Experiments for NTCIR-8 Technical Trend Map Creation Subtask at Hitachi

Yusuke Sato
Hitachi, Ltd., Information & Telecommunication Systems Company
NOF Tameike Bldg., 1-14, Akasaka 1-Chome,
Minato-ku, Tokyo, 107-0052 Japan
+81-3-3584-9915
yusuke.sato.pr@hitachi.com

Makoto Iwayama
Hitachi, Ltd., Central Research Laboratory
1-280, Higashi-koigakubo
Kokubunji-shi, Tokyo, 185-8601 Japan
+81-42-323-1111
makoto.iwayama.nw@hitachi.com

ABSTRACT
This paper reports on an experiment to evaluate the extraction of effect expressions from patents and papers (in Japanese) at the subtask of Technical Trend Map Creation in NTCIR-8 Patent Mining Task. To obtain a more detailed structure for the expressions, we defined that effect expressions consist of TARGET, SCALE and IMPACT elements. We created training data based on these elements and assigned tags by supervised learning. Then, on the basis of conversion rules and dependency relationships, we converted these independently defined tags to the ATTRIBUTE, VALUE and EFFECT tags.

Categories and Subject Descriptors
H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing – *Linguistic Processing.*

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Information filtering.*

General Terms
Languages, Performance.

Keywords
Viewpoint extraction, Support vector machines, 3-tuple tag set, Dependency parsing

1. INTRODUCTION
A technical trend map is a type of patent map that classifies in tabular form a set of documents according to differences in viewpoints of invention such as problem to be solved, solution, and effects. Technical trend maps have been widely used as a tool for surveying technology trends among competitors. However, it can be very costly to accurately grasp the content of each and every patent in a huge collection of patents for creating a map. The need for automating such a process has been felt.

There have been several researches in recent years on methods for extracting the effect expressions of inventions in patent documents with the aim of automatically generating technical trend maps. Nishiyama et al. [1] conducted research on extracting expressions of problem to be solve and effect using a keyword list for each technology field and syntactical patterns based on sentence-ending expressions (e.g., ～[助詞]+向上する) and on using such expressions to predict the business impact of the associated technologies. Sakai et al. [2] proposed a method for collecting expressions of problem to be solved and effect by repeating alternate extraction of both expressions on the basis of their co-occurrence statistics with clue expressions.

In this paper, we have proposed an approach for extracting effect expressions on the basis of a 3-tuple syntactic structure consisting of IMPACT, SCALE, and TARGET. We created training data based on these independently defined tags and assigned tags using a support vector machine (SVM). We then chunked our tags into EFFECT tags, which are NTCIR-defined tags, using dependency relations, and also converted our tags to ATTRIBUTE and VALUE tags, which are also NTCIR-defined tags, using several rules.

This paper is organized as follows. Section 2 describes the 3-tuple tag set in detail. Section 3 describes the learning and assignment flow for each tag. Section 4 describes evaluation data and experimental results. Section 5 concludes the paper.

2. Extraction of Technical Effect Phrase Based on 3-Tuple Expression
We deconstruct an effect expression into a structure having the form “A する B を C する”, where A, B, and C are defined to be TARGET, SCALE, and IMPACT, respectively. TARGET is usually a verb or a noun expressing an action (such as a verbal noun), SCALE is a word like “speed” or “concentration,” and IMPACT is a word that modifies SCALE and TARGET.

In the sentence “重金属イオンの回収効率を向上させる”, for example, “回収効率を向上” is an effect expression. This sentence can be deconstructed into the structure “重金属イオンを回収する効率を向上させる” where TARGET is “回収”, SCALE is “効率” and IMPACT is “向上”.

There are two reasons for adopting the above definitions:

1. Each tag has more consistent grammatical elements than NTCIR-defined tags.
2. Our tags can more clearly divide effect expressions into those commonly appear across technology fields and those do not.

Here, we consider the following two examples: “<ATTRIBUTE>信頼性</ATTIBUTE>の<VALUE>高い</VALUE>” and “<ATTRIBUTE>原NIFSK 波形</ATTRIBUTE>を正確に<VALUE>復元</VALUE>する”. In the case of NTCIR-defined tags, VALUE tag is assigned to an adjective (“高い”) as well as to a noun expressing action (“復元”). In many cases, assignment is performed on the basis of morphological information, and rules and models that assign a same tag to words with different grammatical elements can become complicated, which makes learning difficult.

On the other hand, in the case of the definition of our tags, tags are assigned to above two examples as follows: "<SCALE>信頼性</SCALE>の<IMPACT>高い</IMPACT>" and "<SCALE>原 NIFSK 波形</SCALE>を正確に<TARGET>復元</TARGET>する". TARGET tags are assigned to words expressing an action (e.g, "復元") and IMPACT tags are assigned to words that modify SCALE (e.g., "高い"). Each tag is defined not to have multiple grammatical elements. Thus, by reducing the number of type of grammatical elements that correspond to each tag, the semantic structure of an effect expression can be detected finely and accurately.

In addition, as there are many technology fields in patent documents, the target of tag assignment has huge vocabulary. According to the definition of NTCIR tags, the number of vocabulary words for ATTRIBUTE is very large because it corresponds to a variety of words such as "変換出力" and "再生速度". The phrase corresponding to ATTRIBUTE can be deconstructed into an element (SCALE) corresponding to words that appear commonly across diverse technology fields and an element (TARGET) corresponding to words that do not. In short, words appearing commonly in diverse fields such as "出力" and "速度" correspond to SCALE, which means that the number of vocabulary words of the target of tag assignment should be relatively small and tag assignment should be easy. On the other hand, TARGET corresponds mostly to words unique to individual technology fields, but it should be possible to assign the TARGET tag by using words close to SCALE and IMPACT.

An effect expression does not necessarily have to consist of all three tags. For example, the structure "活性炭の<TARGET>再生</TARGET> を <IMPACT> 確 実 </IMPACT> か つ <IMPACT> 容 易 </IMPACT>に行える" has only TARGET and IMPACT.

3. Our Approach

We assigned TARGET, SCALE and IMPACT tags to effect expressions by using independently developed training data assigned these three tags manually. Models for detecting each tag are generated respectively by supervised learning. The following describes the flow of tag assignment, features for learning, and conversion of our independently defined tags to NTCIR tags.

3.1 The Flow of Tag Assignment

Tags are assigned in the order of IMPACT, SCALE and TARGET (Fig. 1). These tags appear in the same sentence in many cases and are usually close to each other. In patent documents that cover a variety of technology fields, the IMPACT corresponds to words like "向上", "高い" and "経済的" that are common to diverse technology fields, which should make assignment more accurate than other tag assignments. For this reason, tag assignment begins with IMPACT followed by SCALE and TARGET in that order. In this way, the presence of IMPACT and SCALE tags, for example, can be used as a feature in the assignment of TARGET tags. This achieves greater accuracy than independent assignment of each tag.

3.2 Features for Learning

An input sentence is deconstructed into morphemes by ChaSen [3] and tag assignment is performed targeting morphemes with following parts of speech: nouns, verbs, adjectives, adnominals, adverbs, prefixes and unknowns. The following features are used for learning with respect to the morphemes with above parts of speech.

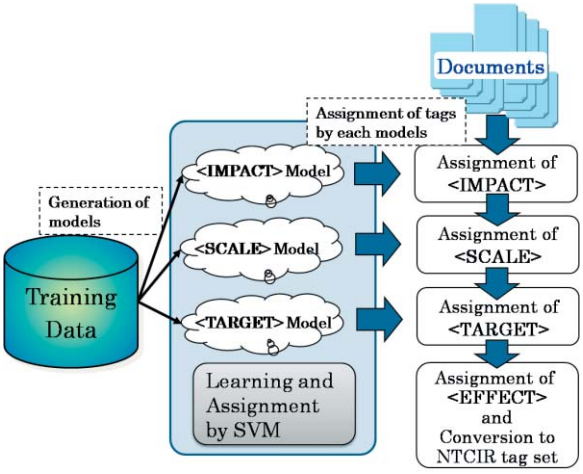

Figure 1. The flow of our tag assignment

1. Features for morpheme

These features, which are obtained by ChaSen, consist of the character string, broad classification of part of speech, detailed classification of part of speech, conjugation type and conjugated form. In addition to features on the intended morpheme, features on the two morphemes on either side of that morpheme are also used as a feature for learning.

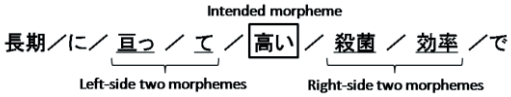

Figure 2. Example of features for morpheme

2. Features for SCALE/IMPACT dictionary

The SCALE/IMPACT dictionary consists of manually extracted words corresponding to SCALE and IMPACT. These words are extracted from a last sentence at the paragraph of effect of the invention in about 10,000 documents from about ten years worth of patent documents. Whether to match with words in this dictionary was used as a feature for learning.

3. Features for SCALE/IMPACT-expression prefix/suffix single-kanji

This is a dictionary of single-kanji (Chinese-character) prefixes and suffixes such as "耐～" and "～化" extracted manually from the above SCALE/IMPACT dictionary. Whether to match with words in this dictionary was used as a feature for learning.

4. Features for Morpheme of head in modified/modifying segment

If the segment to which an intended morpheme belongs has a modified or modifying segment, the morpheme features of head of that segment are used for learning. Here, we use CaboCha [4] as a dependency parser.

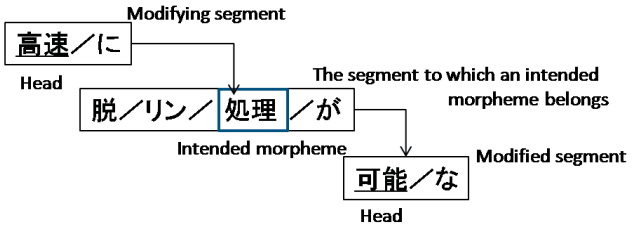

Figure 3. Example of features for Morpheme of head in modified/modifying segment

5. Features for results of IMPACT/SCALE assignment

For an intended morpheme, the results of IMPACT/SCALE tag assignments are used as features, those assignments are for two morphemes on either side of the intended morpheme or morphemes of the head in a modified/modifying segment belonging to it. For example, at the TARGET tag-assignment step, if SCALE tag has already been assigned to the immediately preceding morpheme and IMPACT tag has been assigned to the morpheme of head in a modified segment, then the feature of the presence of the tags is added in the intended morpheme.

6. Features indicating to be effect sentence

There are five features for a sentence to which an intended morpheme belongs as described below.

End-of-sentence clue-phrase match: This feature indicates whether the sentence to which an intended morpheme belongs to matches an end-of-sentence formulaic phrase like "～が可能となる". For this purpose, 235 formulaic phrases have been manually extracted.

Paragraph type: This feature indicates what type of paragraph an intended morpheme belongs to: issue, solution means, or effect.

Sentence position: Sentence position within the paragraph.

Sentence length: the number of characters included in the sentences.

Numeric character ratio within sentence: This feature quantifies the ratio between the number of numerals and characters included in the sentence. The reason for this feature is that a sentence in a solution-means paragraph tends to include many numbers corresponding to an invention's constituent elements such as "該シート吸着部１０２".

3.3 Assignment of EFFECT Tag and Conversion of Our Tags into NTCIR Tags

We specify the area for assigning the EFFECT tag by using dependency relationships between our tags, and we convert our tags to NTCIR tags by using rules according to the order of our tags.

■ <EFFECT> identification

We merge segments based on dependency relationships until segments having our independently defined tags no longer exist and take the result to be the EFFECT area. Examples are given below.

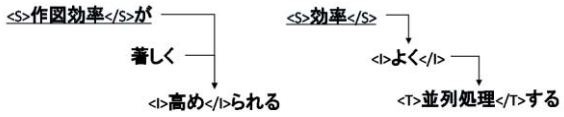

Figure 4 Example of <EFFECT> identification

■ Conversion rules

We created eight rules for converting a combination of TARGET, SCALE and IMPACT tags (<T>, <S> and <I>) to NTCIR tags VALUE and ATTRIBUTE (<V> and <A>). Examples of these rules are given below.

・{<I><S>}<T> → <V><A>

Example: <I>高</I><S>精度</S>の<T>処理</T> → <V>高精度</V>の<A>処理</A>

・<I>{<T><S>} → <V><A>

Example: <I>高い</I><T>払拭</T><S>性能</S> → <V>高い</V><A>払拭性能</A>

・{<T><S>}<I> → <A><V>

Example: <T>破断</T><S>強度</S>をより<I>大きく</I> → <A>破断強度</A>をより<V>大きく</V>

4. Evaluation

4.1 Independently Developed Training Data

We created training data by manually assigning tags on the basis of the definition of our tags. Then, on the basis of the independently developed training data, we performed learning and tag assignment for our tags, and finally assigned NTCIR tags according to the rules given in section 3.3. Table 1 summarizes the four types of training data assigned our independently defined tags. All of data have common data and each extended data.

Table 1. Independently developed training data

	Data1	Data2	Data3	Data4
Common Data	Abstracts in patent specifications ・Water-purifying technology (C02F 1/28) : 100 ・Learning and classification technology (G06F 17/30) : 98 ・Mixed data A : A61B : 10, B41J : 20, C08L : 10, D01F : 10, E02D : 10, F02D : 10, G06T : 20, H04N : 20			
Extended Data	Mixed data B B : 50 G : 50 H : 50	Mixed data B+ B : 50 G : 200 H : 200	Abstracts in Papers 200	

These four data consist of abstracts in patent specifications, which has documents for three different technology fields (water-purifying technology, learning and classification technology, and Mixed data A), Mixed data B/B+ created from the three paragraphs, problems to solve, means for the invention and effect of the invention, and abstracts in papers distributed by NTCIR1 and NTCIR2. Mixed data A consists of documents having the corresponding subclasses as main IPC code. Mixed data B/B+

consists of documents having corresponding sections as main IPC code. Mixed data B consists of data to which we have assigned the independently defined tags, while for Mixed data B+ 150 documents is added to Mixed data B, which belong to the G and H sections in IPC system and have been assigned by a non-specialist. The objective of Mixed data B+ is to evaluate accuracy, when including data with low reliability for tag assignment while this is a large volume of documents. We also used 200 abstracts in papers.

4.2 Experiment

Based on a preliminary evaluation using training data of formal-run, we determined candidates for optical combinations of features. In the end, we submitted four patent runs and two paper runs as shown in Table 2. In the table, the training data corresponds to the data shown in Table 1 and the features correspond to the same-numbered features described in section 3.2. With features #1–#3 used in all runs, we evaluated differences in accuracy based on combinations of the other features.

We used the TinySVM [5] as an SVM tool for tag learning, and decided on the use of a linear kernel based on preliminary experiments. No NTCIR-provided training data was used in the experiment. Here, we targeted only the EFFECT tag and its internal tags in tag assignment; we assigned no TECHNOLOGY tags.

Table 2. Features and training data for our submissions

	#	ID	Training data (Table. 1)	Features		
				#4	#5	#6
Patent	1	HTC_1_1	Data1	✓	✓	✓
	2	HTC_1_2		✓		✓
	3	HTC_2_1	Data2	✓	✓	✓
	4	HTC_2_2		✓		✓
Paper	5	HTC_1	Data3	✓	✓	
	6	HTC_2	Data4		✓	

4.3 Results and Discussion

4.3.1 NTCIR-8 defined tag set

Results of assignment accuracy using topics of formal-run are listed in Table 3. In the table, R, P, and F stand for Recall, Precision, and F-value, respectively, and #1–#6 correspond to the numbers in the "#" column of Table 2. No results for the TECHNOLOGY tag are listed here since it was not taken up in this experiment.

First, for patent documents, we compare the results of Data1 (#1, #2) with those of Data2 (#3, #4) and see that Data1 has a slightly higher F-value for VALUE, EFFECT, and Ave., while Data2 has a higher F-value for ATTR. This result indicates that high quality is required in training data because the accuracy for Data2 was not improved in spite of having three times the number of documents.

Next, accuracy for papers was low compared to patents. Various factors can be considered for this result: for examples, VALUE tags in paper at NTCIR tend to have numerical expressions and it is hard to assign the tags to such expressions or there were few sentences that match the end-of-sentence clue-phrase in effect sentences.

Table 3. Assignment accuracy of NTCIR-defined tags

		Patent				Paper	
		#1	#2	#3	#4	#5	#6
ATTR.	R	25.1%	24.1%	24.7%	23.7%	14.9%	11.5%
	P	24.1%	23.6%	28.2%	27.3%	16.4%	11.1%
	F	24.6%	23.9%	26.3%	25.4%	15.6%	11.3%
VALUE	R	58.0%	57.2%	52.1%	50.8%	20.7%	23.8%
	P	43.4%	43.2%	46.2%	45.5%	21.0%	20.6%
	F	49.6%	49.2%	49.0%	48.0%	20.9%	22.1%
EFFECT	R	16.4%	15.5%	15.3%	14.5%	5.5%	5.8%
	P	22.3%	21.7%	23.6%	22.8%	11.2%	9.9%
	F	18.9%	18.1%	18.6%	17.7%	7.3%	7.3%
Ave.	R	23.3%	22.7%	21.5%	20.9%	10.0%	10.0%
	P	34.6%	34.4%	38.0%	37.3%	18.8%	16.1%
	F	27.8%	27.4%	27.5%	26.8%	13.1%	12.3%

4.3.2 Our independently defined tag set

We show the assignment accuracy of our independently defined tags. We used Data1 and abstracts in 200 papers in Table 1 and all features #1-6 in section 3.2. The accuracy was estimated by leave-one-out cross-validation. In this case, although a single fold should correspond to a single morpheme, we set a single document as a single fold in order to reduce time of repeating estimation by all of morphemes. In other words, if we have 100 documents, the process that all of morphemes in 99 documents are used as training data and all of morphemes in the other document are used as test data is repeated 100 times.

The accuracy of assignment for patent and paper is shown in Table 4.

Table 4. Assignment accuracy of independently defined tags

		Patent (Data1)	Paper (Abstracts in 200 papers)
TARGET	R	45.0%	7.9%
	P	58.7%	19.6%
	F	50.9%	11.3%
SCALE	R	54.3%	19.5%
	P	63.4%	33.8%
	F	58.5%	24.7%
IMPACT	R	64.9%	28.0%
	P	68.4%	38.4%
	F	66.6%	32.4%

As expected, the accuracy of TARGET, for which there are relatively few words common to diverse fields, is low. We can expect to improve the accuracy by combining our approach with other methods, such as increasing the weight of words having a high degree of co-occurrence with SCALE and IMPACT. Assignment of tags for papers was harder than that of patent because of numerical expressions or end-of-sentence clue-phrase mentioned in 4.3.1.

4.3.3 Relation between NTCIR-defined tags and our defined tags

Our results reveal that the accuracy of assignment of ATTRIBUTE tag is lower than that of VALUE. One reason is that our independently defined tags are defined to be assigned to a single morpheme such as "ギヤモータの <TARGET> 用途 </TARGET><SCALE> 範囲 </SCALE> を <IMPACT> 拡げる </IMPACT>" while Assignment of ATTRIBUTE tag is often made for long phrases "<ATTRIBUTE>プログラム作成上の制限</ATTRIBUTE>を <VALUE>解消</VALUE>する". We expect to improve accuracy through the use, for example, of sequence tagging methods as used in named entity extraction.

Our conversion rules were lack of flexibility. When some of independently defined tags in an effect expression were assigned incorrectly, these tags were not fitted to a conversion rule and then could not be converted to NTCIR defined tags correctly. For example, an effect expression "破断開始時期を早める" is assigned our tags as "<TARGET>破断</TARGET>開始<SCALE>時期 </SCALE> を <IMPACT> 早 め る </IMPACT>" and is converted to NTCIR defined tags as "<ATTRIBUTE>破断開始時期</ATTRIBUTE> を <VALUE> 早 め る </VALUE>" by a rule {<T><S>}<I> → <A><V>. However, if the morpheme "破断" is assigned IMPACT tag incorrectly, tags are not matched the rule. Thus a method to apply a conversion rule stochastically is needed.

5. Conclusion

We defined that expressions which indicate the effect of an invention consist of a certain combination of TARGET, SCALE and IMPACT elements and assigned NTCIR tags on the basis of these three elements. TARGET is a verb or a noun expressing action, SCALE is a word like "speed" or "concentration" and IMPACT is a word that modifies SCALE and TARGET. We assigned our tags in the order of IMPACT, SCALE and TARGET because it is easier to detect them in order of having words common to many technical fields of patents. In addition, we converted our three independently defined tags to NTCIR defined tags using conversion rules and dependency relationships.

Our independently defined tags target short phrases while NTCIR correct tags tend to target long phrases. This has been a source of assignment error. In response to this problem, we plan to study the effects of combining our approach with other methods such as sequence tagging methods.

6. REFERENCES

[1] Risa Nishiyama, Hironori Takeuchi, Hideo Watanabe, Tetsuya Nasukawa, Technology Survey Assistance Tool Focusing on Their Advantages, Journal of Japanese Society for Artificial Intelligence, vol.24, no.6, pp.541-548, 2009. (in Japanese)

[2] Hiroyuki Sakai, Hirohumi Nonaka, Shigeru Masuyama, Extraction of Information on the Technical Effect from a Patent Document, Journal of Japanese Society for Artificial Intelligence, vol.24, no.6, pp.531-540, 2009. (in Japanese)

[3] ChaSen:http://chasen-legacy.sourceforge.jp/

[4] CaboCha:http://chasen.org/~taku/software/cabocha/

[5] TinySVM:http://chasen.org/~taku/software/TinySVM/