Association for Information Systems
AIS Electronic Library (AISeL)

CONF-IRM 2008 Proceedings

International Conference on Information Resources Management (CONF-IRM)

5-2008

Knowledge Sharing from Domain-specific Documents

Eiko Yamamoto
Kobe University, eiko@mech.kobe-u.ac.jp

Hitoshi Isahara
National Institute of Information and Communications Technology, isahara@nict.go.jp

Akira Terada
Japan Airlines Co., Ltd., akira.terada@jal.com

Yasunori Abe
Japan Airlines Co., Ltd., yasunori.abe@jal.com

58F. Knowledge Sharing from Domain-specific Documents

Eiko Yamamoto
Kobe University
eiko@mech.kobe-u.ac.jp

Hitoshi Isahara
National Institute of Information and Communications Technology
isahara@nict.go.jp

Akira Terada
Japan Airlines Co., Ltd.
akira.terada@jal.com

Yasunori Abe
Japan Airlines Co., Ltd.
yasunori.abe@jal.com

Abstract
Recently, collaborative discussions based on the participant generated documents, e.g., customer questionnaires, aviation reports and medical records, are required in various fields such as marketing, transport facilities and medical treatment, in order to share useful knowledge which is crucial to maintain various kind of securities, e.g., avoiding air-traffic accidents and malpractice. We introduce several techniques in natural language processing for extracting information from such text data and verify the validity of such techniques by using aviation documents as an example. We automatically and statistically extract from the documents related words that have not only taxonomical relations like synonyms but also thematic (non-taxonomical) relations including causal and entailment relations. These related words are useful for sharing information among participants. Moreover, we acquire domain-specific terms and phrases from the documents in order to pick up and share important topics from such reports.

Keywords
eCollaboration, Knowledge Sharing, Term Extraction.

1. Introduction
Recently, collaborative discussions based on the participant generated documents, e.g., customer questionnaires, aviation reports and medical records, are required in various fields such as marketing, transport facilities and medical treatment, in order to share useful knowledge which is crucial to maintain various kind of securities, e.g., avoiding air-traffic accidents and malpractice. In this article, we introduce several techniques in natural language processing for extracting information from such text data and verify the validity of such techniques by using aviation documents, such as “captain report,” as an example.

We automatically and statistically extract from the documents related words that have not only taxonomical relations like synonyms but also thematic (non-taxonomical) relations including causal and entailment relations. These related words are useful for sharing information among participants. Moreover, we acquire domain-specific terms and phrases from the documents in order to pick up and share important topics from such reports.

Nowadays, we can access huge amount of text data including web documents and participant generated documents which are informative knowledge sources. They are written by many participants and the terms used in the documents are not controlled. Extracting useful information from such document, classifying them, and arranging them in order are crucial for knowledge management. Extracting relations among terms in the documents and picking up important contents from the documents will be key technology for knowledge management.

2. Extracting related terms useful for handling domain-specific documents

We attempted to automatically and statistically extract domain-specific related words that have not only taxonomical relations but also thematic (non-taxonomical) relations by measuring the semantic distance between words in aviation documents which are Japanese documents containing many English words and abbreviations.

We extract such related words in two steps: (1) characterizing each word by a feature vector which represents collocation relations between words, and (2) estimating the semantic distance between each two words with a statistical measure and extracting pairs of specifically related words.

In step 2, we utilized the Complementary Similarity Measure (CSM), which can determine the relation between two words in a text data by estimating inclusive relations between two vectors representing each appearance pattern for each word. Details of this extracting method are described in (Yamamoto et al. 2007; Yamamoto & Isahara, 2008)

2.1 Experimental results

As a first task in this experiment, we used a collection of aviation reports, e.g., captain reports written in Japanese but contain many English words and abbreviations. They contain 6,427 reports from 1992 to 2003. Each of the reports includes fixed information such as departure place and arrival place, and the content (including the title, the report body by a pilot, and the reply to the report) described in free style. We processed only the content described in free style, deleting the personal information.

We show two typical results here: extraction of taxonomical (mainly synonymic) relations and extraction of thematic relations. As for the taxonomical relations, we extracted word pairs whose CSM values were very high, i.e. both words appeared in a very similar context. The top 10 word pairs extracted are shown in Figure 1, where each of the last columns is the relation between the two words judged by humans. “Synonym” judged by humans also includes abbreviations.

Because the aviation reports are written by many pilots, the terms used in the report are not controlled. Therefore, extraction of synonyms and abbreviations are crucial to share useful information among staffs of airline companies.

junbi (preparation)	*syuppatsu-junbi* (preparation for departure)	hyponym
FLT (flight)	*hiko* (flight)	synonym
sagyo (work)	*seibi* (maintenance)	synonym
sagyo (work)	*syuppatsu-junbi* (preparation for departure)	hyponym
seibi-shochi (repair treatment)	*seibi-sagyo* (maintenance work)	synonym
chakuriku (landing)	ATB (Air Turn Back)	hyponym
unko (flight)	FLT (flight)	synonym
sagyo (work)	ENG-Start (Engine Starting)	non-taxonomic
kaizen (improvement)	*zensho* (taking proper measures)	synonym
chosa (investigation)	*kento* (examination)	synonym

Figure 1: The top 10 word pairs, with the relation judged by humans

As for the thematic relations, we also extracted word pairs whose CSM values were very high, but using another kind of linguistic data. The top 10 word pairs extracted are shown in Figure 2, with their CSM values. Words in these word pairs tended to appear in the same sentences and were thematically related.

1.000000	*kansha* (gratitude)	*i* (feelings)
0.782266	*konkai* (this time)	*kesu* (case)
0.660146	*kyukyusha* (ambulance)	*tehai* (arrangement)
0.641839	*ishi* (doctor)	*shinsatsu* (consultation)
0.623064	*ishi* (doctor)	*shindan* (diagnosis)
0.619127	*konkai* (this time)	*jirei* (example)
0.560951	*okyakusama* (customer)	*gomeiwaku* (trouble, nuisance)
0.533606	*gen'in* (cause)	*kyumei* (investigation)
0.489483	*ryokyaku* (passenger)	*shippei* (disease)
0.485799	*hassei* (occurrence)	*kyubyonin* (emergency patient)

Figure 2: The top 10 word pairs extracted from another kind of data, with their CSM values

We show below some of the related word sets extracted by connecting word pairs based on their CSM values. They seem to have a thematic relation among terms composing each of them.

- *jikan* (time) – *seibi* (maintenance) – *tenken* (check) – *tochaku-go* (after arrival)
- *kokan* (substitution) – *buhin* (parts) – *cyotatsu* (supply)
- *hokoku* (report) – *ryokyaku* (passenger) – *zaseki* (seat) – *se* (back)
- *joho* (information) – *jizen* (prior) – *kisho-joho* (weather information)
- *jokyo* (situation) – *henka* (change) – Cabin-PRESS
- *hokoku* (report) – *itami* (pain) – *senaka* (back)

2.2 Extracting related terms from English documents

We also try to apply this method to extract related terms from English documents. Japanese case-marking particles define not deep semantics but rather surface syntactic relations between words/phrases; therefore, we utilized not semantic meanings between words, but classifications by case-marking particles. Therefore, our method is applicable to other languages when a syntactic analyzer that classifies relations between elements, such as subject, direct object, and indirect object, exists for the language.

In this experiment, we used a collection of airplane operation manuals in English. First, we parsed the documents with HPSG-based English parser *Enju* Version 2.2 (2007) and made linguistic data based on dependency relations between terms in a sentence. Next, we collected dependency relations in each sentence and compiled linguistic data using collocations between a verb and its direct object and a verb and its subject. Then, from these linguistic data, similar to Japanese documents, we tried to extract the pairs of related terms with the method based on CSM in order to obtain taxonomically related terms. Our experiment is still in the beginning stage; however, we show some of extracted word pairs in Figure 3.

0.821655	door	lavatory
0.819880	door	cargo door
0.815458	procedure	checklist
0.771699	lavatory	waste system
0.765525	fuel	lb
0.756569	seat	passenger seat
0.751148	procedure	correction
0.746368	position	latch
0.731082	door	compartment
0.729337	be	become
0.721839	position	switch
0.711045	system	compartment
0.698445	message	status message

Figure 3: Examples of the word pairs extracted from airplane operation manuals in English

3. Extracting salient nouns and phrases

We also tried to extract salient terms including compound nouns and noun phrases from the aviation documents.

There are several methods to acquire new words from large amount of text and some of them showed high performance for compound nouns (Nakagawa & Mori, 2003). Our aim is to acquire technical terms which include not only compound nouns but also longer phrases such as “Knowledge Sharing from Domain-specific Documents” in Japanese. The method handles morpheme based n-grams to save processing time and space.

Our term acquisition method consists of two processes: an extraction of candidate terms (“Candidate Selection”) and a guess as to terms (“Unithood Checking”). First, the statistical indicators we defined are used to select all one-morpheme to ten-morpheme strings that appear at least once in a large number of documents, and also appear repeatedly in several documents. This enables a computer to emulate the human ability to recognize and understand unknown terms. Next, the strength of connection between the constituent morphemes of each candidate term is assessed to arrive at a guess as to whether or not it is in fact a term. Each process is described below in detail.

4.1 Candidates selection

For selecting term/phrase candidates, we considered that terms/phrases that characterize the document collection are judged by the two different standards: terms that represent certain documents in the collection, and terms that represent the entire collection. It is reported that for terms that represent the document, their *df2/df* value tends to be in a certain range (Church, 2000). *df* and *df2* here indicate document frequency and document frequency for appearing more than once, respectively. On the other hand, terms that represent the entire collection are distributed through the collection, but not too widely because such terms include functional words. The idea is expressed by the *df/cf* value within the certain range, where cf indicates collection frequency. According to these considerations, we assume terms whose *df2/df* and *df/cf* values were within a certain range to be candidates. In our experiment, we listed up all the morpheme strings from bi-gram to 10-gram, and selected the ones empirically.

4.2 Unithood checking

Next, the candidates are narrowed down by checking the "unithood" (the appropriateness as a word unit (Kageura & Umino, 1996)).

One of the functions to check the unithood is Tanaka's function (Tanaka-ishii et al. 2003), which is a variation of C-value (Frantzi & Ananiadou, 1996).

$$F(Z) = \log(ml(Z)+1)\cdot \log(cf(Z))\cdot\left(1-\frac{1}{cd(Z)}\right) \tag{1}$$

Here, Z is an n-gram string, $ml(Z)$ is the number of morphemes in Z, and $cd(Z)$ is the number of different morphemes adjacent to Z. The first term $\log(ml(Z)+1)$ in function (1) is the length term, the second term $\log(cf(Z))$ is the frequency term, the third term $(1\text{-}1/cd(Z))$ is the term for the number of adjacent different morphemes. We have tested variations of function (1) and found function (2) shown below performed best.

$$F'(Z) = \log\left(cf(Z)\cdot\frac{cd(Z)}{(cf(Z)+cd(Z))}\right) \tag{2}$$

Note that in function (2), the length term in (1) is eliminated and the term for the number of different morphemes are corrected to reduce the effect of the frequency. To pass the process, the candidate must have the higher $F'(Z)$ value than the one of the 1-morpheme shortened strings. It is applied for both directions. The condition removes the candidate from the list that contains the more appropriate word unit inside.

4.3 Experimental results

We extracted from the aviation documents the salient terms including compound nouns and noun phrases. Some of the interesting results are shown below, where words in italic are originally in Japanese and were translated into English for explanatory purpose. Our method can extract various kinds of long phrases.

Phrases consisting of Japanese words only

- *Calculation of the maximum landing weight*
- *Serious situation which makes it difficult to continue the flight*

Phrases consisting of English words only

- Cargo Conditioned Air Flow Rate selector
- Maximum Takeoff Weight Balanced Field Length Limit

Compound nouns

- Default RNP *value*
- KD *staff*

Phrases with both Japanese words and English words

- *Check that* FMC Position *is* updated *by* GPS
- *Trouble of* Fuel Control System

Phrases with coordinate conjunctions

- Auto pilot *and/or* Auto throttle
- Balance Manifest *and* Takeoff Data

We will continue examination of the effects and the characteristics of the experimental results extracted by our method from the aviation documents as an example, and evaluate their validity to support information sharing and knowledge management.

References

Church, K. W., "Empirical Estimates of Adaptation: The chance of Two Noriega's in close to p/2 than p^2", Proceedings of Coling 96, 1:180-186, 1996.

Enju, http://www-tsujii.is.s.u-tokyo.ac.jp/enju/, Version 2.2, 2007.

Frantzi, K. and Ananiadou, S., "Extracting Nested Collocations", Proceedings of Coling 96, 41-46, 1996.

Kageura, K. and Umino, B., "Methods of Automatic Term Recognition: A Review", Terminology, 3(2): 259-289, 1996.

Nakagawa, H. and Mori, T., "Automatic Term Recognition based on Statistics of Compound Nouns and their Components", Terminology, 9(2): 201-219, 2003.

Tanaka-Ishii, K., Yamamoto, M. and Nakagawa H., "Kiwi: A Multilingual Usage Consultation based on Internet Searching", Proceedings of the Interactive Posters/Demonstrations, ACL-03, 105-108, 2003.

Yamamoto, E. and Isahara, H., "Extracting Word Sets with Non-Taxonomical Relation", Proceedings of the ACL 2007 Demo and Poster Sessions, 141-144, 2007.

Yamamoto, E. and Isahara, H., "Extraction of Word Set for Increasing Human-Computer Interaction in Information Retrieval", Proceedings of Conf-IRM 2008, 2008.