n-gram, FrameNet, this study proposes the use of the dependency relations of terms in a sentence. The advantage of this representation is that dependency relations refer to what the original sentence is saying (with a certain degree of human interpretation), and therefore are useful to keep track of information conveyed by the sentences.

Based on the sentence representation, the problem of sentence extraction is formalized as a combinational optimization problem that determines a set of sentences containing as much important information fragments as possible under the constraint of the summarization ratio. The presented algorithm chooses sentences in source documents one by one. Because source documents often contain redundant information, the algorithm reduces the importance of information fragments that has been included in the sentences chosen previously. The implemented system employs a beam search to find a quasi-optimal solution in a reduced search cost as the extraction problem belongs to the NP-hard class.

The presented system achieved a good result on TSC-3 evaluation corpus. For both high and low summarization ratios, the system took the 3rd place among the systems participated in TSC-3. The result was far better than that of a baseline system, which features the lead extraction strategy. The comparison among sentence representation demonstrated that the proposed representation using pair-wise dependency relations performed better than bag-of-words and co-occurrence representations.

Chapter 5 examines a method to arrange sentences that are extracted by important sentence extraction. A summary with improperly ordered sentences confuses a reader and degrades the quality/reliability of the summary itself. However, ordering a set of

sentences for a coherent text is a non-trivial task. For example, identifying rhetorical relations in an ordered text has been a difficult task for computers, whereas the ordering task is more complicated, i.e., reconstructing such relations from unordered set of sentences. Source documents for MDS may have been written by different authors, by different writing styles, on different dates, and based on different background knowledge. We cannot expect that a set of extracted sentences from such diverse documents is coherent on their own.

When asked to arrange sentences, a human may perform this task without difficulty just as we write out thoughts in a text. However, we must consider what accomplishes this task because computers are, by their nature, unaware of ordering. The most common strategy for sentence ordering is *chronological ordering*, which arranges sentences in the order of their publication dates. However, this study addresses the problem of chronological ordering: some sentences may lose the presupposition assumed in their original documents. Thus, chronological ordering sometimes obscures what a sentence is intended to convey.

In order to deal with the problem case with chronological ordering, this study proposes the use of precedence relations for coherent arrangement. The proposed method improves chronological ordering by resolving precedent information of arranging sentences. This study also proposes an evaluation metric that measures *sentence continuity* and an amendment-based evaluation task. The proposed method achieved good results in a rating task, raising poor chronological orderings to an acceptable level by 20%. Amendment-based evaluation outperformed an evaluation that compares an ordering with an answer made by a human. The sentence continuity metric, when applied to the amendment-based task, showed good agreement with the rating result.

Although several strategies to decide a sentence ordering have been proposed in the previous work, the appropriate way to combine these strategies to achieve more coherent summaries remains unsolved. This chapter also formalizes four criteria to capture the association of sentences. These criteria are integrated into a criterion by a supervised learning approach. The chapter also proposes a bottom-up approach to arrange sentences, which repeatedly concatenate textual segments until we obtain the overall segment with all sentences arranged. Our experimental results showed a significant improvement over existing sentence ordering strategies.

Chapter 6 addresses abbreviation recognition for MDS. Abbreviations result from a highly productive type of term variation which substitutes fully expanded terms (e.g.,

European Union) with shortened term-forms (e.g., *EU*). Abbreviations hinder automatic text summarization as an acronym could be expressed in different forms across documents, e.g., *European Union (EU)*; *European Union*; or *EU*. An MDS system should be aware of associations between shortened term-forms and fully expanded terms so that a summary chooses a consistent term form in describing the same concept/entity.

In practice, no generic rules or exact patterns have been established for dealing with abbreviation creation. Thus, abbreviation recognition aims to extract pairs of short forms (acronyms or abbreviations) and long forms (their expanded forms or definitions) occurring in text. Except for a few studies, most studies focus on parenthetical expressions and locate a textual fragment with an abbreviation definition by using a letter-matching algorithm. However, the letter matching approach cannot deal with a long form whose short form is arranged in a different word order, e.g., *water activity (AW)*. In addition, the letter matching approach is not applicable to Japanese acronyms, which does not necessarily share the same letters between a short/long-form pair because of its foreign origin, e.g., *Yakan Ri-chakuriku Kunren (NLP)*.

This study assumes a word sequence is a possible long-form if the word sequence co-occurs frequently with a specific abbreviation and not with other surrounding words. Satisfying a validation rule for being a long form, the word sequence is stored in the abbreviation dictionary. This approach detects the starting point of the long form without using letter matching. In order to validate a short/long-form pair in English, this study uses a refined letter-matching algorithm that can recognize shuffled abbreviations. This study also examines the number of paraphrase instances in the source documents for validating short/long-form pairs in Japanese. The proposed method outperformed the base-line systems, achieving 99% precision and 82-95% recall on MEDLINE evaluation corpus.

In conclusion, this thesis reports novel methods for sentence extraction, sentence ordering, and acronym recognition. The effectiveness of these methods is shown with some experimental evidences. Even though this thesis targets at multi-document summarization as a problem exemplar, the outcomes of this study will contribute to various NLP applications.