Technical Trend Analysis by Analyzing Research Papers' Titles

Kondo Tomoki, Hidetsugu Nanba, Toshiyuki Takezawa

Hiroshima City University
{hirahara, nanba, takezawa}@ls.info.hiroshima-cu.ac.jp

Manabu Okumura

Tokyo Institute of Technology
oku@pi.titech.ac.jp

Abstract

The history of the elemental technologies (underlying technologies) used in a particular research field is essential for analyzing technical trend in the field. However, it is too costly and time-consuming to collect and read all of the papers in the field for the purpose of this analysis. Therefore, we have constructed a system that can recognize the application of elemental technologies to any research field. We focus on the structure of research papers' titles for the extraction of elemental technologies. In research papers' titles, particular expressions, such as "using" or "is based on", are often used. The terms immediately after these expressions are considered elemental technologies. Therefore, we used these expressions as cue phrases, and extracted elemental technologies from both English and Japanese titles. We conducted experiments to investigate the effectiveness of our method for analyzing the structure of titles. We obtained Recall and Precision scores of 0.825 and 0.816, respectively, for the analysis of Japanese titles, and scores of 0.735 and 0.780, respectively, for English titles. Finally, we constructed a system that creates a technical trend map for a given research field.

1. Introduction

The application of the elemental technologies (underlying technologies) used in a particular research field is essential for analyzing technical trends in the field. However, it is costly and time-consuming to collect and read all of the papers in the field for the purpose of this analysis. Therefore, we have studied the automatic analysis of technical trends.

For the extraction of a history of elemental technologies, we focus on the structure of research papers. In research papers' titles, particular expressions, such as "using" or "is based on", are often used. The terms immediately after these expressions are considered to refer to elemental technologies in most cases. Therefore, we use these expressions as cue phrases, and extract elemental technologies from the titles. For example, if the title "Morphological analysis based on HMM" is given, we focus on the cue phrase "based on," and extract "HMM" as an elemental technology. If we regard the head noun phrase "morphological analysis" as a theme of the paper, we can obtain "HMM" and "morphological analysis" as a theme/elemental technology pair.

In the next step, a technical trend map for a given research field can be obtained by the following procedure.

1. Extract theme/elemental technology pairs from all titles in a research paper database.
2. Collect all pairs whose themes match the given field.
3. Plot this data on a graph whose x-axis gives the publication year for each paper and whose y-axis shows the elemental technologies.

For creating more comprehensive technical trend maps, it is necessary to extract theme/elemental technology pairs from titles written in various languages. In this paper, we propose a method that analyzes the structure of titles written in either Japanese or English.

The remainder of this paper is organized as follows. Section 2 shows the system behavior in terms of snapshots. Section 3 describes related work. Section 4 explains our method for analyzing the structure of Japanese and English titles. To investigate the effectiveness of our method, we conducted some experiments. Section 5 reports on the experiments, and discusses the results. We present some conclusions in Section 6.

2. System Behavior

In this section, we describe our system for visualizing technical trends. Figure 1 shows a technical trend map when the research field "speech recognition" was given to the system. In this figure, several elemental technologies used in the "speech recognition" field, such as "HMM" (Hidden Markov Model), are listed on the left hand. These technologies were extracted automatically from research papers in this field, and each paper was shown as a dot in the figure. The x-axis in the figure indicates the publication year for the research papers. If a user's cursor moves to a dot, bibliographic information about the research paper is shown in a pop-up window.

If the user clicks on an elemental technology in the figure, the list of research fields for which the elemental technology was used, is shown. Figure 2 shows a list of research fields for which "HMM" is an elemental technology, and this list is displayed when the user clicks on "HMM" in Figure 1. As shown in Figure 2, we discover that "HMM" was used in the speech recognition field in the 1980s, and that this technology was also used in image processing, such as handwritten character recognition, and in natural language processing, such as morphological analysis, in the 1990s.

3. Related Work

3.1. Utilization of Research Papers' Structures

Taniguchi and Nanba (Taniguchi and Nanba, 2008) studied the automatic construction of a multilingual citation index by collecting Postscript and PDF files from the Internet. They proposed a method for identifying bibliographic information duplicated in Japanese and English, which would be an indispensable

module for the construction in a multilingual citation index. However, a research paper's title is difficult to translate using general machine translation systems because, in general, a title is a large noun phrase.

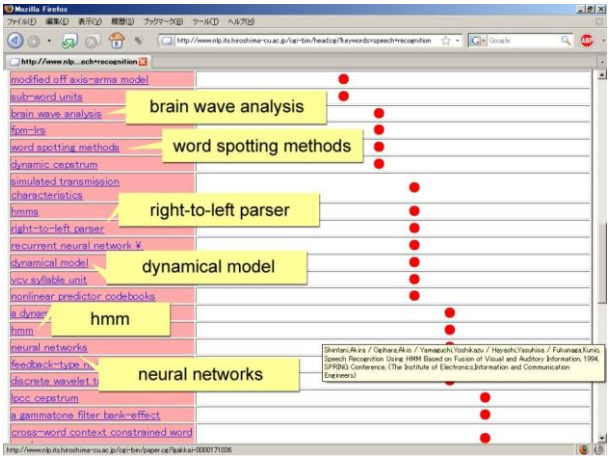

Figure 1: A list of elemental technologies used in a "speech recognition" field

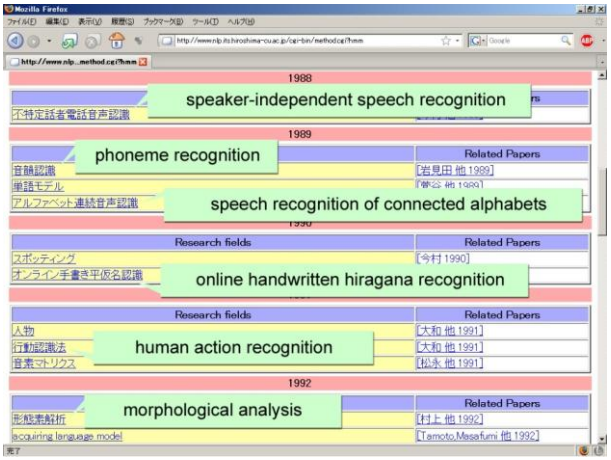

Figure 2: A list of research fields that uses "HMM" as an elemental technology

Therefore, they analyzed structures of both Japanese and English titles, translated technical terms in Japanese titles into English, and identified bibliographic information by comparing translated Japanese terms with English ones.

3.2. Automatic Generation of Survey Articles and Technical Trend Maps

Recently, many researchers have studied automatic generation of survey articles from a set of research papers in a particular research field (Mohammad *et al.*, 2009; Elkiss *et al.*, 2007; Teufel and Moens, 2002; Nanba and Okumura, 1999). Our task is considered a kind of multi-paper summarization in terms of element technologies, though our method generates technical trend maps instead of summary documents.

The interest in systems that analyze technical trends is very high. However, few systems are actually in use. Aureka[1] of Thomson Reuters is introduced as one such system. Aureka is fundamentally a patent analysis system. One function can express quotation relations as a tree. Alternatively, they can be displayed in an aerial view, called a ThemeScape map, which relates the patent to a given patent set. Import of paper data in various formats, such as PDF and MS Word, is possible with this system. Therefore, a paper can be mapped and analyzed via the ThemeScape map of a patent.

[1] http://science.thomsonreuters.com/training/aureka/

4. Analysis of Research Papers' Titles

4.1. Analyzing the Structure of Japanese Titles

We use information extraction based on machine learning to extract any information, such as the elemental technology or topic, from titles. First, we define the tags used in our examination.

- **HEAD** tag includes a topic or a research field of the paper.
- **METHOD** tag includes an elemental technology or data used in the paper.
- **GOAL** tag includes the purpose or the goal of the paper.
- **OTHER** tag includes other words.

This is a tagged example.

[original]
<METHOD> サポートベクトルマシン </METHOD><OTHER> を用いた </OTHER><HEAD>重要文抽出</HEAD><OTHER>に関する研究</OTHER>
[translation]
<OTHER>A Study of</OTHER> <HEAD> Sentence Extraction</HEAD><OTHER>based on </OTHER><METHOD> Support Vector Machines </METHOD>

We formulated the analysis of the structure of titles as a sequence-labeling problem, and analyzed and solved it using machine learning. For the machine learning method, we investigated the Conditional Random Fields (CRF) method, whose empirical success has been reported recently in the field of natural language processing. The CRF-based method identifies the class (tag) of each word. The features and tags given in the CRF method are: (1) the k tags occurring following a target entry, (2) k features occurring before the target entry, and (3) the k features following a target entry (see Figure 3). We used a value of k=5, which was determined in a pilot study. Here, we used the following features for machine learning.

- A word[2].
- Its part of speech[3].
- Whether the word is a method cue (F1).
- Whether the word is a goal cue (F2).
- Whether the word is a final word (F3).

[2] A sequence of nouns (a noun phrase) was treated as a noun.

[3] We used MeCab as a Japanese morphological analysis tool. (http://mecab.sourceforge.net/)

Word	POS	F1	F2	F3	Tag
サポートベクトルマシン (support vector machines)	Noun	0	0	0	
を	Particle	1	0	0	
用い (using)	Verb	1	0	0	
た	Particle	1	0	0	
重要文抽出 (sentence extraction)	Noun	0	0	0	B-HEAD
に	Particle	0	0	0	I-OTHER
関する	Verb	0	0	0	I-OTHER
研究 (a study)	Noun	0	0	1	B-OTHER

target

k

parsing direction

Figure 3: Features and tags given to the CRF

In the following, we describe method cues, goal cues, and final words in more detail.

Method cues
Method cues are the phrases that often appear immediately after the METHOD tag. We prepared 37 phrases to act as method cues. {ex: "を用いた" (using), "に基づく" (is based on), "による" (by)}

Goal cues
Goal cues are the phrases that often appear immediately after the GOAL tag. We prepared 23 phrases to act as goal cues. {ex: "に向けて" (towards), "のための" (for)}

Final words
The HEAD tag is often assigned to the last or penultimate noun phrase in a Japanese title. For example, the HEAD tag is assigned to the penultimate noun phrase "重要文抽出" (sentence extraction) in Figure 3, because the last noun phrase "研究" (a study) is a final word. To collect final words efficiently, we collected the last noun phrases from 255,960 Japanese titles, which we will describe further in Section 5.1. We selected the final words from them manually, obtaining 6,482 final words. {ex: "研究" (study), "実験" (examination), "開発" (development)}

4.2. Analyzing the Structure of English Titles

In a pilot study, we analyzed the structure of English titles in the same way as for Japanese titles. We prepared 18 phrases as method cues, and four phrases as goal cues. We also prepared 924 final words. From the experimental results, we found that the English titles[4] were not analyzed as accurately as the Japanese titles, mainly for the following two reasons:

(1) The complicated structure of English titles
In Japanese titles, the HEAD tag is often assigned to the last or penultimate noun phrase, whereas the tag is assigned to various words in the English titles. In the following examples, the HEAD tag is assigned to the first noun phrase in Ex. 1, whereas the tag is assigned to the third noun phrase in Ex. 2. In Ex. 3, the tag is assigned to the whole title.

[Ex. 1] <HEAD> Electric Field Distribution</HEAD> of Helix LCX on the Ground
[Ex. 2] The Result of Propagation Test on <HEAD>Transmission Line Parallel Sorting </HEAD>on Multi-stage Network
[Ex. 3] <HEAD>GaInAsP/InP High-speed Optical Intensity Modulator</HEAD>

(2) Ambiguity of the cue phrases
Most of the noun phrases immediately before the method cues, such as "を用いた" (using), in Japanese titles and immediately after the method cues, such as "based on", in English titles are element technologies. However, noun phrases immediately after the method cues, such as "with" or "by", in English titles are not necessarily elemental technologies.

To resolve these problems, we used bilingual knowledge and the method for analyzing Japanese titles. In general, some technical terms related to methodology, such as "Hidden Markov Model" (HMM) or "Support Vector Machine" (SVM), have a high degree of probability of being assigned the METHOD tag, while the HEAD tag is likely to be assigned to terms that indicate research fields, such as "information retrieval" or "machine translation". Although the method for English titles cannot analyze as accurately as the method for Japanese titles[5], we can improve the method for English titles by the following procedure[6].

1. Analyze a large number of Japanese titles using the method for the Japanese titles.
2. Extract the noun phrases to which the METHOD tag was assigned, sort them by frequency, and obtain a list of METHOD tags.
3. Obtain lists for HEAD and GOAL, in the same way.

[4] We describe the details of the experiments in Section 5.
[5] We report on the experimental results in Section 5.
[6] Zitouni *et al.* proposed an information extraction method from Chinese, Arabic, and Spanish texts using machine translation techniques and an information extraction tool for English. They experimentally confirmed that their method could improve a simple machine-learning-based approach using Chinese, Arabic, and Spanish tagged texts. However, we did not employ their approach, because we did not have a tool for translating research papers' titles.

4. Translate the top 3,000 terms[7] in each list using bilingual knowledge, which we will describe later.
5. Use the following features for machine learning in addition to the features described in Section 4.1.

Features used in the analysis of English titles
- Whether the word is in the METHOD list.
- Whether the word is in the GOAL list.
- Whether the word is in the HEAD list.

As resources for bilingual knowledge, we used the following.

(i) Statistical machine-translation tool for technical terms
We used a statistical machine-translation tool, developed for translating technical terms (Taniguchi and Nanba, 2008).

(ii) Bilingual lexicon of technical terms created from the NTCIR test collection
We used bilingual lexicon of technical terms, which was created from a document set used in the Cross-lingual Information Retrieval tasks in the first and second NTCIR workshops (Kando *et al*., 1999, Kando *et al*., 2001). It contains 255,960 records of Japanese-English paired documents, with each record comprising a title, the author(s), an abstract, keywords, a publication year, and a conference name. We extracted 710,000 Japanese-English paired keywords from the document set and used them in our task.

(iii) Bilingual dictionary of technical terms
We used a Japanese-English dictionary[8] comprising 450,000 technical terms.

5. Experiments

5.1. Experimental Method

Data sets and experimental settings
We used a document set from the CLIR tasks in the first and second NTCIR workshops. It comprises 255,960 records of Japanese-English paired documents hosted by 65 Japanese academic societies. We randomly selected 1,000 Japanese and English titles from the records, and manually assigned tags to them. For the machine-learning package, we used CRF++[9] software.

Evaluation

$$Recall = \frac{\text{The number of tags that the system could detect correctly}}{\text{The number of tags that should be detected}}$$

$$Precision = \frac{\text{The number of tags that the system could detect correctly}}{\text{The number of tags that the system detected}}$$

Analyzing the structure of Japanese titles
- **J-RULE** (baseline method)
  Japanese titles were analyzed by a rule-based method (Taniguchi and Nanba, 2008). The method assigned the HEAD tag to the last noun phrase in a title. The METHOD and GOAL tags were assigned to noun phrases immediately before a method cue and a goal cue, respectively.
- **J-ML** (our method)
  Japanese titles were analyzed by our machine-learning-based method.

Analyzing the structure of English titles
- **E-RULE** (baseline method)
  English titles were analyzed by a rule-based method (Taniguchi and Nanba, 2008). The method assigned the HEAD tag to the first noun phrase in a title. The METHOD and GOAL tags were assigned to noun phrases immediately after a method cue and a goal cue, respectively.
- **E-ML** (our method)
  English titles were analyzed by our machine-learning-based method.
- **E-ML+MT** (our method)
  English titles were analyzed by our machine-learning based method. To enhance the machine learning, the bilingual knowledge described in Section 4.3 was used.

5.2. Experimental Results

The evaluation results for the analysis of Japanese and English titles are shown in Tables 2 and 3, respectively. As shown in Table 3, our method (J-ML) improved Recall and Precision scores by 0.222 and 0.371, respectively. For the analysis of English titles (see Table 3), E-ML (our method) improved Recall and Precision scores by 0.229 and 0.425, respectively. The Recall and Precision scores for E-ML were 0.078 and 0.041 lower, respectively, than those for J-ML. However, E-ML+MT improved the Recall and Precision scores of E-ML by 0.006 and 0.005, respectively. In particular, the Recall and Precision scores for METHOD in E-ML+MT were both improved, by 0.015 and 0.014, respectively. These results indicate that using J-ML and bilingual knowledge can contribute to the analysis of the structure of English titles.

	J-RULE (baseline method)		**J-ML (our method)**	
	Recall	**Precision**	**Recall**	**Precision**
GOAL	**0.895**	0.333	0.842	**0.842**
HEAD	0.353	0.342	**0.774**	**0.770**
METHOD	0.837	0.736	**0.909**	**0.888**
OTHER	0.328	0.369	**0.776**	**0.762**
Average	0.603	0.445	**0.825**	**0.816**

Table 2: Evaluation results for analyzing Japanese titles

[7] We conducted a pilot study using the top 3,000 terms and the top 6,000 terms. In the experimental results, we obtained higher Recall and Precision scores when using the top 3,000 terms.
[8] "Kagakugijyutsu 45 mango taiyakujiten" Nichigai Associates, Inc., 2001.
[9] http://www.chasen.org/~taku/software/CRF++

	E-RULE (baseline method)		E-ML (our method)		E-ML+MT (our method)	
	Recall	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**
GOAL	0.703	0.474	**0.820**	**0.811**	**0.820**	0.801
HEAD	0.451	0.362	0.726	0.688	**0.731**	**0.693**
METHOD	0.591	0.302	0.736	0.904	**0.751**	**0.918**
OTHER	0.326	0.262	0.704	0.698	**0.711**	**0.708**
Average	0.518	0.350	0.747	0.775	**0.753**	**0.780**

Table 3: Evaluation results for analyzing English titles

5.3. Discussion

Typical errors in the analysis of Japanese titles

There were two typical errors in the analysis of Japanese titles: (1) lack of final words(32.0%) and (2) ambiguity in cue phrases(28.9%). We describe error (1) as follows.

(1) Lack of final words (32.0%)

In the following examples, the HEAD tag was mistakenly assigned to "流通" (Distribution) instead of to "地震データ" (Earthquake Data), because "流通" (Distribution) was not a final word. As we described in Section 4.1, a final word list was created semi-automatically, and most of the frequently used final words, such as "研究" (study) or "実験" (experiment), were already contained in the list. Therefore, we consider that expanding the final word list would be costly and time consuming.

[original]
(Correct) CD-ROM による<HEAD>地震データ</HEAD>の流通
(Analysis result) CD-ROM による地震データの<HEAD>流通</HEAD>

[translation]
(Correct) Distribution of <HEAD>Earthquake Data</HEAD> with CD-ROM
(Analysis result) <HEAD>Distribution </HEAD> of Earthquake Data with CD-ROM

Typical errors in the analysis of English titles (E-ML+MT)

(i) Ambiguity in cue phrases (55.6%)

This error is similar to error (2) in the analysis of Japanese titles. An example is shown below. In this example, the GOAL tag was assigned to "Multiple-valued P1a", because a goal cue "for" appears immediately before it. However, the "HEAD" tag should have been assigned.

(Correct) A Minimization Technique for <HEAD>Multiple-valued Pla</HEAD>
(Analysis result) <HEAD>A Minimization Technique</HEAD> for <GOAL>Multiple-valued Pla</GOAL>

6. Conclusion

We have proposed a method that analyzes the structure of research papers' titles written in either Japanese or English using a machine-learning-based information extraction technique. From our experimental results, we obtained Recall and Precision scores of 0.825 and 0.816, respectively, for the analysis of Japanese titles, and scores of 0.735 and 0.780, respectively, for the analysis of English titles. Finally, we have constructed a system that creates a technical trend map for a given research field.

References

Aaron Elkiss, Siwei Shen, Anthony Fader, Güscedil, Erkan, David States, and Dragomir Radev. (2007). Blind men and Elephants: What do Citation Summaries Tell Us about a Research Article? Journal of the American Society for Information Science and Technology, 59 (1): (pp. 51-62).

Noriko Kando, Kazuko Kuriyama, Toshihiko Nozue, Koji Eguchi, Hiroyuki Kato, and Soichiro Hidaka. (1999). Overview of IR Tasks at the first NTCIR Workshop, Proceedings of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition: (pp. 11-44).

Noriko Kando, Kazuko Kuriyama, and Masaharu Yoshioka. (2001). Overview of Japanese and English Information Retrieval Tasks (JEIR) at the second NTCIR Workshop, Proceedings of the 2nd NTCIR Workshop Meeting: (pp. 4-37 - 4-60).

Saif Mohammad, Bonnie Dorr, Melissa Egan, Ahmed Hassan, Pradeep Muthukrishan, Vahed Qazvinian, Dragomir Radev, and David Zajic. (2009). Generating Surveys of Scientific Paradigms, Proceedings of HLT-NAACL 2009, Boulder, CO.

Hidetsugu Nanba and Manabu Okumura. (1999). Towards Multi-paper Summarization Using Reference Information, Proceedings of the 16th International Joint Conferences on Artificial Intelligence, (IJCAI '99): (pp. 926-931).

Yuko Taniguchi and Hidetsugu Nanba. (2008). Identification of Bibliographic Information Written in both Japanese and English, Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries, (ECDL 2008): (pp. 431-433).

Simone Teufel and Marc Moens. (2002). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status, Computational Linguistics: 28 (4): (pp. 409-445).

Imed Zitouni and Radu Florian. (2008). Mention Detection Crossing the Language Barrier, Proceedings of Conference on Empirical Methods on Natural Language Processing (EMNLP 2008).