mantic characteristics is shown in Table 1. Types listed here are mostly based on t n e96 , and the percentages of each type were derived from sample television news

Table 1: Types of television news captions. Numbers cited in ‘Grammatical Characteristics’ refer to those listed in 2.1.1.

	Types	Ratio	Grammatical Characteristics
(a)	Locational, Organizational	30%	5, 6
(b)	Personal	15%	5, 6
(c)	Title	14%	1 - 6
(d)	Speech (Include summaries and translations)	10%	1, Colloquial
(e)	Temporal	7%	6
(f)	Technical (live’)	3%	6
(g)	Descriptive	2%	2, 3, 5
(h)	Others	19%	1 - 6

programs, where 464 captions appeared in total. Grammatical characteristics associated with semantic characteristics are also cited in the table.

Among these types, (a) and (b), which cover 45% of the captions, represent simply what is actually shown in the image. This allows them to almost directly become candidate keywords. Nevertheless, when attempting to index conscientiously, it is important to analyze the semantics of the candidates. In this case, captions should be analyzed whether they are (1) personal, (2) locational / organizational, or (3) others.

Although (c) does not necessarily reflect the graphical contents of the video, it is an important piece of information which explains the topic. Thus, occupying nearly 60% of the captions, (a), (b) and (c) could be directly used as candidate keywords.

2.2 Semantic Analysis based on Suffixable Nouns

As related researches, several methods to analyze nouns that satisfy our purpose to some extent exist.

Similar to our task, the ‘Named Entity task’ defined for the Message Understanding Conference (MUC) [NEtask] assigns the participants to classify personal, organizational, locational, temporal and numerical phrases. The difference is that our aim is not limited to proper nouns, while the Named Entity task limits the tagging to personal, organizational and locational phrases to proper names.

Nasukawa [Nasukawa 97] proposed a semantic disambiguation method that determines semantic attributes of proper nouns (whether a proper noun indicates a place or a person) referring to the context of neighboring sentences. On the other hand, Watanabe [Watanabe96] proposed a method that analyzes television news captions by referring to both location and grammatical characteristics.

Although these methods perform fairly well, Nasukawa’s method is difficult to serve our purpose since (1) captions do not have enough neighboring information to analyze contexts, and (2) it is mainly purposed to handle proper nouns. Watanabe’s method does serve our purpose, but is not generally applicable to various news programs, which have different designing policies where layouts of captions vary.

Considering these issues, we decided to analyze captions on their own. As an example, it is impossible even for a human being to distinguish whether ‘橋本’ (Hashimoto) alone indicates a person or a place or even a common noun, which means ‘the foot of a bridge’. Then how do we distinguish them? We know that ‘橋本-首相’ (Prime Minister Hashimoto) indicates a person and ‘橋本-市’ (Hashimoto City) indicates a place from the suffixable nouns, ‘首相’ (Prime Minister) and ‘市’ (City), respectively.

As shown in this example, it is generally possible in Japanese (and also presumably in other East Asian languages) to analyze semantic attributes of noun phrases referring to suffixable nouns. Accepting this assumption, we will analyze captions that are mostly noun phrases, by referring to suffixable nouns.

Although in this example and also in the following sections, issues on suffixable nouns are mainly discussed, our method also includes analysis to individual nouns such as ‘volunteer’ and ‘closet’ that on their own indicate people and places, respectively.

3 Collection of Suffixable Nouns from Text Corpora

First of all, we need to know which suffixable noun determines the semantic attributes, (1) personal or (2) locational / organizational, of the whole noun phrase. There are several renowned concept oriented Japanese dictionaries such as the EDR electronic word / concept dictionaries [EDR] and the Classified Lexical Table [Lexical], but these are not structured in the way that satisfy our purpose, in the sense that they are not organized considering the use of words. This makes it difficult to determine which concept class in these dictionaries contains the appropriate suffixable nouns. So, we first collected such nouns automatically from two text corpora based on simple rules.

The following two corpora we used are morphologically analyzed manually beforehand, which ensures the basic reliability of the collection process.

C
Consists of approximately 27,000 sentences selected from Mainichi newspaper’s 1994 edition.

K C
Consists of approximately 20,000 sentences from the Mainichi newspaper’s 1995 edition. Strictly speaking, morphological analysis was carried out first automatically using the Japanese morphological analysis system JUMAN [JUMAN] and the Japanese structural analysis system KNP Parser [KNP], and then corrected manually.

Both of these corpora consist of texts from newspapers, which should include vocabularies similar to those used in television news captions.

Although these contain quite an amount of data, the number of the collected nouns is still relatively small. We expanded the vocabulary by gathering nouns from the same concept classes (consists of 10 words on an average) of the Classified Lexical Table (Enlarged Edition) in which the collected nouns belong. This expansion was done based on the assumption that although the structure of the table does not satisfy our purpose, each concept class should contain similar words.

In this Chapter, we will introduce the methods and the results of the collection of the two groups of nouns.

Table 2: Result of personal noun collection. The top twenty frequent nouns among the 1,630 collected nouns (include overlaps) are listed.

Rank	Noun		Frequency	Rank	Noun		Frequency
1		(person)	160	11		(congressman)	28
2		(children)	100	12		(chief)	26
3		(I)	91	13		(executive)	21
4	者	(person)	76	14	教授	(professor)	18
5	氏	(Mr.)	57	14	首相	(prime minister)	18
6	さん	(Mr./Ms.)	53	16	学生	(student)	17
7	自分	(myself)	36	16		(housewife)	17
8		(chairman)	32	16		(women)	17
9		(athlete)	31	19		(-er as in photograph ’)	15
10		(children)	30	19		(corporate president)	15

Table 3: Result of locational / organizational noun collection. The top twenty frequent nouns among the 4,861 collected nouns (include overlaps) are listed.

Rank	Noun		Frequency	Rank	Noun		Frequency
1	市	(city)	302	11		(meeting)	51
2		(inside)	161	12		(language)	41
3		(person)	142	13		(fight)	40
4		(ward)	123	14		(station)	39
5		(side)	111	15		(state)	37
6		(prefecture)	85	16		(between)	36
6		(town)	85	16		(region)	36
8		(conference)	80	18		(countries)	35
9		(government)	58	19		(problem)	34
10		(inside the city)	53	20		(army)	33

3.1 Collection of Personal Nouns

In this section, the method and the result of collecting personal nouns”; suffix able nouns that indicate people like Mr. —’ and doctor’, are discussed.

3.1.1 Collection Rules

Personal nouns were collected by the following rules:

1. Search for specific suffixes and that only succeed personal nouns
2. If the preceding morpheme is a common noun or suffix, collect it as a personal noun”

Following this rule, from the following phrase, 首相’ (Prime Minister) a Common Noun will be collected.

橋本 (Hashimoto) Proper Noun
- 首相 (Prime Minister) Common Noun
- ” (=) Suffix

Note that proper nouns are not collected in rule 2.

3.1.2 Result

As a primary result, 343 different personal nouns were collected including and used as seeds. The top twenty nouns are shown in Table 2. Among these, 287 were recorded in the Classified Lexical Table, and the remaining 56 were not. The recorded 287 were classified in 251 sub-classes under 102 classes. 3,737 nouns were recorded in these sub-classes, and as a final result, a total of 3,793 personal nouns were collected.

3.2 Collection of Locational / Organizational Nouns

In this section, the method and the result of collecting locational nouns” and organizational nouns”; suffix able nouns that indicate locations and/or organizations like —station’ and kitchen’, are discussed.

3.2.1 Collection Rules

Locational / organizational nouns were collected by the following rules:

1. Search for a place name (proper noun)[1]
2. Break down succeeding nouns continuously
3. When the chain of nouns ends, collect the last common noun if the succeeding morpheme is one of the following post-position particles that are related to locations: (from), (at), (to), (from), and (at)

Following this rule, from the following phrase, 市’ (City) a Common Noun will be collected.

Note that, again proper nouns are not collected in rule 3.

---

[1] This information is available from the corpora.

橋本	(Hashimoto)	Proper Noun
-市	(City)	Common Noun
-で"	(= at)	Post-positional Particle

3.2.2 Result

As a primary result, 1,030 different locational / organizational nouns were collected. The top twenty nouns are shown in Table 3. Among these, 930 were recorded in the Classified Lexical Table, and the remaining 100 were not. The recorded 930 were classified in 1,049 sub-classes under 399 classes. At this point, we excluded all the sub-classes that were assigned to personal nouns in the previous experiment (for example, sub-class derived from 人' (person) in Table 3). This ended in reducing the sub-classes to 926 under 379 classes. In these sub-classes, 11,070 nouns were recorded, and as a final result, a total of 11,166 locational / organizational nouns were collected.

The exclusion may delete potential locational / organizational noun classes, but the mal-effect from the wrongly collected classes seems to be more harmful.

4 Evaluation

In order to evaluate the suffixable nouns collected in the previous Chapter, we applied each dictionary to actual television news programs. Five 20 minute news programs were used, in which 587 captions appeared.

4.1 Caption Analysis Procedure

For each caption:

1. Apply the Japanese morphological analysis system JUMAN JUMAN to it. If the utmost tail morpheme is analyzed as person's, place, or organization name at this point, the analysis terminates here.
2. Else if the utmost tail morpheme is a common noun, look up for it in the dictionaries collected in the last Chapter to distinguish its semantic attribute.

For example, JUMAN will analyze a phrase as follows:

田中	(Tanaka)	Person's Name
-東京	(Tokyo)	Place Name
-大学	(university)	Common Noun
-教授"	(professor)	Common Noun

This does not give us any information whether the phrase indicates a person, location / organization or something else.

On the other hand, the proposed method looks up for the last noun of the above result in the collected dictionary. Since the collected personal noun dictionary has the noun 教授' (professor), the whole phrase is considered as a personal noun phrase.

4.2 Results

The result of the analysis is shown in Tables 4 and 5. The true answers were given by a third person, and was compared with the reported answer to evaluate the result.

Precision and recall are defined as follows:

$$Precision = \frac{N_{Correct}}{N_{Reported}}$$

$$Recall = \frac{N_{Correct}}{N_{TrueAnswer}}$$

Table 4: Result of Caption Analysis (personal).

	Proposed Method Alone	JUMAN Alone	Combined
$N_{Correct}$	63	16	79
$N_{Reported}$	75	24	99
$N_{TrueAnswer}$	—	—	83
Precision	84.0%	66.7%	
Recall	(75.9%)	+ (19.3%)	=

Table 5: Result of Caption Analysis (locational / organizational).

	Proposed Method Alone	JUMAN Alone	Combined
$N_{Correct}$	80	60	140
$N_{Reported}$	176	61	237
$N_{TrueAnswer}$	—	—	160
Precision	45.5%	98.4%	
Recall	(50.0%)	+ (37.5%)	=

In the Tables, Proposed Method Alone" indicates the result derived solely from the collected noun dictionary, and JUMAN Alone" indicates the proper nouns analyzed by JUMAN as person's, place, or organization names. Since it is difficult to count the true answers for each method separately, individual recalls are calculated as how much each method contributed to the overall recall.

4.3 Examination

Misrecognition and oversight are due to the following reasons:

1. (JUMAN)
   There are proper nouns that belong to both categories in JUMAN proper noun dictionary.
2. (Proposed method)
   Inappropriate nouns that sneaked in during the collection process, due to the incompleteness of the collection rules, and to the mismatch of the classification policy of the Classified Lexical Table with our purpose. This is prominent in locational / organizational nouns.
3. (Proposed method)
   Some nouns were essentially applicable to both categories.
4. (JUMAN)
   JUMAN's proper noun dictionary did not have enough vocabulary.
5. (Proposed method)
   Our dictionary did not have enough vocabulary either.

Among these, we will have to wait for the upgrade of JUMAN's dictionary to solve problem 4, but problem 2 will be solved by reconsideration of the collection rules, and problem 5 by collecting nouns from larger corpora. It is difficult to solve problems 1 and 3, since they arise from diversity; an essential property of words.

5 Conclusion

We proposed and evaluated an automatic television caption semantic analysis method in this paper. First, we showed that a large number of captions could be used almost directly as keyword candidates by analyzing captions manually. Next, we collected suffix table nouns that could be used as keys when analyzing semantic attributes of noun phrases, and accordingly, of most captions.

As a result, although there is still room for improvement of the dictionaries, the evaluation shows fairly good results.

The proposed method is a basis for using captions as candidate keywords for automatic indexing. We have been applying the method to actual indexing [Ide98], which shows limited but promising results.

Acknowledgments

The Japanese Morphological Analysis System JUMAN is a free software developed at and distributed by Nagao Lab., Kyoto University and Matsumoto Lab., Nara Institute of Science and Technology. Kyoto University Text Corpus is a product of Nagao Lab., Kyoto University.

The Classified Lexical Table (Enlarged Edition) was used for monitoring under a licensed agreement with National Language Institute of Japan. RWC text database is a product of Real World Computing Partnership (RWCP), and was used under a licensed agreement.

References

[Ide98] Ide, I., Hamamoto, K., and Tanaka, H.; “Automatic Video Indexing Based on Shot Classification”; , Nov 1998 ( ).

[Nasukawa97] Nasukawa, T.; “Keyword Categorization Based on Discourse Information”; , pp.348-349, Jun 1997 ( ).

[Hauptmann97] Hauptmann, A. G. and Witbrock, M. J. “News-on-Demand: Using Speech Recognition to Create a Digital Video Library”; , pp.120-126, May 1997.

[Hisamitsu97] Hisamitsu, T. and Niwa, Y.; “Acquisition of Personal Names from Newspaper Articles by Lexical Knowledge and Co-occurrence Analysis”; 97-NL-118, Vol.97, No.29, pp.1-6, Mar 1997 ( ).

[Smith97] Smith, M. A. and Kanade, T.; “Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques”; CMU-CS-97-111, Feb 1997.

[Watanabe96] Watanabe, Y., Okada, Y. and Nagao, M.; “Semantic Analysis of Telops in TV Newscasts”; 96-NL-116, Vol.96, No.89, pp.107-114, Nov 1996 ( ).

[Wactlar96] Wactlar, H. D., Kanade, T., Smith, M. A. and Stevens, S. M.; “Intelligent Access to Digital Video: Informedia Project”; , Vol.29, No.3, pp.46-52, May 1996.

[NEtask] United States Defense Advanced Research Projects Agency (DARPA), Information Technology Office; “Named Entity Task Definition, Version 2.1”; , pp.317-332, Nov 1995.

[KNP] Kurohashi, S. and Nagao, M.; “Japanese Structural Analysis System KNP version 2.0 beta6”; Jul 1998 ( ).
Available from http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/knp-e.html

[KUC] Kurohashi, S., Saito, . and Nagao, M.; “Kyoto University Corpus version 2.0”; Jun 1998 ( ).
Available from http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/corpus.html

[JUMAN] Kurohashi, S. and Nagao, M.; “Japanese Morphological Analysis System JUMAN version 3.5”; Mar 1998 ( ).
Available from http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman-e.html

[EDR] Japan Electronic Dictionary Research Institute, Ltd.; “EDR Electronic Dictionary”; Mar 1996 ( ).

[Lexical] National Language Research Institute of Japan; “The Classified Lexical Table (Enlarged Edition) Monitoring Version”; Mar 1996 ( ).

[RWC] Real World Computing Partnership (RWCP); “RWC Text Database”; Mar 1996 ( ).

[Informedia] “The Informedia Project”; http://www.informedia.cs.cmu.edu/.