社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ＦＩ－ 84 (21)
2006－ＮＬ－175 (21)
2006／9／13

中国語を対象とした句表現要約技術

胡柏，上田良寛，岡 満美子

E-mail: {Hu.Bai, ueda.yoshihiro, oka.mamiko} @ fujixerox.co.jp

富士ゼロックス(株) サービス技術開発本部

句表現要約は、検索結果のふるい分けに適した要約である。日本語版のアルゴリズムをベースに、中国語を対象とした句表現要約技術を開発した。日本語が係り受けの関係を基本単位として、重要な関係を組み合わせることで句を構成するのに対して、中国語では重要なpredicate-argument構造をひとつ特定してそれにパターンを適用して句を生成する方式を採る。解析結果の表現形式としてLFG (Lexical-Functional Grammar)を用いた。中国語を母国語とする被験者による予備評価で、生成した要約句が文法的に受け入れられることがわかった。

Phrase-Representation Summarization Method for Chinese

HU Bai*, UEDA Yoshihiro, OKA Mamiko
Service Technologies Development Group, Fuji Xerox, Co., Ltd
{Hu.Bai, Ueda.Yoshihiro, oka.mamiko}@fujixerox.co.jp

Phrase-represented summary is suitable for the process of sifting information retrieval results. We have developed the phrase-representation summarization algorithm for Chinese based on the Japanese version. While the algorithm for Japanese creates phrases from syntactic subtree constructed by selecting a core relation and attaching required relations, the method developed for Chinese applies an appropriate pattern to the important predicate-argument structure selected from all the analysis results of input sentences. LFG (Lexical-Functional Grammar) is used as the analysis module. The generated summary phrases have been accepted as grammatically well-formed by the native Chinese speakers who participated in the preliminary evaluation.

1. Introduction

Summaries are often used to support fast and accurate judgment when selecting relevant information from IR results. Ueda et al. [1] proposed "at-a-glance" summary that emphasizes brevity (short in length) and simplicity (less embedded sentences) for such an "indicative" purpose and developed phrase-representation summarization method instead of the important sentence selection adopted by many summarization systems.

We assume such phrase-representation summarization method will be also required for Chinese IR systems and thus we decided to develop this. Because the method has been designed mainly for Japanese and the Chinese linguistic characteristics is rather different, we must reconsider from the linguistic formalism. Linguistic formalism change requires the modification of phrase construction algorithm. In this paper, we make a brief review of the

* He is a graduated student of Shanghai Jiao Tong University joining the 8th visiting fellowship program of Fuji Xerox

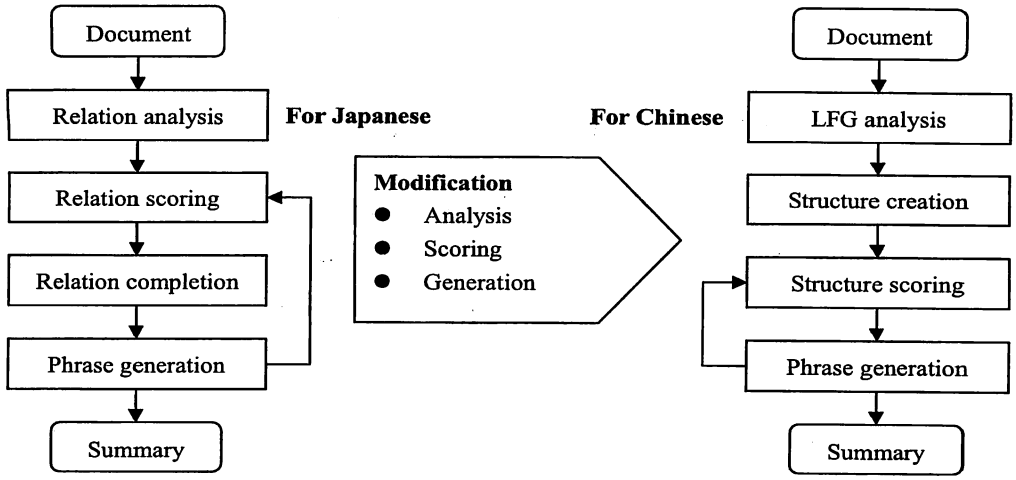

Fig.1 Outline of the basic strategy of Algorithm

phrase representation summarization algorithm for Japanese. Then we will discuss the strategy of modifications to apply it to Chinese. Detailed descriptions will be given to each step.

2. Basic Strategy of Algorithm

The left part of Fig.1 is the brief flow chart of Japanese phrase-represented summary method. First, a document is analyzed to extract the relations between words. Then, a score is calculated to every obtained relation and one core relation is selected based on its score. After that, a sub tree is constructed by attaching additional information to the selected core relation if necessary. A phrase is produced by simply concatenating words in the sub tree.

There are several differences in linguistic features between Japanese and Chinese. Please see the following sample in Fig.2.

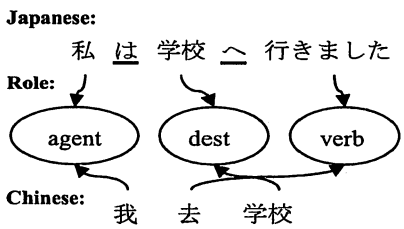

Fig.2 Language differences

In the agglutinative languages such as Japanese, the role of each content word in a sentence is indicated by the function word (the postpositional particle such as "は" and "へ") attached to it. On the other hand, in isolating languages such as Chinese, the roles are indicated by the positions of the content words.

We have made the following decisions to apply the summarization method to Chinese.

1) Analysis

Japanese summarization uses a dependency analyzer that utilizes postpositional particles to analyze the relationship between content words. Such an analyzer cannot be applied to Chinese that lacks such functional words and an alternative analyzer must be selected. The alternative analyzer should utilize the semantic information to reduce the ambiguity. We have adopted LFG (Lexical-Functional Grammar) [2] [3] [4], because it has such good features and the Chinese grammar is also under development in another research group.

One output form produced by LFG, f-structure, is shown in a simplified way as (a) in Fig.3. It includes several predicate-argument (pred-arg) structures,

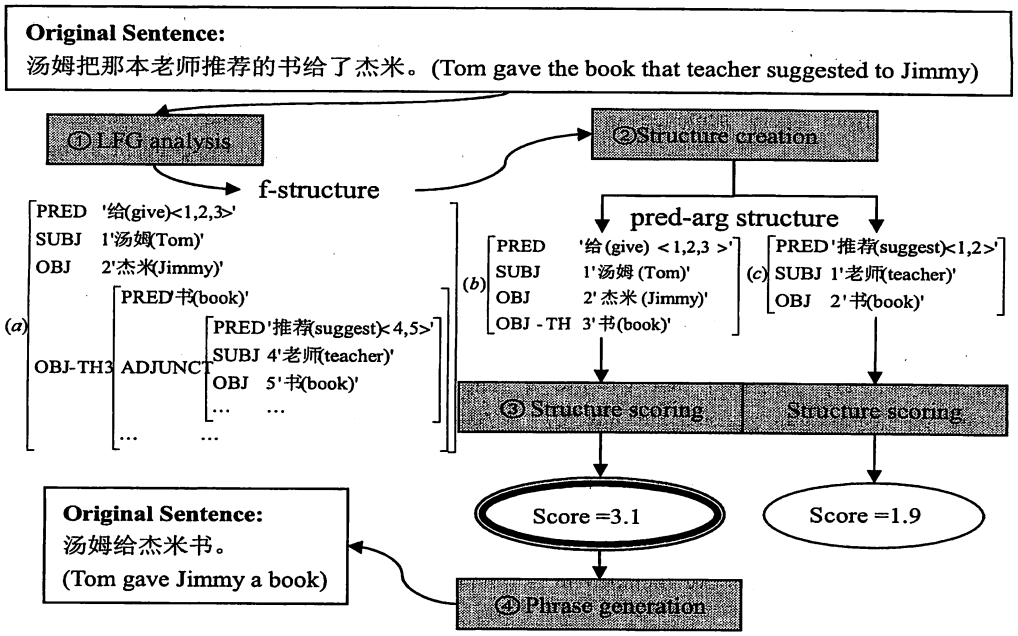

Fig.3 Outline of the phrase-represented summarization for Chinese

as (b) and (c) in Fig.3. pred-arg structure (b) is constructed by a predicate "给" ("give") and its arguments "Tom" (SUBJ/agent), "Jimmy" (OBJ/recipient) and "book" (OBJ-TH/object).

2) Scoring

Japanese summarization gives a score to each relation and selects the relation that has the highest score. The pred-arg structure adopted here is a combination of several relations and no single relation should be isolated from the structure. Thus, we have decided to give score to each pred-arg structure and to select the structure with the highest score.

In Fig.3, two pred-arg structures ((b) and (c)) are extracted and each of them will be given a score. Here we assume that the give-Tom-Jimmy-book structure has the highest score and be selected as the source for a summary phase.

3) Generation

The predicate-argument structure does not reserve the original surface word sequence. Thus, we need to construct a suitable sequence from the select pred-arg structure. The sequence to be generated can be decided by the number and the kinds of the arguments. This rule is encoded to the patterns, each of which has an argument matching pattern and a surface string generation pattern.

The rescoring step that appear in Japanese summarization is also required in Chinese summarization in order to generate multi-phrase summaries

3. Further description

LFG analysis

Lexical-Functional Grammar (LFG) is a linguistic theory that studies the various aspects of linguistic structure and the relations between them. LFG analyses focus on two syntactic structures; constituent structure (c-structure) that represents word order and phrasal groupings and functional structure

(f-structure) that represents grammatical functions like subject and object. Here, LFG analysis is applied to generate f-structure like (a) in Fig.3 from every sentence in the document

Structure creation

For each sentence, one f-structure would be generated. Each summary phrase candidate has its source structure, which is a substructure of this f-structure. We must decide which part is extracted from the sentence f-structure.

We have already described that we use pred-arg (predicate-argument) structure like (b) and (c) in Fig.3 as the source for the summary phrase. In other words, we select substructure that contains predicate with one or more argument(s) and the features that play roles of the corresponding arguments are included in the structure.

The Table 1 shows the important features, i.e. features that can be arguments to be included in the pred-arg structural candidates for phrases.

Elements	Meaning
PRED	Predicate
SUBJ	Subject
OBJ	Object
COMP	Closed complement
XCOMP	Open complement
OBJ-TH	Used for things to be treated by the main verb
OBL-AG	Possible semantic agent in the passive sentence

Table 1

Structure refinement

The pred-arg structure candidates directly extracted from f-structure are not always sufficient to generate a readable phrase. Refinements like word replacement or adding are necessary to generate phrases containing enough information.

The followings are the modification criteria that are used most frequently.

- Generic pronouns modification

Because all of the pronouns in the document are treated as "pro" in f-structure, the pronoun "pro" should be changed to corresponding words like "she", "it", "there", and so on. Besides that, it is necessary to accompany adjunctive information to the concept brought by generic nouns like "的" (sometimes means "thing").

In the f-structure (a) and the extracted pred-arg structure (b) in Fig. 5, both the "SUBJ" and the "OBJ" are treated as a "pro" which itself has no definite meanings. Therefore, SUBJ value should be changed to "他(he)" using the information of SUBJ value in the original f-structure, while OBJ value should be changed to a concrete word "的" and added with the adjunct part "红" to obtain the final form " 红 的 (red one)". By these modifications, (c) is generated.

- Location words adding.

Some other cases like location words are needed to be revised.

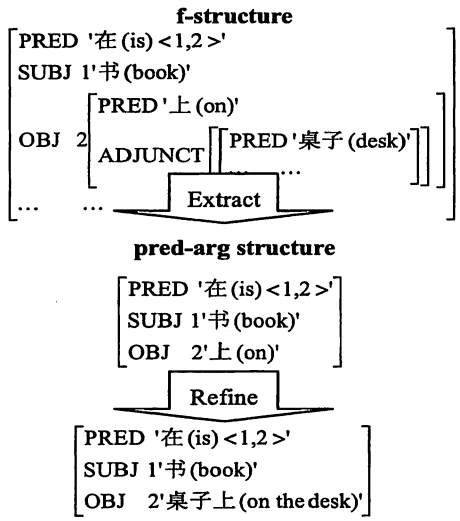

Fig.4 Location word example

As shown in the Fig. 4, Chinese LFG grammar treats "上" (on) as object and "桌子" (desk) as the adjunct part of "上" (on). "桌子" (desk) should also be included in the readable summary phrase, thus we add it in the pred-arg structure by attaching it to "上".

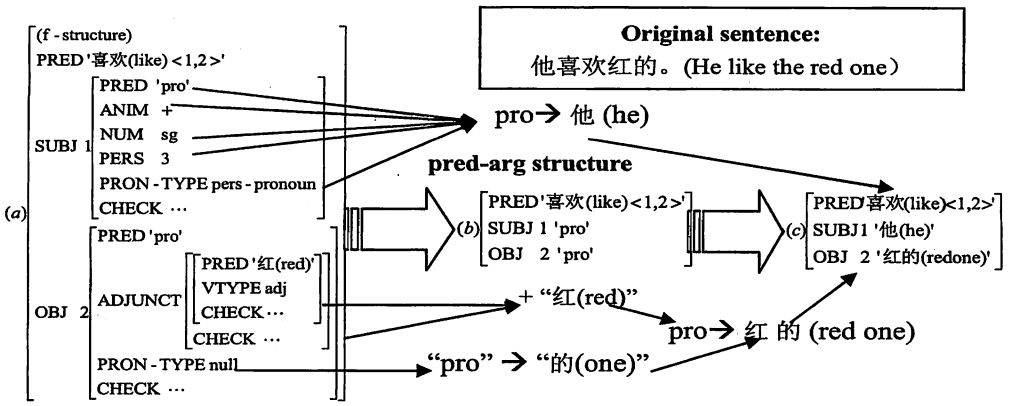

Fig.5 Example of pronoun modification

It is important to mention that the pred-arg structure is independent of LFG. Therefore, all the following steps don't depend on the analysis realization. We also have a plan to apply other analysis tools to this algorithm.

Structure scoring

Every candidate is given an importance score.

The score of every word is calculated by its importance based on its frequency. Usually, the word score is calculated based on the $tf \cdot IDF$ value. Here tf is the term frequency while IDF is the inverted document frequency.

We introduce Gtf (Global term frequency) that is a word statistic obtained from newspaper to replace IDF because of our limitation of the Chinese word resource.

The word score (ws) to word w_i is defined as:

$$ws(w_i) = \frac{tf(w_i)}{Gtf(w_i)} \quad \cdots\cdots (1)$$

For more efficient usage, word score is normalized.

$$ws(w_i) = \frac{ws(w_i)}{\sqrt{\sum_{j=1}^{m} ws(w_i)^2}}, i = 1,2,\cdots m \quad \cdots\cdots (2)$$

In equation 2, m is the number of words in the target text which want to be summarized. Then, the structure score is calculated as follows:

$$score = \sum_{i=0}^{n} C_i \cdot ws(w_i) + C_{sub} \cdot \sum_{j=0}^{p} C_j \cdot ss_j, \quad (3)$$
$$i = 1,2,\cdots n, j = 1,2,\cdots p$$

In equation 3, C_i is the weight of ith word while n is the basic level word number (not including the words in a sub-structure) in the pred-arg structure candidates. And C_j is the weight of ith sub-structure; C_{sub} is the weight of sub structure; p is the total number of the sub-structures that the basic structure has; and ss_j is the score of jth sub-structure. Currently, all words are equally treated ($C_i = C_j = C_{sub} = 1,$). For improvement, C_i and C_j, are hoped to depend on the grammatical role of the argument. Besides, the current C_{sub}, is set to 1 and tends to select bigger pred-arg structure from the f-structure. We also hope to find a right value

$0 < C_{sub} < 1$ in order to increase the possibility of small structure selection.

Finally, the structure candidate with highest score among all the candidates is selected.

Phrase generation

With a pred-arg structure, using predefined phrase patterns, summary phrase could be rightly generated.

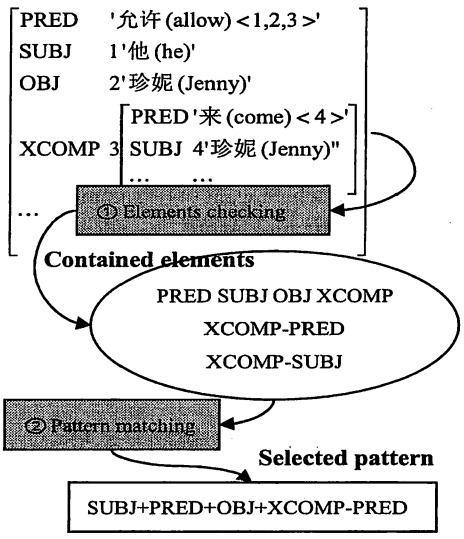

Fig.6 Phrase generation

In Fig.6, because the structure contains "PRED", "SUBJ", "OBJ", "XCOMP", "XCOMP-PRED" and "XCOMP-SUBJ", a phrase will be produced by matching one of the total 40 patterns we created. The matched generation pattern is:

SUBJ+PRED+OBJ+XCOMP-PRED

Therefore, the generated phrase would be:

他允许珍妮来。 (He allows Jenny to come.)

Re-scoring of the candidates

In order to produce multi-phrase summary, these steps including scoring and selection are repeated until the termination condition given in the last step is satisfied. The scores for the words used in the generated phrase are decreased to give chances for other words to be used in the next phrase. This score reduction is achieved by multiplying a decrease ratio R (0<R<1) by the scores of the words have used and recalculating candidates' scores with the new word scores

4. Implementation

We have developed a prototype based on the algorithm of the phrase-represented summary for Chinese in this paper. The input of the prototype currently is XML format file. Every sentence has a corresponding analyzed XML file. Directory is used to represent all the sentences in a document.

The development language is Java and the system is working on Windows XP. The time consumed by summarization process is in proportion to the text length and it takes about 3 seconds (including 325 sentences xml files loading and parsing time) to generate a summary for 325 sentences document (About 2000 Chinese characters) using a PC with Pentium processor (3.2 GHz).

5. Evaluations

Because the LFG grammar is also currently under development, we only have analysis result for about 300 sample sentences, each of which is independent of other sentences. Thus we have not yet evaluated the summarization result for document constructed from multiple sentences.

We have decided to make sentence-level tests to see if the system produces the results of expected quality.

Several summary phrase candidates are produced from one sentence input. We have asked several native Chinese speakers:

1) If the result phrases include appropriate short phrases that carry the essential meaning of the original sentence.
2) If the result phrase with the highest score carries the essential meaning of

the original sentence.

Most subjects satisfied all results of (1) (95.7%) but some from (2) results are not satisfying (91.7%). All the results are presumed to be grammatically well-formed because all native Chinese-speakers accepted the result.

The result shows that our next target of enhancement might be the scoring algorithm. Besides that, it is also necessary to integrate LFG analysis modules into the system and make it produces summary phrases to one document with many sentences in it.

The effect of phrase summary must be evaluated after integrating LFG analysis modules into the system and we can produce summaries for any documents.

6. Related work

Although most summarization studies (including Zechner [5], 1996; Gong and Liu, 2001 [6]) are based on important sentence selection and extraction, some research focus on inter-sentence structure and try to making sentences to shorter ones. However, their purpose seems to reduce the sentence length by removing modifiers with less important meaning. Kaplan's "a note-taking" method [7] is an example of such systems.

Summarization studies for Chinese language seem to have not yet been emerged or the enhancement of morphological analysis might be under way to develop good summarization systems with sentence pickup method

7. Conclusion

Under the concept of "at-a-glance" summary, we produced a Chinese phrase-representation summarization method. This method using LFG to extracts pred-arg structures from the original document and generates phrases as summary. Sentence level evaluation has been done to prove that the result is grammatically acceptable.

We continue to improve the prototype for better performance and have document level evaluation. Besides that, other parser is possible to be used in this algorithm.

Acknowledgement

We should thank Masuich Hiroshi and Ohkuma Tomoko for their important suggestions and helps.

Reference

[1] Yoshihiro UEDA, Mamiko OKA, Takahiro KOYAMA and Tadanobu MIYAUCHI: Toward the "At-a-glance" Summary: Phrase-representation Summarization Method. *Proceedings of COLING-2000:878-884, 2000.*

[2] Stuart M. Shieber: AN INTRODUCTION TO UNIFICATION-BASED APPROACHES TO GRAMMAR *CSLI Lecture Notes Number 4, 1988*

[3]Miriam Butt, Tracy Holloway King, Maria-Eugenia Nino, and Frederique Segond: A GRAMMAR WRITER'S COOKBOOK *CSLI Lecture Notes Number 95, 1999*

[4] Michael T. Wescoat 1989: Practical Instructions for Working with the Formalism of Lexical Functional Grammar *URL .http://www-lfg.stanford.edu/lfg/lfg-introductions/pracinstrucsforlfg.ps.*

[5] Zechner, K. Fast Generation of Abstracts from General Domain Text Corpora by Extracting Reveleant Sentences. *Proceedings of. COLING-96: 986-989,1996*

[6] Yihong Gong, Xin Liu. 2001. Generic text summarization using relevance measure and latent semantic analysis. *In Proceedings of ACM SIGIR'01, pages 19-25, ACM, New York.*

[7] Ronald M.Kaplan, Richard Crouch, Tracy Holloway King, Michael Tepper, Danny Bobrow: A Note-taking Appliance for Intelligence Analysts *2005 Intel Conf on Intelligence Analysis*