Producing an Encyclopedic Dictionary Using Patent Documents

Atsushi Fujii

Graduate School of Library, Information and Media Studies
University of Tsukuba
1-2 Kasuga, Tsukuba, 305-8550, Japan
fujii@slis.tsukuba.ac.jp

Abstract

Although the World Wide Web has of late become an important source to consult for the meaning of words, a number of technical terms related to high technology are not found on the Web. This paper describes a method to produce an encyclopedic dictionary for high-tech terms from patent information. We used a collection of unexamined patent applications published by the Japanese Patent Office as a source corpus. Given this collection, we extracted terms as headword candidates and retrieved applications including those headwords. Then, we extracted paragraph-style descriptions and categorized them into technical domains. We also extracted related terms for each headword. We have produced a dictionary including approximately 400 000 Japanese terms as headwords. We have also implemented an interface with which users can explore our dictionary by reading text descriptions and viewing a related-term graph.

1. Introduction

Term descriptions that have been carefully organized in hand-compiled dictionaries and encyclopedias provide valuable linguistic knowledge for human use and knowledge-intensive computer systems, developed in the human language technology community. However, as with other types of linguistic knowledge relying on human introspection and supervision, compiling encyclopedias is expensive. It is often the case that users find it difficult to obtain descriptions for new terms and new definitions for existing terms. Therefore, the quantity problem is inherent in conventional encyclopedias.

The World Wide Web, which contains an enormous volume of up-to-date information, is a potential source of encyclopedic knowledge. It has become common practice to consult the Web for specific keywords, instead of consulting dictionaries and encyclopedias. However, existing Web search engines, despite improvements during the past decade, still retrieve extraneous pages. It is often time consuming for users to identify pages that satisfy their information needs. In addition, unlike hand-compiled encyclopedias, in which descriptions are organized using domains and word sense, descriptions in independent Web pages are not related to one another and are not organized. Therefore, the quality problem is crucial in using the Web as an encyclopedia.

To solve the quantity and quality problems described above, we have been proposing a method to produce encyclopedic dictionaries automatically from the Web (Fujii and Ishikawa, 2000; Fujii and Ishikawa, 2001; Fujii and Ishikawa, 2004; Fujii et al., 2002; Fujii et al., 2005). Our method extracts paragraph-style term descriptions from the Web and classifies those descriptions into domains. Our method also extracts related terms for each headword and summarizes multiple descriptions into a single text. We have produced an encyclopedic dictionary including approximately 750 000 Japanese terms as headwords, which is available at a Web search site called "CYCLONE"[1]. CYCLONE has been used for various research purposes.

At the same time, we have identified that descriptions of technical terms associated with high technology are not necessarily found on the Web. Example terms are "photosensitive lithographic printing plate", "tracking error signal", and "magenta coupler". Even in Wikipedia, which is a large encyclopedia on the Web, a number of high-tech terms are not explained.

However, patent documents (either of unexamined patent applications or patents granted by a government patent office) often include high-tech terms and their descriptions. Even in patent documents, dictionary-style definitions are often provided for terms that are important to describe the invention in question.

In this paper, to produce an encyclopedic dictionary for high-tech terms and enhance CYCLONE, we apply our previous method for the Web to patent documents. In TREC, the definition question answering task intended to extract term definitions from newspaper articles (Voorhees, 2003). However, our research is the first serious effort to produce a dictionary using patent documents.

Section 2. describes the method to produce an encyclopedic dictionary and Section 3. shows working examples of our interface for utilizing the resultant dictionary.

2. Methodology

2.1. Overview

Figure 1 shows the overall design of our system, which produces an encyclopedic dictionary from a collection of patent documents. Our system, which currently targets Japanese, consists of six modules: "term recognition", "retrieval", "extraction", "organization", "related term extraction", and "summarization". We explain the entire process in terms of Figure 1.

For a given collection, the term recognition module identifies a headword and the retrieval module searches the collection for documents containing that headword. The extraction module analyzes the layout of the retrieved documents to extract specific segments that potentially describe

[1] http://cyclone.slis.tsukuba.ac.jp/

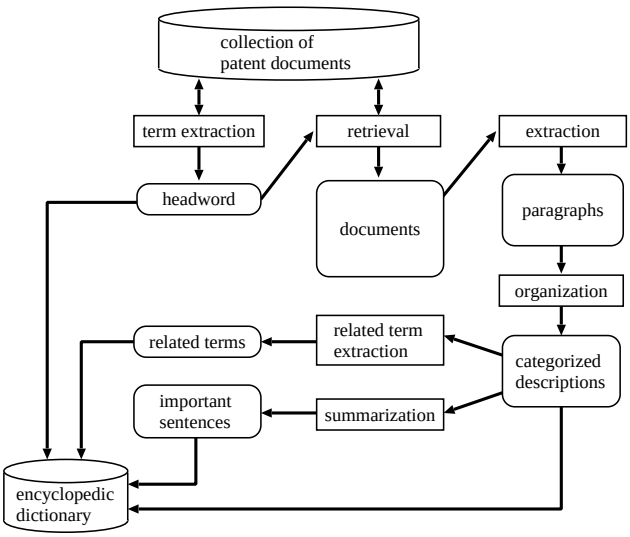

Figure 1: The overall design of our system for producing encyclopedic dictionary.

that headword. The organization module selects high-quality segments as descriptions for that headword and classifies the selected descriptions into domains. The related term extraction module extracts terms that are highly associated with that headword. Related terms can be used as feedback terms to refine the user's focus. Finally, the summarization module extracts important sentences from multiple descriptions and generates a single text for that headword. For each headword, the resultant descriptions, the related terms, and the summary are indexed in an encyclopedic dictionary. Because patent documents have progressively been published from a government patent office, such as the Japanese Patent Office (JPO), we periodically perform the above method and update a dictionary. We have produced an encyclopedic dictionary including approximately 400 000 Japanese terms as headwords.

In Sections 2.2.–2.6., we explain each module, respectively. While the summarization module is the same as that proposed for the Web (Fujii and Ishikawa, 2004), we modified the other modules for patent documents. Thus, we do not explain the summarization module in this paper.

2.2. Term Recognition

The term recognition module searches the patent collection for compound nouns that are not indexed as headwords in our dictionary. To extract compound nouns, we use ChaSen[2] to perform morphological analysis and use hand-crafted rules to identify compound nouns based on part of speech. Each of the extracted compound nouns is a candidate of headwords. However, because patent documents are described by patent-specific wording, extracting headword candidates from patent documents is different from that for the Web. In Japanese patent documents, technical terms are often concatenated with prefixes, such as "*jouki* (above-mentioned)" and "*tougai* (concerned)", and suffixes, such as identification numbers for each component of the invention in question. To resolve this problem, we manually produced a list of stopwords that are not used as a part of headwords and also modified our hand-crafted rules to extract headwords.

[2]http://chasen.naist.jp/hiki/ChaSen/

2.3. Retrieval

For each headword candidate, the retrieval module searches the patent collection for the documents containing that headword. For indexing purposes, we use ChaSen to perform morphological analysis, and use content words, such as nouns and verbs, as index terms. We use Okapi BM25 (Robertson et al., 1994) as the retrieval model, and this ranks retrieved documents according to a TF.IDF-like score. We use 13 years of unexamined Japanese patent applications published by the JPO in 1993–2005. The number of documents is approximately 4 600 000.

2.4. Extraction

The extraction module identifies specific document segments that contain descriptions for the headword in question. The description of a headword is usually only a part of a document, because a patent document is long and its purpose is not to define a specific term but to describe an invention. At the same time, because a term description usually consists of more than one sentence, it is desirable to extract a logical text unit, such as a paragraph. For this purpose, the layout of documents can be useful.

In Japanese patent applications, paragraphs are identified and annotated with specific SGML-style tags by applicants. To extract paragraphs in patent documents, we use a tool produced for the Patent Retrieval Task at the NTCIR Workshops (Fujii et al., 2004; Fujii et al., 2006) and this process is fully automated.

We also use linguistic properties for extraction purposes, because sentences in Japanese term descriptions typically contain specific patterns, such as "X *toha* Y *dearu* (X is Y)". These patterns can be effective clues in selecting desirable term descriptions from many candidate segments. However, there are a number of cases where "*toha*" is not used for definition purposes. During the extraction process, we determine whether a sentence containing "*toha*" is a definition of the headword in question or not. For this purpose, we produced manually annotated examples including "*toha*" and use SVM (Support Vector Machine) to perform a binary decision. Our experiments showed that the accuracy of the binary decision for "*toha*" was 93%. SVM also outputs the score for each decision, which will be used in the organization module.

2.5. Organization

The extraction module potentially outputs extraneous paragraphs that do not actually describe the headword in question. To resolve this problem, the quality of each paragraph has to be evaluated by using language properties. In addition, to produce an encyclopedic dictionary that resembles conventional hand-compiled encyclopedias, descriptions related to different word senses have to be distinguished.

Although many methods have been proposed to identify word sense, such as a method based on the vector space

model (Schütze, 1998), it is still difficult to identify word sense accurately without a dictionary that defines word-sense candidates. In addition, because word sense is often associated with domains, word sense can be distinguished indirectly by determining the domain related to each paragraph. In other words, in a single domain a word is usually used in a consistent meaning (Yarowsky, 1995). For example, the word "pipeline" means a processing method and a transportation pipe in the computer and construction domains, respectively.

In summary, the basis of the organization module is to select appropriate paragraphs based on different criteria and classify those paragraphs into domains. We use a classification method (Iwayama and Tokunaga, 1994) and classify each paragraph into 20 technical domains. The resultant descriptions for each headword are the paragraphs classified into domains.

We compute the score of each paragraph and select paragraphs with a high score. For this purpose, we combine the score of Okapi BM25 in the retrieval module (Section 2.3.), the score of SVM in the extraction modules (Section 2.4.), and the score of the classification.

For producing an encyclopedic dictionary on the Web, we also used the link structure on the Web and the HTML layout of each page. However, these properties cannot be used for patent documents.

2.6. Related Term Extraction

In the related term extraction, we use the same method in the term recognition (Section 2.2.) to extract compound nouns from the descriptions of the headword in question. These extracted compound nouns are candidates of related terms and we select only candidates with a high score. The score of a related term becomes high when that term frequently appears in descriptions with a high score, which is computed in the organization module.

3. Working Examples

Figure 2 shows a retrieval result in response to the term "tracking error signal". In the bottom half of this figure, two descriptions extracted from patent documents are presented. Below the input box are extracted related terms, such as "optical disk" and "objective lens", which can be used as feedback terms to refine the user's focus.

Because a related term for a headword can also be a headword, we can produce a graph, in which a term and a term relation are a node and an edge, respectively. However, because the number of nodes is approximately 400 000, it is difficult to view the entire graph in a single screen. Thus, a user can choose a headword and view only a fragment of a graph associated with that headword. In the current implementation, because each text description as in Figure 2 is linked to its corresponding graph fragment, a user can view a fragment of the graph only by clicking on the link. Figure 3 shows a fragment of related-term graph for the term "dichlorvos", which is linked from the text description for "dichlorvos". In Figure 3, small rectangles denote terms, each of which is linked to its corresponding text description. The length of the link between two nodes denotes the similarity between those nodes.

By surfing on a link structure consisting of text descriptions and graph fragments, a user can perform an exploratory search. First, a user searches our dictionary for the text description of a target term. Second, after surveying that term, the user moves on to the graph fragment associated with that term and browses related terms. Third, by choosing one of the related terms in the graph, the user can read the text description for that term. By repeating the second and third stages, the user can survey descriptions of terms that he/she was not aware of in the first stage.

4. Conclusion

We have proposed a method to produce an encyclopedic dictionary from 13 years of patent applications and have produced an encyclopedic dictionary including approximately 400 000 Japanese terms as headwords. We have also implemented an interface with which users can utilize our dictionary by reading text descriptions and viewing fragments of a related-term graph. Users can also explore our dictionary across text descriptions and graph fragments.

5. Acknowledgments

This research was supported in part by the Industrial Technology Research Grant Program from the New Energy and Industrial Technology Development Organization (NEDO) of Japan (Grant No. 05A14001a).

6. References

Atsushi Fujii and Tetsuya Ishikawa. 2000. Utilizing the World Wide Web as an encyclopedia: Extracting term descriptions from semi-structured texts. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 488–495.

Atsushi Fujii and Tetsuya Ishikawa. 2001. Organizing encyclopedic knowledge based on the Web and its application to question answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 196–203.

Atsushi Fujii and Tetsuya Ishikawa. 2004. Summarizing encyclopedic term descriptions on the Web. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 645–651.

Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. 2002. Producing a large-scale encyclopedic corpus over the Web. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 1737–1740.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2004. The patent retrieval task in the fourth NTCIR workshop. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 560–561.

Atsushi Fujii, Katunobu Itou, and Tetsuya Ishikawa. 2005. Cyclone: An encyclopedic Web search site. In *Special Interest Tracks & Posters of the 14th International World Wide Web Conference*, pages 1184–1185.

Atsushi Fujii, Makoto Iwayama, and Noriko Kando. 2006. Test collections for patent retrieval and patent classification in the fifth NTCIR workshop. In *Proceedings of*

Figure 2: Example descriptions for the term “tracking error signal”.

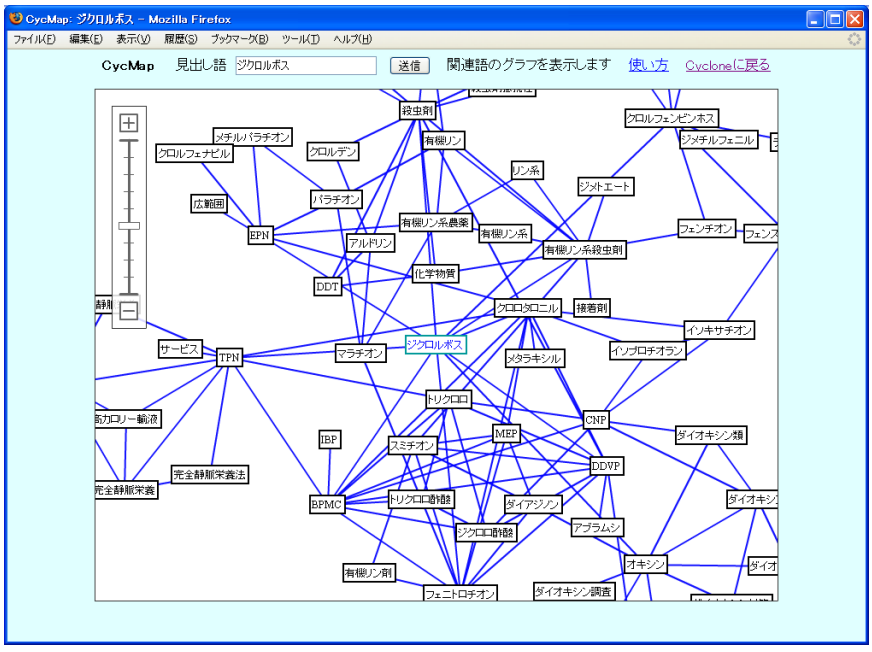

Figure 3: Related-term graph for the term “dichlorvos”.

the 5th International Conference on Language Resources and Evaluation, pages 671–674.

Makoto Iwayama and Takenobu Tokunaga. 1994. A probabilistic model for text categorization: Based on a single random variable with multiple values. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 162–167.

S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. 1994. Okapi at TREC-3. In *Proceedings of the 3rd Text REtrieval Conference*.

Hinrich Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Ellen M. Voorhees. 2003. Evaluating answers to definition questions. In *Companion Volume of the Proceedings of HLT-NAACL 2003*, pages 109–111.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.