Rhetorical Structure Analysis of Japanese Patent Claims using Cue Phrases

Akihiro SHINMORI
Department of Computational Intelligence and Systems Sciences,
Tokyo Institute of Technology, and
INTEC Web and Genome Informatics Corp.
shinmori@isl.intec.co.jp

Manabu OKUMURA　Yuzo MARUKAWA
Precision and Intelligence Laboratory
Tokyo Institute of Technology
{oku,maru}@pi.titech.ac.jp

Makoto IWAYAMA
Precision and Intelligence Laboratory
Tokyo Institute of Technology, and
Hitachi, Ltd.
iwayama@pi.titech.ac.jp

Abstract

The most important part of patent specification is where the claims are written. It is common that claims written in Japanese are described in one sentence with peculiar style and are difficult to understand for ordinary people. We are investigating NLP technologies to improve readability of patent claims. To do so, it is necessary to present the structure of patent claims in a readable way. We found that there are several typical phrases used in claim descriptions and that they can be used as clues to analyze the rhetorical structure of patent claims. We propose a method to analyze the rhetorical structure of patent claims by using these cue phrases and report the result of evaluation.

Keywords: *RST (Rhetorical Structure Theory), Cue Phrase, Claim Readability.*

1 Introduction

In the good and old days, only specialists such as patent attorneys or product engineers in specific fields were dealing with patent. But with the advent of "business-model patent", more and more business persons are concerned about patent.

Patent is described in patent specification. The most important part of patent specification is where the claims are written, because the claims declare and define the scope of the patent. It is common that Japanese patent claims are described in one sentence with peculiar style and are difficult to understand for ordinary people.

操作手段によりアクチュエータを駆動して所望の作業を行なう作業機において、前記作業機の作業機構に作用する負荷を検出する 負荷検出手段 と、この 負荷検出手段 の検出値に応じた周波数の信号を出力する 第１の周波数変換器 と、当該負荷検出手段の検出値に応じた周波数のパルスを出力する 第２の周波数変換器 と、前記 第１の周波数変換器 から出力される信号を前記 第２の周波数変換器 からのパルスの出力期間だけ間欠的に出力する変調手段と、この変調手段の出力信号に応じて振動を発生する振動発生手段とを設けたことを特徴とする作業機の操作用仮想振動生成装置。

Figure 1. A sample Japanese patent claim (Publication Number=10-011111)

We are investigating NLP technologies to improve readability of patent claims[13]. In this paper, we propose a method to analyze the rhetorical structure of patent claims and report the result of evaluation.

2 Characteristics of Patent Claim

A typical Japanese patent claim taken from the first claim of a patent is shown in Figure 1.

As can be seen from Figure 1, the salient characteristics of Japanese patent claims from the viewpoint of readability are as follows:

1. The length of sentence is long.
2. The style of description is peculiar.
3. The structure of description is complex.

To examine the first point, we extracted all of the first claims of the sample data (59968 patents)

in the NTCIR3 patent collection [2], and calculated the average sentence length. We found that it is 242 characters and confirmed that Japanese patent claims are unusually long.

With regard to the second and third point, we surveyed several books and articles written for patent applicants to explain how to draft patent claims[4, 5]. Based on the survey, we can say that there exist three fundamental description patterns in Japanese patent claims:

- Process sequence style
- Element enumeration style
- Jepson-like style

It is summarized in Table 1. Note that these patterns are not mutually exclusive. For example, the known or the precondition part of the Jepson-like style may be written in the process sequence style or in the element enumeration style.

Because of these characteristics, the well-known Japanese parser KNP [6] fails to process most of Japanese patent claims. KNP's dependency analysis works by detecting parallel structure utilizing thesaurus and dynamic programming, but it does not work well for patent claims because there often exist chain-like descriptions in which one concept is first defined and next another concept is defined using the first. For the claim in Figure 1, although the "負荷検出手段 (load detection method)", the "第１の周波数変換器 (frequency transfer device no.1)", the "第２の周波数変換器 (frequency transfer device no.2)", the "変調手段 (modulation method)", and the "振動発生手段 (oscillation generation method)" need to be recognized as parallel, it cannot be recognized due to the existence of "chain-like" expression designated by the underline.

3 Rhetorical Structure Analysis for Readability

To improve the readability of Japanese patent claims, we claim that the structure of description needs to be presented in a readable way. To do so, the structure needs to be analyzed first.

Japanese patent claims are described in such a way that multiple sentences are coerced into one sentence[5]. In other words, a claim is composed of multiple sentences that have some kind of relationships with each other. Therefore, we decided to apply the RST (Rhetorical Structure Theory) [7] which was proposed to analyze discourse structure composed of multiple sentences.

RST was first proposed in the 1980's and has been applied to automatic summarization[8] successfully. A Tcl/Tk-based interactive tool[10] was developed to support manually edit and visually show the structure.

```
－ 操作手段によりアクチュエータを駆動して所望の作業を行なう
    作業機
        において、
＋ 前記作業機の作業機構に作用する負荷を検出する負荷検出手段と、
＋ この負荷検出手段の検出値に応じた周波数の信号を出力する第１の周波数変換器と、
＋ 当該負荷検出手段の検出値に応じた周波数のパルスを出力する第２の周波数変換器と、
＋ 前記第１の周波数変換器から出力される信号を前記第２の周波数変換器からのパルスの出力期間だけ間欠的に出力する変調手段と、
＋ この変調手段の出力信号に応じて振動を発生する振動発生手段と
を設けた
    こと
        を特徴とする
            作業機の操作用仮想振動生成装置。
```

Figure 3. An example of newline insertion, itemization, and indentation

For the rhetorical structure analysis of Japanese patent claims, we defined six rhetorical relations as in Table 2. Two of them are multi-nuclear where composing elements are equally important, and four of them are mono-nuclear where one element is nucleus, the other is satellite, and the nucleus is more important than the satellite. In the "Example" column of Table 2, the regions enclosed with "[" and "]" are segments and the underlined ones are nucleus.

Given the patent claim in Figure 1, we can analyze its rhetorical structure and present it visually by using RSTTool[10] as in Figure 2.

In addition, if the rhetorical structure is analyzed, the original patent claim can be inserted with newlines, itemized, and indented as in Figure 3.

To present a patent claim in the form of Figure 2 or Figure 3 helps readers to understand the claim. We believe it is a first step toward readable patent claims.

4 Rhetorical Structure Analysis using Cue Phrases

4.1 Cue-phrase-based Approach

To analyze the rhetorical structure of Japanese patent claims, we took a similar approach to [8]. We collected cue phrases which can be used for segmenting long claims and establishing rhetorical relations among segments.

Table 1. Description style pattern of patent claim

Style	Description
Process Sequence Style	As in "... し (processing), ... し (processing), ... した (and processed)...", the sequence of processes is described. Mainly used in method invention.
Element Enumeration Style	As in "... と (and), ... と (and), ... とからなる (consisting of), ...", the set of element is described. Mainly used in product invention.
Jepson-like Style	As in "... において (in), ... を特徴とする (characterized by), ...", either the known or the precondition part is first described, then either the new or the main part is described. Note that this style includes more vague claims than the rigidly-defined "Jepson claim" where the know part and the new part are rigidly declared.

Table 2. Rhetorical relations for Japanese patent claims

Type	Rhetorical Relation	Explanation	Example
Multi-Nuclear	PROCEDURE	Process Sequence Style	[〜し、][〜し、][〜する]X
Multi-Nuclear	COMPONENT	Element Enumeration Style	[〜と、][〜と、][〜と] を
Mono-Nuclear	ELABORATION	S elaborates N.	[X を Y した][Z の A]
Mono-Nuclear	FEATURE	Characterization	[X である Y][を特徴とする]
Mono-Nuclear	PRECONDITION	Jepson-like Style	[X であって、][Y した Z]
Mono-Nuclear	COMPOSE	Composition	[〜と、〜と、〜と][を備えた]X

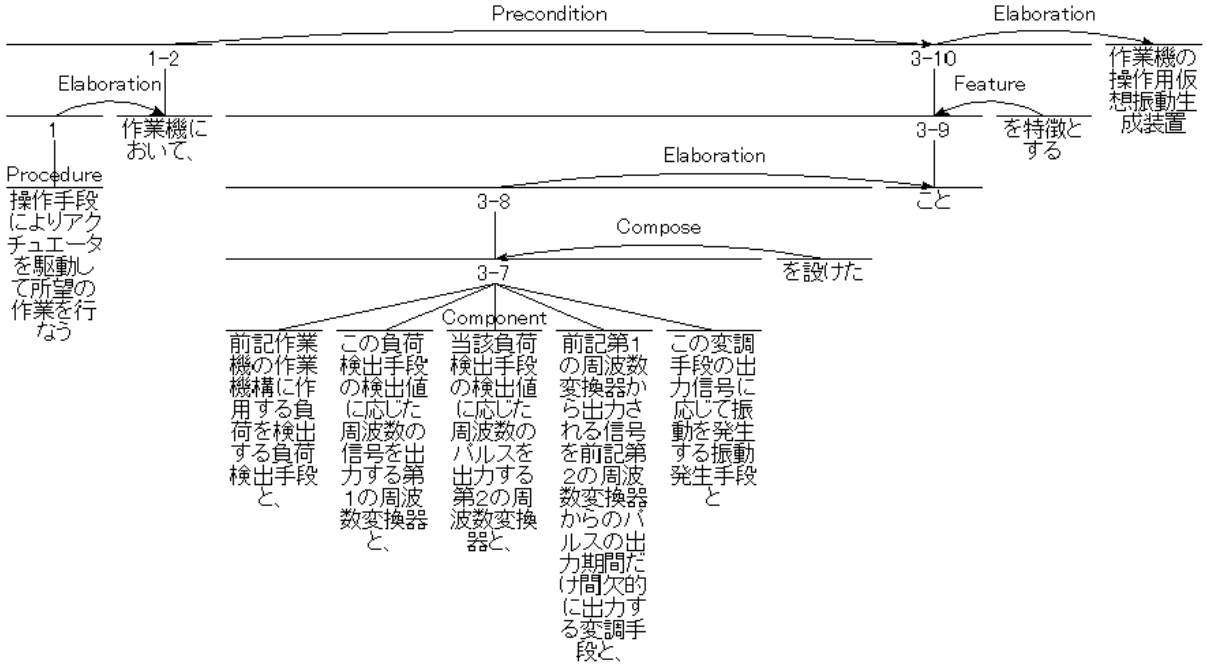

Figure 2. A result of rhetorical structure analysis of patent claim (using RSTTool v2.7)

Table 3. Description pattern just before the newlines in claims which newline are explicitly inserted

No	Pattern	Ratio
1	(Noun|Symbol) と (、|，)	46.1%
2	(Verb-Cont-Form| AuxVerb-Cont-Form)(、|，)	17.5%
3	(Noun|Symbol) において (、|，)	16.4%
4	(Noun|Symbol) であって (、|，)	7.2%

原稿が載置される原稿台と、<nl>
この原稿台に対して主走査方向に移動する走査光学手段と、<nl>
この走査光学手段上に配置され原稿を副走査方向に照明する照明手段と、を備えた画像読取装置において、<nl>
前記照明手段は、前記走査光学手段に対して走査移動平面に略平行に回動自在に取付けられることを特徴とする画像読取装置。

Figure 4. An example of claim which newlines are explicitly inserted (<nl> means newline.)

Cue phrases were first collected manually by reading patent claims. Then we found that about half of the claims are inserted with newlines at seemingly segment boundaries as in Figure 4.

We investigated all of the extracted first claims of the sample data and 48.5% of them are newline-inserted claims. It seems that the drafters of patent claims explicitly inserted those newlines for readability for themselves. We checked the description pattern of the last three morphemes just before each newline of those claims. The result is shown in Table 3. In Table 3, "Verb-Cont-Form" means "動詞連用形" (verb in continuous form) and "AuxVerb-Cont-Form" means "助動詞連用形" (auxiliary verb in continuous form). Note that the description patterns are expressed in the regular expression notation of Perl.

Summarizing the above, we came up with cue phrases in Table 4. In Table 4, "Verb-Basic-Form" means "動詞基本形" (verb in basic form) and "AuxVerb-Basic-Form" means "助動詞基本形" (auxiliary verb in basic form).

4.2 Algorithm and Implementation

We designed an algorithm for rhetorical structure analysis of independent claims [1].

Although patent claims are written in natural language, it's not written in a free form and is restricted in a sense that there are description styles established in the community as in Table 1. So, we designed an algorithm composed of the lexical analyzer and the parser as in the formal language processors[12].

First, the input claim is analyzed with the morphological analyzer "chasen"[9]. Because some patent claims contain newlines, we used "-j" option setting the sentence delimiter as "。：；" in ".chasenrc".

Next, the output from chasen is analyzed with the lexical analyzer. The main point of our algorithm is the context-dependent behavior of the lexical analyzer as follows:

- The lexical analyzer outputs two types of token: cue phrase token and morpheme token.
- Outputting morpheme tokens is done depending on some contextual conditions to avoid ambiguities in the parsing.
- For other morphemes whose context did not satisfy the above conditions, a single morpheme token (WORD) is output.

Next, the output from the lexical analyzer is processed with the parser generated from a context-free grammar (CFG) by using "Bison"[1]-compatible parser generator.

Finally, a rhetorical structure tree is constructed in the form of ".rs2" file used in RSTTool v2.7. By using RSTTool, the output is visually displayed as in Figure 2.

The detail of the lexical analyzer is described in [12]. The CFG designed to analyze Japanese patent claims is shown in Appendix.

5 Evaluation

The evaluation was done by using the collection of the first claims of patent specifications of 1999 in NTCIR3 patent data collection [2]. The evaluation data was different from the one used to collect the cue phrases and to create the CFG.

The evaluation was done in the following points:

- Accept Ratio
- Processing Speed
- Accuracy
  - Indirect evaluation
  - Direct evaluation

The accept ratio was more than 99.77%. The processing speed was 0.30 second per each claim. So, it is almost real-time.

[1] Independent claims are claims which do not refer to any other claims.

[2] First claims are always independent claims.

Table 4. Cue phrases which can be used to analyze patent claims

Token Name	Cue Phrase
JEPSON_CUE	に (お | 於) いて、 であって、 にあたり、 に当 (た)?り、
FEATURE_CUE	を特徴と (した | する)(、)?
COMPOSE_CUE	を搭載して構成され (た | る | ている) (、|，)? を (、|，)?(具 | 備 | そな) え (た | る | ている) (、|，)? を (、|，)?具備 (した | する | している | してなる) (、|，)? (で | から) 構成され (た | ている) (、|，)? を (、|，)?有 (する | した) (、|，)? を (、|，)?包含 (する | した) (、|，)? を (、|，)?含 (む | んだ) (、|，)? から (、|，)?(なる | なった | なっている) (、|，)? を (、|，)?設け (た | ている) (、|，)? を (、|，)?装備 (する | した | している) (、|，)?
NOUN POSTP_TO PUNCT_TOUTEN	The sequence of “(Noun|Symbol) と (、|，)”
VERB_RENYOU PUNCT_TOUTEN	The sequence of “(Verb-Cont-Form|AuxVerb-Cont-Form)(、|，)” which exist before “(Verb-Basic-Form|AuxVerb-Basic-Form)(Noun|Symbol)”

5.1 Indirect Evaluation on Accuracy

By specifying a command-line switch, our program can be run without utilizing the originally inserted newlines. The newline insertion positions can be predicted by the result of rhetorical structure analysis and some heuristics. So, indirect evaluation was done by comparing the newline insertion positions between the originally newline-inserted claims and the automatically newline-inserted claims utilizing the result of rhetorical structure analysis. The recall(R), the precision(P), and the F-measure(F) are calculated by the followings, where c is the number of correctly-inserted newlines, n is the number of newlines in the original claim, and i is the number of inserted newlines.

$$R = \frac{c}{n} \quad (1)$$

$$P = \frac{c}{i} \quad (2)$$

$$F = \frac{2 * R * P}{R + P} \quad (3)$$

The baseline was set in that the newlines are inserted mechanically at the end of every sequence of “(NOUN|SYMBOL)(、|，)” and “(Verb-Cont-Form|AuxVerb-Cont-Form)(、|，)”.

Note that newlines are sometimes inserted at the positions that are not segment boundaries in the meaning of RST. For example, it is often the case that at the end of “は、” (a postpositional particle representing the subject), newlines are inserted. So, our newline-insertion prediction algorithm has the inherent upper limit whose recall is 0.873. The result is shown in Table 5.

Table 5. Evaluation result (Indirect)

Index	Baseline	Newline Insertion utilizing RST	Upper Limit
Recall(R)	0.478	0.674	0.8736
Precision(P)	0.374	0.663	N/A
F-measure	0.420	0.669	N/A

5.2 Direct Evaluation on Accuracy

The direct evaluation was done by using randomly-selected 100 claims. All of these claims are the first claims. We checked the field distribution by the IPC (international patent code) and found it’s almost the same as that of all patent application data in 1999 published by the Japan Patent Office.

The 100 claims were analyzed by our program and the visually-displayed outputs like Figure 2 were presented to a subject who had some experience in reading patent specifications. The subject evaluated the result by the following criteria.

- when the claim is in the Jepson-like style, whether it is correctly analyzed.

Table 6. Evaluation result (Direct)

Category	Count	Percentage (Except "No judgment")
Correct	60	64.52%
Partially Correct	22	23.66%
Incorrect	11	11.82%
No judgment	7	-

- when the claim is in the Jepson-like style, whether the top-level structure is correctly analyzed for the known or the precondition part.
- when the claim is in the Jepson-like style, whether the top-level structure is correctly analyzed for the new or the main part.
- when the claim is not in the Jepson-like style, whether the top-level structure is correctly analyzed for the whole.

The result is shown in Table 6.

In Table 6, "Partially Correct" means the results that are almost correct but partial mistakes as in the followings are reported:

Under-segmentation For example in Figure 5, the subject reported that the leftmost segment should be further segmented around "と共に"(additionally).

Over-segmentation For example in Figure 6, the subject reported that the 6th segment and the 7th segment should be merged.

In Table 6, "No judgment" means the result in which it was not possible to judge because of typographical errors in the original claims.

5.3 Discussion

Patent claims are not always written by specialist such as patent attorney. There are ones written by "individual inventor" and the ones translated from foreign languages. It is often the case that those claims are written in unusual styles. Also there are claims which contain typographical errors. Most of these claims constitute not-accepted claims.

Another reason of the not-accepted claims is the error of morphological analyzer. For example, chasen did not recognize "おむつ" (diaper) as noun and it caused the following algorithm to fail.

The under-segmentation problem can be solved by adding a new cue phrase "と (共 | とも) に"(additionally). The over-segmentation problem is tough because it requires us to incorporate statistical dependency analysis technique.

In any case, both problem can be manually corrected by using interactive operation in RSTTool.

Based on the evaluation and the above discussion, we can say that our approach is adequate.

6 Related Work

A NLP research for patent claim is already reported in [3]. It is directed toward dependency analysis of patent claims. Although it is proposed to support "analytic reading" of patent claims, the evaluation result for large-scale real patent data is not reported. Our approach is different from [3] in that the top-level rhetorical structure is analyzed.

7 Conclusion

We have proposed a cue-phrase-based algorithm to analyze the rhetorical structure of Japanese patent claims. The evaluation result suggest that our approach is robust and practical.

It is not only a first step toward readability to analyze the rhetorical structure of Japanese patent claims and to present it visually, but it can also lead to more challenging task of automatic patent map generation[11] because it would be possible to automatically extract composing elements of patent claims.

Acknowledgement

The NTCIR3 patent data collection was used in our research.

References

[1] C. Donnelly and R. Stallman. *Bison: The YACC-compatible Parser Generator, Version 1.25*, 1995.

[2] M. Iwayama, A. Fujii, A. Takano, and N. Kando. Patent retrieval challenge in ntcir-3. In *IPSJ SIGNotes Natural Language*, number 063. Information Processing Society of Japan, 2001. (in Japanese).

[3] M. Kameda. Support functions for reading japanese text. In *IPSJ SIGNotes Natural Language*, number 110. Information Processing Society of Japan, 1995. (in Japanese).

[4] Y. Kasai. *Manual for Drafting Patent Claims*. Kougyo Chosakai, 1999. (in Japanese).

[5] Y. Kasuya. On the description style of patent claims and the techniques to draft them. *Patent*, 52(2), 1999. (in Japanese).

[6] S. Kurohashi. Knp - japanese parsing for real. *IPSJ MAGAZINE*, 41(11), 2000. (in Japanese).

[7] B. Mann. An introduction to rhetorical structure theory (rst), 1999. http://www.sil.org/ mannb/rst/rintro99.htm.

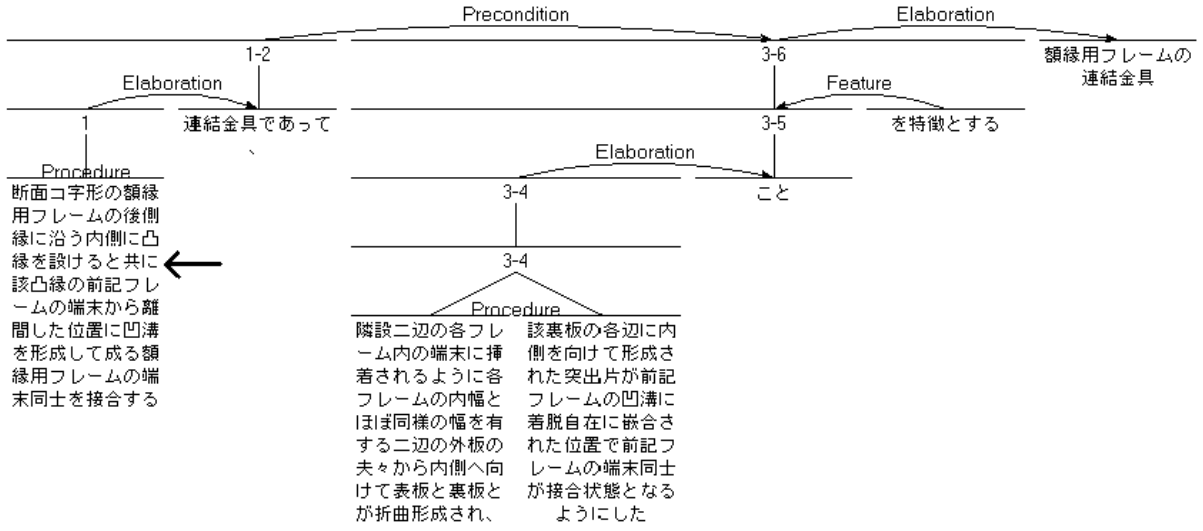

Figure 5. An example of under segmentation. (Publication Number=11-4737) (It was reported that the leftmost segment should be further segmented at the point designated by the arrow.)

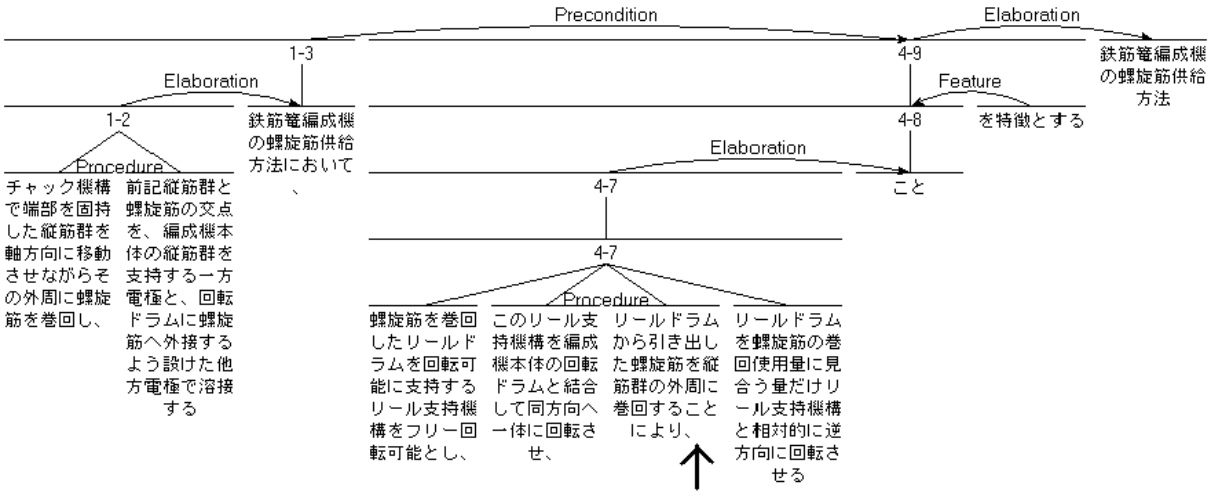

Figure 6. An example of over segmentation. (Publication Number=11-19742) (It was reported that the 6th segment designated by the arrow should be merged with the next segment.)

[8] D. Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, 2000.

[9] Y. Matsumoto, A. Kitauchi, T. Yamashita, Y. Hirano, H. Matsuda, K. Takaoka, and M. Asahara. *Morphological Analysis System ChaSen version 2.2.9 Manual*. Nara Institute of Science and Technology, 2002.

[10] M. OD'onnel. Rst-tool: An rst analysis tool. In *The 6th European Workshop on Natural Language Generation*, 1997.

[11] S. G. on Patent Map. *Patent Map and Information Strategy*. Japan Institute of Invention and Innovation, 1990. (in Japanese).

[12] A. Shinmori, M. Okumura, Y. Marukawa, and M. Iwayama. Rhetorical structure analysis of japanese patent claims using cue phrases. In *IPSJ SIGNotes Natural Language*, number 149. Information Processing Society of Japan, 2002. (in Japanese).

[13] A. Shinmori, S. Saitou, and M. Okumura. Toward automatic paraphrasing of patent claims for readability. In *Workshop on Paraphrasing at the 7th Annual Meeting of the Association for Natural Language Processing*, 2001. (in Japanese).

Appendix: CFG for the Rhetorical Structure Analysis

```
%token JEPSON_CUE FEATURE_CUE COMPOSE_CUE
%token POSTP_NO POSTP_TO
%token NOUN
%token VERB_RENYOU VERB_KIHON
%token PUNCT_TOUTEN
%token WORD
%%
claim_spec:
  after_jepson
  | before_jepson JEPSON_CUE
after_jepson
  ;
before_jepson:
  word_noun_group
  | word_noun_group compose_phrase word_noun_group
  | composed word_noun_group
  | processed
  | processed word_noun_group
  ;
after_jepson:
  word_noun_group
  | word_noun_group compose_phrase word_noun_group
  | composed word_noun_group
  | processed word_noun_group
  | word_noun_group feature_phrase
word_verb_noun_group
  | word_noun_group compose_phrase word_noun_group
feature_phrase word_verb_noun_group
  | composed word_noun_group feature_phrase
word_verb_noun_group
  | processed word_noun_group feature_phrase
word_verb_noun_group
  | word_noun_group feature_phrase word_noun_group
compose_phrase word_verb_noun_group
   ;
composed:
   youso_rekkyo_seq COMPOSE_CUE
   | youso_rekkyo_seq each_youso_complex
PUNCT_TOUTEN COMPOSE_CUE
   ;
processed:
   verb_group
   | shori_rekkyo_seq verb_group
   ;
youso_rekkyo_seq:
   each_youso_complex youso_connect
   | youso_rekkyo_seq each_youso_complex
youso_connect
   ;
youso_connect:
   POSTP_TO
   | POSTP_TO PUNCT_TOUTEN
  ;
each_youso_complex:
   word_noun_group
   ;
shori_rekkyo_seq:
   each_shori_complex
   | shori_rekkyo_seq each_shori_complex
   ;
each_shori_complex:
   word_seq VERB_RENYOU PUNCT_TOUTEN
   | VERB_RENYOU PUNCT_TOUTEN
   ;
feature_phrase:
   FEATURE_CUE
   | PUNCT_TOUTEN FEATURE_CUE
   ;
compose_phrase:
   COMPOSE_CUE
   | PUNCT_TOUTEN COMPOSE_CUE
   ;
word_seq:
   WORD
   | word_seq WORD
   ;
word_verb_noun_group:
   word_noun_group
   | verb_group noun_group
   ;
word_noun_group:
   noun_group
   | word_seq
   | VERB_KIHON noun_group
   | word_seq noun_group
   ;
noun_group:
   NOUN
   | noun_group NOUN
   | noun_group POSTP_NO NOUN
   ;
verb_group:
   word_seq VERB_KIHON
   ;
%%
```