Extraction of Informative Expressions from Domain-specific Documents

Eiko YAMAMOTO[1,2], Hitoshi ISAHARA[1,2], Akira TERADA[3], Yasunori ABE[3]
[1]Graduate School of Engineering, Kobe University
1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan
[2]National Institute of Information and Communications Technology
3-5 Hikari-dai, Seika-cho, Souraku-gun, Kyoto 619-0289, Japan
[3]Japan Airlines Co., Ltd.
Terminal 1, 3-2 Haneda Airport 3-chome, Ota-ku, Tokyo 144-0041, Japan
E-mail: eiko@mech.kobe-u.ac.jp, isahara@nict.go.jp, akira.terada@jal.com, yasunori.abe@jal.com

Abstract

What kinds of lexical resources are helpful for extracting useful information from domain-specific documents? Although domain-specific documents contain much useful knowledge, it is not obvious how to extract such knowledge efficiently from the documents. We need to develop techniques for extracting hidden information from such domain-specific documents. These techniques do not necessarily use state-of-the-art technologies and achieve deep and accurate language understanding, but are based on huge amounts of linguistic resources, such as domain-specific lexical databases. In this paper, we introduce two techniques for extracting informative expressions from documents: the extraction of related words that are not only taxonomically related but also thematically related, and the acquisition of salient terms and phrases. With these techniques we then attempt to automatically and statistically extract domain-specific informative expressions in aviation documents as an example and evaluate the results.

1. Introduction

Recently, thanks to the development of high-performance computers and large-capacity storage devices, huge amounts of domain-specific documents that used not to be available are being generated and stored in all fields, such as marketing, transport facilities and medical treatment. Typical examples of such documents include customer questionnaires, aviation reports, and medical records. Although these data contain much useful knowledge, it is not obvious how to extract such knowledge efficiently from these documents. We need to develop techniques for extracting hidden information from such domain-specific documents. These techniques do not necessarily use state-of-the-art technologies and achieve deep and accurate language understanding, but are based on huge amounts of linguistic resources, such as domain-specific lexical databases.

What kinds of lexical resources are helpful for extracting useful information from domain-specific documents? Although (Kiyota & Nakagawa, 2006) tried to extract similar case frames from aviation documents, we examine processes by which humans extract informative knowledge from web documents. In this paper, we introduce two techniques for extracting informative expressions from documents: the extraction of related words that are not only taxonomically related but also thematically related, and the acquisition of salient terms and phrases.

As for the extraction of related words, we calculate similarities among words based on the inclusive relations between appearance patterns of words and construct meaningful sets of words by connecting related word pairs. Such related terms help to lead users to high-quality information, and the word sets themselves can be regarded as informative expressions in domain-specific documents. Furthermore, because our approach can extract informative relations from relatively few documents, it can help solve the problem of data sparseness.

As for the acquisition of salient terms and phrases, we extract salient terms including compound nouns and longer noun phrases from domain-specific documents. These terms and phrases are useful for informing users of noteworthy topics in the domain.

In this paper, we attempt to automatically and statistically extract domain-specific informative expressions in aviation documents. Section 2 describes the extraction of related words from documents in Japanese, while Section 3 does the same for English documents. Section 4 examines the acquisition of salient terms and phrases from Japanese documents.

2. Extracting Related Terms

We try to extract related terms useful for information extraction. As for the relations among words, there are at least two kinds of relation: the taxonomical relation and the thematic relation. The former is a relation representing the physical resemblance among objects, which is typically a semantic relation such as a hierarchical, synonymic, or antonymic relation[1]; the latter is a relation between objects through a thematic scene, such as "milk" and "cow" as recollected in the scene "milking a cow," and "milk" and "baby," as recollected in the scene "giving a baby milk," which include a causal relation and an entailment relation. Wisniewski & Bassok (1999) showed that both relations are important in recognizing those objects in cognitive psychology.

[1] The taxonomical relation which is, for example, provided by WordNet (Fellbaum, 1998) corresponds to the "classical" relation by Morris & Hirst (2004), and the thematic relation corresponds to the "non-classical" relation.

In lexical database research, progress is being made in the extraction of relations between words in non-specific domain documents, notably with taxonomical semantic lexical databases such as WordNet (Fellbaum, 1998) and the EDR electronic dictionary (1998) which are used for natural language processing research worldwide. These databases are essential for enabling computers, and even humans, to fully understand the meanings of words because lexicons are the origin of language understanding and generation. However, they mainly focus on related words with taxonomical relations such as synonyms, hypernyms-hyponyms, and antonyms, and it is not easy to apply them to practical domain-specific tasks because they are lexical resources for general words. Related to this problem, many researchers in natural language processing have developed many methodologies for extracting various relations from corpora. Several methods exist for extracting relations such as "is-a" (Hearst, 1992), "part-of" (Girju, 2006), causal (Girju, 2003), and entailment (Geffet & Dagan, 2005).

We extract such related words in two steps: (1) we characterize each word by a feature vector which represents collocation relations, which are based on dependency relations between words, and (2) we estimate the relation between each two words by using a statistical measure and extract pairs of specifically related words.

2.1 Linguistic Data

In step 1, we make several kinds of linguistic data based on a modifiee/modifier relationship in documents and characterize each word by a feature vector, which represents collocation relations for each linguistic data.

The Japanese language has case-marking particles that indicate the semantic relation between two elements in a dependency relation, which is a kind of modifiee/modifier relationship. For our experiment in Japanese, we used such particles and extracted the data from the documents we gathered. First, we parsed sentences with the KNP[2]. From the results, we collected dependency relations matching one of the following five patterns of case-marking particles. With A, B, P, Q, R, and S as nouns (including compound words); V as a verb; and <*X*> as a case-marking particle with its role in parentheses, the five patterns are A <*no* (of)> B, P <*wo* (object)> V, Q <*ga* (subject)> V, R <*ni* (dative)> V, and S <*ha* (topic)> V.

Suppose we have the following sentence:

> "*Chloe ha Mike ga Judy ni bara no hanataba wo okutta to kiita*
> (Chloe heard that Mike had given Judy a rose bouquet.)."

We can extract from this sentence five dependency relations between words:

> *bara* (rose) <*no* (of)> *hanataba* (bouquet),
> *hanataba* <*wo* (object)> *okutta* (had presented),
> *Mike* <*ga* (subject)> *okutta*,
> *Judy* <*ni* (dative)> *okutta*,
> *Chloe* <*ha* (topic)> *kiita* (heard).

[2] A Japanese parser developed at Kyoto University.

The following types of linguistic data can be compiled from this set of dependency relations:

- ***NN-data*** *based on co-occurrence between nouns.* For each sentence in our document collection, we gathered nouns followed by all five of the case-marking particles we used and nouns proceeded by <*no*>; that is, A, B, P, Q, R, and S. For the above sentence, we can gather *Chloe*, *Mike*, *Judy*, *bara*, and *hanataba*. Each noun is represented by a binary vector showing whether each noun occurs in each sentence. The number of data items equals the number of sentences in the documents.
- ***NV-data*** *based on a dependency relation between noun and verb.* We gathered nouns P, Q, R, and S followed by each of the case-marking particles <*wo*>, <*ga*>, <*ni*>, and <*ha*> for each verb V. We named them *Wo-data*, *Ga-data*, *Ni-data*, and *Ha-data*, respectively. For the verb *okutta* in the above sentence, the *Wo-data* is *hanataba, Ga-data* is *Mike,* and so on. Each noun is represented by a binary vector showing whether each noun occurs with each verb. The number of data items equals the number of kinds of verbs.
- ***SO-data*** *based on a collocation between subject and object.* We gathered subject Q followed by the case-marking particle <*ga*> that depends on the same verb V as the object P for each object followed by the case-marking particle <*wo*>. For the above example, we can gather the subject *Mike* for the object *hanataba* because we have the dependency relations *Mike* <*ga*> *okutta* and *hanataba* <*wo*> *okutta*. The number of data items equals the number of kinds of objects, where each of them co-occurs with a subject in a sentence and depends on the same verb as the subject.

These data are represented by a binary vector which corresponds to the appearance pattern of a noun and these vectors are used as arguments of statistical measure in step 2. Figure 1 shows images of the appearance pattern expressed by the binary vector for each data item. The number of dimensions equals the number of data items for each linguistic data. For *NN-data*, each dimension corresponds to a sentence. The element of the vector is 1 if the noun appears in the sentence and 0 if it does not. Similarly, for *NV-data*, each dimension corresponds to a verb. For *SO-data*, we represent the appearance pattern for each subject with a binary vector whose dimension corresponds to an object.

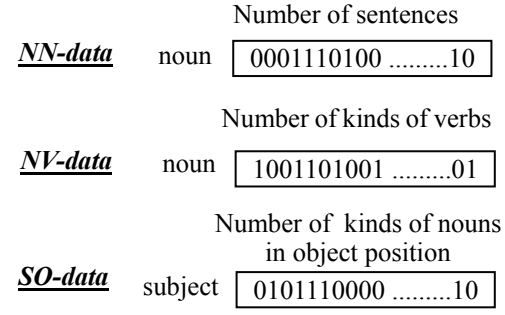

Figure 1: Appearance patterns of a binary vector for a noun in each type of linguistic data

2.2 Related Word Set Extraction Method

In step 2, in order to extract word sets that are useful for information extraction, we applied the method based on the Complementary Similarity Measure (CSM) which can measure inclusive relations between two vectors. This method can extract related words by calculating inclusive relations of the appearance pattern between two words based on the collocation relationship (modifiee/modifier relationship) in Japanese documents (Yamamoto & Isahara, 2007).

The CSM was developed as a means of recognizing degraded machine-printed text (Hagita & Sawaki, 1995). It is known that CSM can be applied to natural language processing and can determine the relation between two words in text data by estimating inclusive relations between two vectors representing each appearance pattern for each word.

We first extract word pairs having an inclusive relation of the appearance patterns between the words by calculating the CSM values. An appearance pattern is expressed as an n-dimensional binary feature vector. When $\boldsymbol{V}_i = (v_{i1}, ..., v_{in})$ is a vector for word w_i and $\boldsymbol{V}_j = (v_{j1}, ..., v_{jn})$ is a vector for word w_j, $CSM(\boldsymbol{V}_i, \boldsymbol{V}_j)$ is defined as follows:

$$CSM(V_i, V_j) = \frac{ad - bc}{\sqrt{(a+c)(b+d)}},$$

$$a = \sum_{k=1}^{n} v_{ik} \cdot v_{jk}, \qquad b = \sum_{k=1}^{n} v_{ik} \cdot (1 - v_{jk}),$$

$$c = \sum_{k=1}^{n} (1 - v_{ik}) \cdot v_{jk}, \quad d = \sum_{k=1}^{n} (1 - v_{ik}) \cdot (1 - v_{jk}).$$

CSM is an asymmetric measure. Therefore, $CSM(\boldsymbol{V}_i, \boldsymbol{V}_j)$ usually differs from $CSM(\boldsymbol{V}_j, \boldsymbol{V}_i)$ exchanged between $\boldsymbol{V}_i$ and $\boldsymbol{V}_j$. According to the asymmetric feature, we can estimate whether the appearance pattern of w_i includes the appearance pattern of w_j. The inclusive relation corresponds to the semantic relation between w_i and w_j. Extracted word pairs are expressed by a tuple <w_i, w_j>, where $CSM(\boldsymbol{V}_i, \boldsymbol{V}_j)$ is greater than $CSM(\boldsymbol{V}_j, \boldsymbol{V}_i)$ when words w_i and w_j have each appearance pattern represented by each binary vector $\boldsymbol{V}_i$ and $\boldsymbol{V}_j$.

As for the comparison between our CSM-based method and methods which were previously proposed for extraction of relations between words, Yamamoto & Umemura (2002) compared CSM with other similarity measures including Cosine and Dice functions used as comparison measures in (Dekang, 1998), and concluded that CSM is useful for determining the hypernym-hyponym relation between two words. Moreover, Yamamoto et al. (2005) compared CSM with the Overlap function and showed the usefulness of CSM for the task.

Next, we connected word pairs with CSM values greater than a certain threshold and constructed word sets. A feature of the CSM-based method is that it can extract not only pairs of related words but also sets of related words because it connects their word pairs consistently.

We connected word pairs with CSM values greater than a certain threshold and constructed word sets. Suppose we have tuples <A, B>, <B, C>, <Z, B>, <C, D>, <C, E>, and <C, F>, which are word pairs having CSM values greater than the threshold in the order of their values. For example, let <B, C> be an initial word set {B, C}. First, we find the tuple with the greatest CSM value among the tuples in which the word C at the tail of the current word set is the left word, and connect the right word behind C. In this example, word "D" in <C, D> is connected to {B, C}, making the current word set {B, C, D}. This process is repeated until no tuples can be chosen. Next, we find the tuple with the greatest CSM value among the tuples in which the word B at the head of the current word set is the right word, and connect the left word before B. This process is repeated until no tuples can be chosen. In this example, we obtain the word set {A, B, C, D}.

Finally, by using a thesaurus, we identify the word sets to which all words are taxonomically related, that is, which agree with the thesaurus. As the rest of the word sets have a non-taxonomical relation among the words, we identify them as word sets with a thematic relation.

2.3 Experimental Results

In this experiment, we used a collection of aviation safety reports in Japanese, which contained 6,427 reports from 1992 to 2003 (3.7 Mbytes). Each of the reports includes fixed information such as departure place and arrival place, and the content (including the title, the pilot's report, and the reply to the report) described in free style. We processed only the content described in free style, deleting the personal information.

In this paper we utilize the results from *NN-data* and Wo-data among all extracted data for the sake of explanation. As for *NN-data*, the number of data items is 36683, which is the number of sentences including dependency relations we used, and 42352 nouns appear in the collection. As for *Wo-data*, the number of data items is 4871, which is the number of kinds of verbs, and 9972 nouns appear in the data, where we treat a verb with a different suffix as a different verb. The number of different verbs in the data is 1983.

First, we extracted related word pairs by calculating the CSM values for all pairs of nouns appearing in each data. 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 from *Wo-data* are shown in Figure 2, where each of the last columns is the relation between the two words judged by humans. "Synonym" judged by humans also includes abbreviations. Using *Wo-data* which is *NV-data* based on the dependency relation between noun and verb for each case-marking particle <*wo* (object)>, the extracted word pairs tended to have a hypernym-hyponym relation, and so could be useful for classifying and understanding the terms. There are also included many synonyms and abbreviations. The terms used in the aviation safety reports are not controlled since the reports were written by many pilots. Therefore,

junbi (preparation)	*shuppatsu-junbi* (preparation for departure)	hyponym
junbi (preparation)	*shuppatsu-junbi* (preparation for departure)	hyponym
FLT (flight)	*hiko* (flight)	synonym
sagyo (work)	*seibi* (maintenance)	synonym
sagyo (work)	*shuppatsu-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 2: The top 10 word pairs extracted from *Wo-data*, with the relation judged by humans

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 3: The top 10 word pairs extracted from *NN-data*, with their CSM values

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

Figure 4: Examples of related word sets extracted from *NN-data*

extraction of synonyms and abbreviations is crucial to enable airline companies to develop efficient text applications such as text mining and information retrieval.

As for the thematic relations, we also extracted word pairs whose CSM values were very high, but using *NN-data*. The top 10 word pairs extracted are shown in Figure 3, with their CSM values. Words in these word pairs tended to appear in the same sentences and were thematically related.

Then, we extracted related word sets by connecting word pairs having the CSM-value over the threshold, which were set empirically. We also show some of the related word sets extracted from NN-data in Figure 4, where the threshold is 0.25 and the number of extracted word sets is 136 in this case. They seem to have a thematic relation among the terms composing each of them.

3. Extracting Related Terms from English Documents

We also try to use 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, so we used 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 Dispatch Deviations Guide in English that differs from the collection used in Section 2.3. The manuals used were the MEL/CDL Manuals (MCM), where MEL is "Minimum Equipment List," and CDL is "Configuration Deviation List." First, we parsed sentences in the documents with the 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 based on collocations between a verb and its direct object, and one based on collocations between a verb and its subject. These linguistic data correspond to *Wo-data* and *Ga-data* in Section 2.1, respectively. Then, from these English linguistic data, similar to the experiment shown in Section 2, we tried to extract the pairs of related terms with the method based on CSM in order to obtain taxonomically related terms. For

0.851329	System	Pack
0.761612	Pack	Window
0.745672	Pack	Engine
0.739358	Door	Pack
0.694023	Pack	ADP
0.683866	Pack	Autopilot_Channel
0.668298	Pack	APU
0.668044	Valve	Pack
0.650525	Position	Pack
0.639627	Pack	Compartment
0.618628	Light	Pack
0.610521	Pack	Side
0.610521	Pack	Portion
0.608480	Pack	All
0.605161	All	Air_Conditioning_Pack
0.603675	Pack	Pump
0.601260	APU	Air_Conditioning_Pack
0.598400	Pack	Fan
0.595798	Pack	Temperature_Indication
0.595798	Pack	Pump_Operation

Figure 5: Examples of related word sets extracted from *Wo-data* for English MCM

1.000000	APU	Pack
0.870120	Pack	ADP
0.757830	Pack	Pump
0.757830	Pack	Light
0.725567	Windmilling_Start	One_Pack_Operation
0.725567	Equipment_Cooling_System	Right_Pack
0.725567	One_Pack_Operation	MASTER_CAUTION_Recall
0.725567	One_Pack_Operation	Contamination_Check
0.723942	Engine_Start	One_Pack_Operation
0.723942	Performance_Adjustment	One_Pack_Operation
0.723942	ADP	Right_Pack
0.722318	Pump	Right_Pack
0.722318	Light	Right_Pack
0.722318	Fuel_Jettison	One_Pack_Operation
0.719069	*Seino-hosho* (Performance Penalty)	One_Pack_Operation
0.719069	Pack	V_NAV
0.719069	Pack	Reverse_Thrust
0.719069	Pack	MAN_Mode
0.719069	Pack	IRU
0.719069	Pack	Equipment_Cooling_System

Figure 6: Examples of related word sets extracted from *Wo-data* for Japanese MCM

comparison, we also used the Japanese MCM to examine sentences that were Japanese translations of most of the sentences in the English MCM. Figures 5 and 6 show some of the extracted word pairs including the term "pack," which appears in both English documents and Japanese documents, i.e. the English noun "pack" is used in Japanese manuals. The term "pack" in the manuals means a part of the air conditioning system. The results shown in these figures are extracted from each *Wo-data*.

The top four terms which have strong relations with "Pack" in the Japanese results, i.e. "APU," "ADP," "PUMP," and "Light," also appear in the English results, as the 5th, 7th, 11th, and 16th terms, respectively. On the other hand, the top four terms from English, i.e. "System," "Window," "Engine," and "Door," do not appear near the top of the results from Japanese. This seems to be because not all English sentences in the documents are translated into Japanese. "System" and "Engine" are included in the collocations (or compound nouns) such as "Engine Start," "Engine_Bleed_Air" and "Engine Bleed." This seems to be because of the difference of the word segmentation, that is, the difference between the English morphological analysis system and the Japanese one. Similarly, this difference would cause "Air Conditioning_Pack" to appear in the English results but not in the Japanese results, and "One_Pack_Operation" and "Right_Pack" to appear in the Japanese results only. Therefore, if the same word segmentation is applied, we could obtain similar results for various languages. This suggests that this method for extracting related words does not depend on language.

4. Extraction of Salient Nouns and Phrases

We also tried to extract salient terms including compound nouns and noun phrases from aviation documents which are written in Japanese, but contain many English words.

Our technique acquires terms from morpheme strings. There are several methods to acquire new words from a

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 "Extraction of Informative Expressions from Domain-specific Documents" in Japanese. The method uses morpheme-based n-grams to save processing time and space compared with previous character-based methods; therefore the acquired terms are compounds of one or more morphemes.

Our term acquisition method consists of two stages: 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. For example, when we guess whether "お (*o*)/ 台 (*dai*)/ 場 (*ba*)" [3] is a term, statistical indicators are used to verify the hypothesis that if "お/台/場(*odaiba*)" is a term, the kinds of morphemes following/preceding it will outnumber those following "お / 台 (*odai*)" or "台 / 場 (*daiba*)". Each process is described below in detail.

4.1 Selection of Candidates

For selecting term/phrase candidates, we considered that terms/phrases that characterize the document collection are judged by 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 documents, 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 throughout the collection, but not too widely. The idea is expressed by the *df/cf* value within the certain range, where *cf* indicates collection frequency, that is, term frequency in the collection. If both *df* and *cf* are high, the terms are distributed too widely, like function words. If both of them are low, the terms do not represent the entire collection. We do not use the number of documents composing the collection because we would like to consider the contribution made by a term that appears in a specified document repeatedly. Accordingly, we consider that terms whose *df2/df* and *df/cf* values are 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 within a range set 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 for checking unithood is Tanaka's function (Tanaka-Ishii et al., 2003), which is a variation of C-value (Frantzi & Ananiadou, 1996):

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

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, and the third term $(1-1/cd(Z))$ is the term for the number of adjacent different morphemes. We have tested variations of function (1) and found that function (2) shown below performed best.

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

Note that in function (2), the length term in (1) is eliminated and the terms for the number of different morphemes are corrected to reduce the effect of the frequency. This was applied in both directions, that is, the shortened string with the last morpheme removed and the shortened string with the head morpheme removed. We extracted as a term/phrase the candidate having the higher $F'(Z)$ value than the values for both the one-morpheme shortened strings.

4.3 Experimental Results

Using this method, we extracted the salient terms including compound nouns and noun phrases from the collection of Japanese documents. When we used the collections of aviation safety reports only, the number of extracted terms was 382, and most of them were compound nouns. We then also conducted the examination for aviation documents including not only aviation safety reports but also various kinds of manuals in Japanese. As a result, we obtained 2157 terms, which are not only compound nouns but also various kinds of long phrases.

Some of the typical results are shown in Figure 7. The words in italics were originally in Japanese and were translated into English for explanation.

The extracted terms and phrases are useful for informing users of noteworthy topics in the domain, because they indicate some concept, but cannot be written in one word. They can be used as a certain unit for machine translation, that is, they could be the entries in dictionaries for translation systems and items in translation memories.

We will examine the effects and the characteristics of the experimental results extracted from the aviation documents as an example, and evaluate their validity for guiding users toward useful information.

5. Conclusion

In this paper, we introduced two techniques for extracting informative expressions from documents: the extraction

[3] "お台場 (*odaiba*)" is a famous Japanese location, but a typical Japanese morphological analyzer extracts the result as a list of three morphemes.

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

Figure 7: Examples of typical results from aviation documents

of related words that are not only taxonomically related but also thematically related, and the acquisition of salient terms and phrases. With these techniques we then attempted to automatically and statistically extract domain-specific informative expressions in aviation documents as an example and evaluated the results. As a result, it was suggested that the informative expressions obtained by using our techniques could usefully assist information extraction by humans.

6. Acknowledgements

We sincerely appreciate the contribution of Mr. Atsushi Ikeno of Oki Electronic Co., Ltd.

7. References

Church, K. W. (2000). Empirical Estimates of Adaptation: The chance of Two Noriega's is closer to p/2 than p^2. In *Proceedings of the Eighteenth International Conference on Computational Linguistics*. pp. 180–186.

Dekan, L. (1998). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the COLING-ACL98*, pp. 768–774.

EDR Electronic Dictionary. (1995). http://www2.nict.go.jp/r/r312/EDR/J_index.html.

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

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Frantzi, K., Ananiadou, S. (1996). Extracting Nested Collocations. In *Proceedings of the Sixteenth International Conference on Computational Linguistics*. pp. 41–46.

Hagita, N., Sawaki, M. (1995). Robust Recognition of Degraded Machine-Printed Characters using Complimentary Similarity Measure and Error-Correction Learning. In *Proceedings of the SPIE – The International Society for Optical Engineering*. 2442, pp. 236–244.

Kageura, K., Umino, B. (1996), Methods of Automatic Term Recognition: A Review. *Terminology*. 3(2), pp. 259–289.

Kiyota, Y., Nakagawa, H. (2006). A Domain Ontology Production Tool Kit Based on Automatically Constructed Case Frames. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*. pp. 1482–1487.

Morris,, J., Hirst, G. (2004). Non-classical lexical semantic relations. In *Proceedings of Human Language Technology Conference of the NAACL*, *Workshop on Computational Lexical Semantics*.

Nakagawa, H., Mori, T. (2003). Automatic Term Recognition based on Statistics of Compound Nouns and their Components. *Terminology*. 9(2), pp. 201–209.

Tanaka-Ishii, K., Yamamoto, M., Nakagawa H. (2003). Kiwi: A Multilingual Usage Consultation Tool based on Internet Searching. In *Proceedings of the Interactive Posters/Demonstrations ACL-03*. pp. 105–108.

Wisniewski, E. J., Bassok, M. (1999). What makes a man similar to a tie? Stimulus compatibility with comparison and integration. *Cognitive Psychology*. 39, pp. 208–238.

Yamamoto, E., Isahara, H. (2007). Extracting Word Sets with Non-Taxonomical Relation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pp. 141–144.

Yamamoto, E., Kanzaki, K., Isahara, H. (2005). Extraction of Hierarchies based on Inclusion of Co-occurring Words with Frequency Information. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*. pp. 1166–1172.

Yamamoto, E., Umemura, K. (2002). A Similarity Measure for Estimation of One-to-Many Relationship in Corpus. *Journal of Natural Language Processing*. 9(2), pp. 45–75.