Error Correcting System
for Analysis of Japanese Patent Sentences

YOKOYAMA Shoichi and KENNENDAI* Shigehiro

Graduate School of Yamagata University (* now Dai-Nippon Printing Co.)
Jonan 4-3-16, Yonezawa, Yamagata 992-8510
JAPAN
yokoyama@yz.yamagata-u.ac.jp

Abstract

It is widely known that Japanese patent sentences, especially those regarding necessary conditions and details, have long and complicated structures. If these sentences are investigated by morphological analysis, most of the morphemes are correctly derived because each morpheme can be separated using the connective rules of parts of speech. However, the relation of modification is very difficult because the length of a sentence is large, and because the relationship is very complicated. Even humans researchers are often unable to extract the correct modification. The authors analyzed the automatic error correcting system for the modification analyzer. Initially, we extracted morphemes using the well-known standard morpheme analyzer "Chasen", and then extracted the modification relations for utilizing the standard software "Cabocha." The system automatically extracts the errors of Cabocha and indicates the corrections. We focused on the parallel phrases in Japanese, and estimated the result.

1. Introduction

It is widely known that Japanese patent sentences have long and complicated structures, with up to 200 Japanese characters (50 to 60 words), such that the modifications among phrases also have complicated and difficult structures. It is difficult for even native speakers to understand and clarify such structures.

If an individual wants to apply for a patent, they must retrieve the large-scale patent database in order to confirm whether or not there are similar patents. Correct and accurate retrieval requires automatic information extraction from the patent database.

Recently, the necessity of global application has increased due to rapid technological progress; thus, information should be shared immediately. A patent granted in one country should be valid in another country. If such system is realized, the request of machine translation for a patent will be increased. Therefore, the correct analysis of modification for patent sentences is necessary.

In this paper, we report a system that finds errors of automatic modification, and corrects these errors automatically. We describe the content of the system and the result of an evaluation (Kennendai, 2007).

2. Material and Background

The material is a DVD database in which all available patent gazettes of the Japanese patent office in 2003 are included (Patent, 2005). We have made a comparison of several Japanese patents and their English translations from the database. We previously reported that the modification errors in analyzing Japanese patent sentences reflect the translation result (Yokoyama, 2005). That is, if the modification is in error, the resulting translation also contains the erroneous modification.

If these errors are corrected, correct information about Japanese patent sentences can be obtained. The development of such a system will enable connection to a Japanese proofreading system.

2.1. Comparison of Modification between Japanese and English

The database stores the titles and abstracts of patents and their machine translations. We determined the existence of modification errors by comparing the machine translation data with the human translation data included in the patent database supported by the Japan Patent Office (Industrial Property).

2.2. Classification of Modification Errors

The content of a patent consists of bibliographical terms (publication number, date of publication of application, inventor, title of invention, etc), abstract and solution, range of the patent, detailed explanation, and a simple explanation of figures.

We previously classified the characteristic patterns of modifications occurring in patent sentences primarily written in the abstract and solution (Yokoyama, 2005). Based on this classification, we selected some patterns of modification errors. Analysis of modification is automatically performed by the "Chasen" modification software, which is commonly used by developing by the researchers at Nara Advanced Institute of Science and Technology (NAIST).

(a) Proper Representation in Patent Sentences

「本発明は～（中略）A である」 (This invention is A, which …) (A: noun)

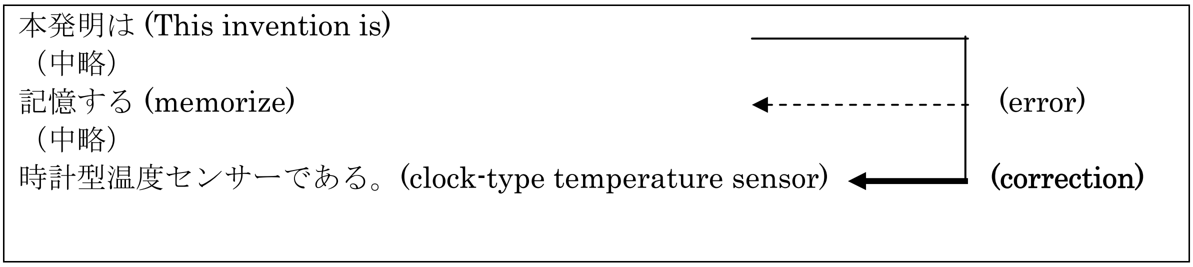
本発明は (This invention is)
（中略）
記憶する (memorize)
（中略）
時計型温度センサーである。(clock-type temperature sensor)
(error)
(correction)

Fig. 1 An example of proper representation in patent sentences

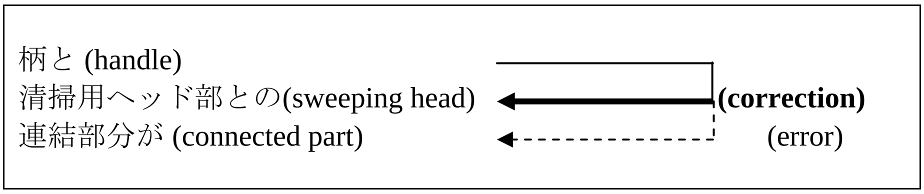
柄と (handle)
清掃用ヘッド部との(sweeping head)
連結部分が (connected part)
(correction)
(error)

Fig. 2 An example of parallel structure in patent sentences

The above pattern is one of the most typical patterns in patent sentences. After the phrase "This invention is A, which," very long modifier(s) would follow. In Fig. 1, the subject "invention" is erroneously analyzed as the predicate modifying the verb "memorize". The correction should be made such that the subject should modify the last phrase "clock-type temperature sensor."

(b) Parallel Structure

「A と B との C が、」 (C of A and B is …) (A, B, and C are nouns.)

As shown in Fig. 2, the Japanese particle "to" ("and") is erroneously analyzed to modify the last noun. Correction is performed by the modification of the parallel property of nouns.

These corrections are usually made by human operators; however, we have developed a system which performs such corrections automatically. Other classifications are conjunctives, subject-verb agreement, modification between subordinate clauses, clause of noun modification, and parallel structure with noun and comma. These categories have not been implemented in the system because of the complexity of the procedure and/or algorithm.

3. System

The flowchart of the correction system, which automatically finds and corrects modification errors, is shown in Fig. 3. First, the patent sentence is input and analyzed by Cabocha. Using Cabocha, the system then finds erroneous candidates among the modifications, primarily through keyword and pattern matching. We also use a Japanese thesaurus (Ikehara, 1997); however, correction at this stage is not sufficiently effective because patent sentences often include many new and unknown words. If modification errors are found, they are then automatically corrected.

An example sentence is shown below. The correction of the sentence belongs to type (b) in the previous section, that is, the parallel structure followed by Japanese particle "to" ("and").

Example (partial) sentence

「製造設備、検査設備の各装置個別のデータ収集とデータ解析を下位のネットワーク上で可能とし、」(to make possible on the sub-network the collection of data and the analysis of data for each device in production facilities and inspection facilities)

0	1D 製造設備、		(production facilities)
1	2D 検査設備の		(inspection facilities)
2	4D 各装置個別の		(each device)
3	4D **<<3 7D>>** データ収集と		(collection of data)
4	7D データ解析を		(analysis of data)
5	6D 下位の		(sub-)
6	7D ネットワーク上で		(on … network)
7	8D 可能とし、		(to make possible)

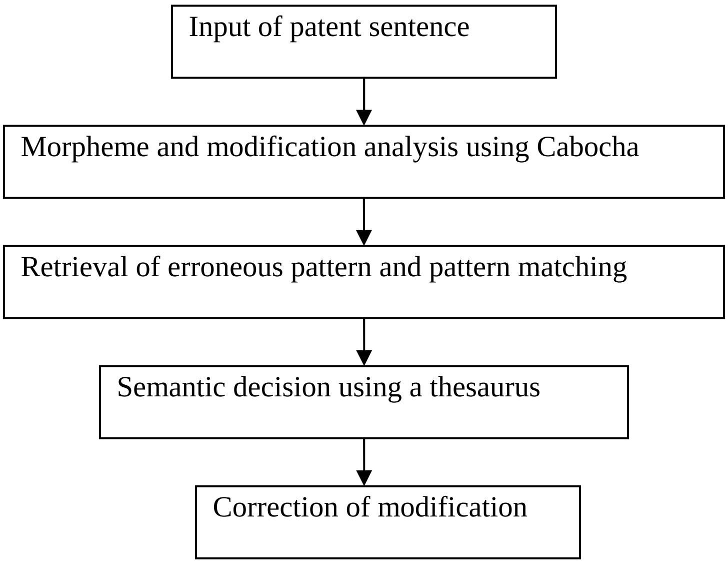
Input of patent sentence
Morpheme and modification analysis using Cabocha
Retrieval of erroneous pattern and pattern matching
Semantic decision using a thesaurus
Correction of modification

Fig.3 Flowchart of the system

The left-most number is the ordering number of the phrase, and the following number ("1D", "2D") is the modified phrase number. That is, "production facilities" correctly modifies "inspection facilities." However, phrase No. 3, "collection of data", erroneously modifies phrase No. 4, "analysis of data." Phrase No. 3 correctly modifies phrase No.7, "to make possible." This is the same pattern as shown in Fig. 2. The procedure for correction begins by finding the particle for parallel structure "to" ("and"). Next, the program retrieves the phrase "tono" (alternatively, "towo", "toni", or "to"). If such a structure is found, then the modification is corrected to the connection from "to" to "towo." In this case, the correction is successfully performed.

4. Evaluation

First, the result of classification of patent sentences by human operators is shown in Table 1.

Table 1 Classification of sample patent sentences

	Total 1228
Proper representation	19
Parallel structure	209
Conjunctives	92
Parallel structure with noun + comma	23
Unclassified error	85
Correct (no errors)	800

These 1228 patents are random files extracted from the DVD database (Patent, 2005). As shown in Table 1, the system deals with 19 proper representations and the parts of parallel structure (34 of 209) for the sentence form "C of A and B". All sample sentences were found and correctly modified and no correct modifications were modified erroneously.

Most of the 175 parallel structures, with 34 exceptions, have structure such as "A and D which C (verb) D." The above correction methodology cannot be applied to such structures. Among 209 parallel structures, the system can only deal with the structure "C of A and B," and cannot correct the similar structure "C of A and (A' and B')", in which B has the embedded parallel structure(s).

5. Concluding Remarks

This paper describes a system for finding and correcting modification errors. However, the system is only a simple prototype for error correction, and should be extended to address other types of errors, as shown in Table 1.

There are similar parallel structures written in the column at the parallel structure with nouns + comma in Table 1. This type of phrase has a complicated parallel structure (e.g., "meats, eggs, vegetables, spinach, eggplant, carrot,…") which sometimes includes a parallel structure with different levels. It is often difficult to clarify such detailed structures. The means to resolve such errors is the use of a thesaurus for semantic interpretation. However, the range and depth of retrieval using a thesaurus is problematic. If the retrieval is too deep, the correct modification is erroneously modified; but, if it is too shallow, the error cannot be corrected satisfactorily.

The use of commas varies for each writer, and decisions on the error or correctness of usage can be difficult even for human operators. We will continue to examine the patterns of such sentences in the future.

If we can classify, detect, and adjust the modification structure of these sentences automatically, we will be able to contribute the improvement of automatic patent translation quality by correcting the modification structure. In addition, the same method can be applied to other type of Japanese sentences with complicated structures as well as patent sentences.

Acknowledgements

The authors thank Mr. Tadaaki OSHIO of the Japan Patent Information Organization (JAPIO) for elucidating the methodology of search and extraction for Japanese and English language patent data. This paper is based on the results of a discussion at the Special Interest Group in AAMT/JAPIO Research Committee.

Bibliographical References

IKEHARA Satoru et al. (eds.): Thesaurus for Japanese Vocabulary (Nihongo Goi Taikei) (in Japanese), Iwanami Publishing Co. (1997).

Industrial Property National Library in Japanese: http://www.ipdl.ncipi.go.jp/ homepg.ipdl

Industrial Property National Library in English: http://www.ipdl.ncipi.go.jp/ homepg_e.ipdl

KENNENDAI Shigehiro and YOKOYAMA Shoichi: A System Correcting Modification Errors in Patent Sentences (in Japanese), Proceeding of the 69th Meeting of the Information Processing Society Japan (IPSJ) (2007) 6Q-3, pp.2-427-8.

Modification Analyzer: “Cabocha”, Nara Institute of Science and Technology.

Morpheme Analyzer: “Chasen”, Nara Institute of Science and Technology.

Patent Database for the Special Interest Group in AAMT/Japio Research Committee, Japio (2005).

YOKOYAMA Shoichi and KANEDA Yuya: Classification of Modified Relationships in Japanese Patent Sentences, Proceedings of Workshop on Patent Translation in the 10th Machine Translation Summit (2005) pp.16-20.