Automatically Extracting Simple Auxiliary Phrases from a Corpus

Hiroyuki Shinnou
Ibaraki University.
Dept. of Systems Engineering
4-12-1 Nakanarusawa,
Hitachi, Ibaraki, 316, Japan
Phone: 0294-35-6101 (ext. 421)
E-mail: shinnou@etlron.etl.go.jp

Hitoshi Isahara
Electrotechnical Laboratory.
Machine Understanding Division
Natural Language Section
1-1-4 Umezono,
Tsukuba, Ibaraki, 305 Japan
Phone: 0298-58-5925
E-mail: isahara@etl.go.jp

Abstract

A new method is proposed in this paper to extract simple auxiliary phrases automatically from a Japanese corpus. A simple auxiliary phrase is a kind of idiomatic expressions to act as an auxiliary verb or a postpositional particle. The point of the proposed method is to utilize the three heuristics concerning simple auxiliary phrases. The method has an advantage to be carried out without a dictionary, a grammar and a syntactic parser. An experiment was performed on the newspaper articles for one month.

Key words: automatic extraction, idiomatic phrase, corpus

1 Introduction

This paper presents a new method to extract simple auxiliary phrases automatically from a Japanese corpus. A simple auxiliary phrase is a kind of idiomatic expression to act as an auxiliary verb or a postpositional particle. In Japanese, it is called *"fuzokugo-teki-hyougen"* [1] (e.g.,"にかんして" , "なければならない")[1].

In case of analyzing a Japanese sentence, first the sentence has to be divided into words, then its syntactic structure is built by the given grammar. But, for the simple auxiliary phrase, it is not efficient to divide the phrase into words.

The meaning of a simple auxiliary phrase cannot be constructed by combination of the meanings of each word, by which the phrase is composed. In addition, a simple auxiliary phrase is idiomatic expression, that is, in which the order of words cannot be changed, and other words cannot be inserted. Thus, it is no use to divide a simple auxiliary phrase into words. It should be more effective to handle a simple auxiliary phrase as one word. Therefore, it is necessary to include simple auxiliary phrases in dictionary as lexical entries.

However, it is difficult to distinguish these simple auxiliary phrases from other general phrases. Each time a system is designed, the distinction between simple auxiliary phrases and general phrases has been subjectively defined at a great cost. Moreover, it is unclear whether the phrases were collected consistently and whether all of the simple auxiliary phrases were extracted.

To overcome the problems mentioned above, we propose a new method to extract simple auxiliary phrases automatically from a corpus. This automatic extraction ensures objective and consistent selection of simple auxiliary phrases. Using a large corpus, all necessary phrases would be extracted.

First, we define σ-sequence as follow.

σ-sequence

> σ-sequence consists of HIRAGANA characters and KANJI characters, but a KANJI character cannot be followed by another KANJI character.

[1] We put *"fuzokugo-teki-hyougen"* into "auxiliary phrase", but it may not be suitable translation.

The point of our method is to utilize the following three heuristics concerning simple auxiliary phrases:

(H1) A simple auxiliary phrase is a σ-sequence.

(H2) Characters which precede or follow a simple auxiliary phrase are limited.

(H3) The words which compose a simple auxiliary phrase are strongly connected.

In our method, all σ-sequences whose length is N are extracted form a corpus, where $N \geq 4$. The set of sequences obtained by this operation is named SET-A. In view of (H1), all simple auxiliary phrases must exist in SET-A. Next, using (H2) and (H3), the sequences which are not simple auxiliary phrases are removed from SET-A. Lastly, redundant phrases are removed,i.e., if $P1$ and $P2$ belong to SET-A, and $P1$ is a subsequence of $P2$, then $P1$ is removed from SET-A. As a result, we can acquire simple auxiliary phrases(Fig.1).

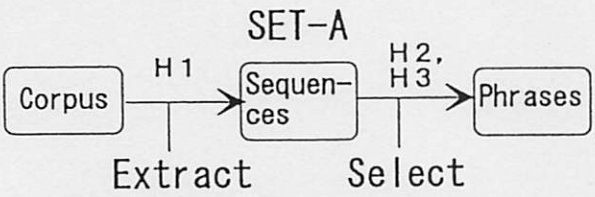

Fig.1 Extraction of simple auxiliary phrase

The proposed method has an advantage to be easily carried out. Recently, many researches on "knowledge acquisition from corpora" have been done[2]. However, most of them are based on a high-cost analysis environment, or a large tagging data (for example [3]). Thus, these method cannot be performed under other environments, and it is hard to expand to a large scale experiment. On the other hand, the proposed method is very simple. Moreover, it can be implemented without a dictionary, a grammar and a syntactic parser, since the proposed three heuristics (H1–H3) are related only to the character type in Japanese.

An experiment of this method was performed on the articles of ASAHI-SHINBUN for one month(about 9Mbyte). We report the result of this experiment in a latter section.

2 Automatic extraction of simple auxiliary phrases

2.1 Classification of auxiliary phrases

Auxiliary phrases are divided into two groups. One is "simple auxiliary phrase". In this group the words which compose the phrase are strongly connected and this phrase can be regarded as one word. The other is "complex auxiliary phrase". In this group the words which compose the phrase may not necessarily appear consecutively [2]. In other word, it is possible to insert other words in a complex auxiliary phrase but not in a simple auxiliary phrase.

Further, simple auxiliary phrases are divided into two groups. One is "relational phrase" to act as a postpositional particle. The other is "auxiliary statement phrase" to act as an auxiliary verb.

This paper deals with only simple auxiliary phrases(Fig.2).

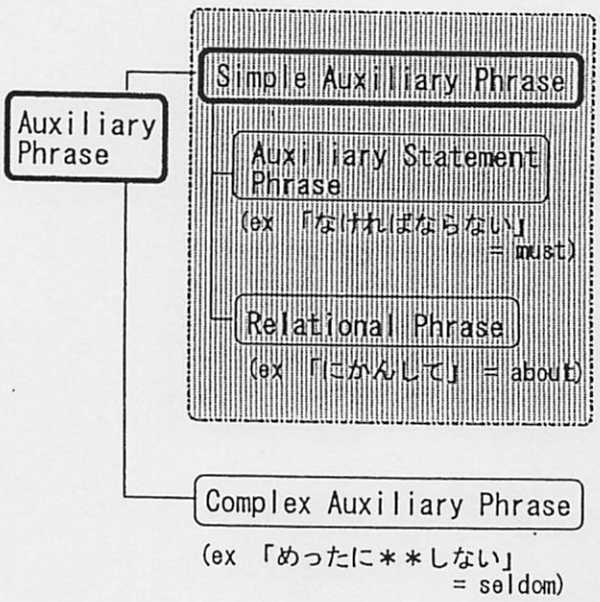

Fig.2 Classification of auxiliary phrase

2.2 Extraction by character types

By using (H1), all sequences which may be a simple auxiliary phrase could be picked up from a corpus. To explain concretely, we pick up all σ-sequences whose length are N (≥ 4) from a corpus. In view of

[2] "simple" and "complex" are the translations of " *ichigosei* " and "*tagosei*", respectively.

[EX1]

その軍事機密に関して、お互いが譲歩し合ったことは、核時代のやむを得ぬ要請であったとはいえ、同時に国家主権の不可侵性に対する修正の可能性を含むものである。

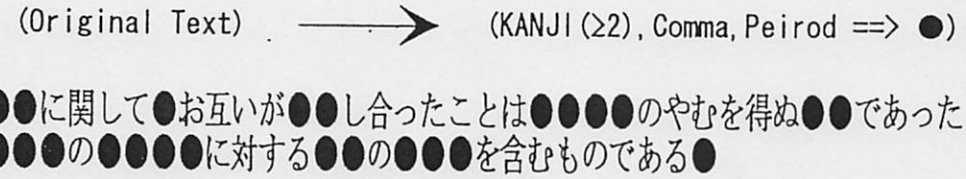

Fig.3 Effect of (H1)

(H1), all simple auxiliary phrases must exist in the σ-sequences picked up.

For example, let us apply the above operation to [Ex1](Fig.3). The following sequences for each length N ($= 4, 5, 6$) are extracted by the operation from [EX1].

$N = 4$ ... set-1

に関して，お互いが，し合った，合ったこ，ったこと，たことは，のやむを，やむを得，むを得ぬ，であった，あったと，ったとは，たとはい，とはいえ，に対する，を含むも，含むもの，むもので，ものであ，のである

$N = 5$ ... set-2

し合ったこ，合ったこと，ったことは，のやむを得，やむを得ぬ，であったと，あったとは，ったとはい，たとはいえ，を含むもの，含むもので，むものであ，ものである

$N = 6$ ... set-3

し合ったこと，合ったことは，のやむを得ぬ，であったとは，あったとはい，ったとはいえ，を含むもので，含むものであ，むものである

In this example, N cannot be greater than 8. The above operation is applied to all sentences in a corpus. The set of sequences extracted by these operations is named SET-A. As mentioned above, in view of (H1), all simple auxiliary phrases are in these sequences. On the matter of fact, "に関して" , "に対する", and "ものである" used in [Ex1], which seem to be simple auxiliary phrases, belong to set-1 or set-2.

2.3 Selection by characters around the phrase

SET-A includes all simple auxiliary phrases, but also includes a large number of junk sequences. Next, we remove these junk sequences from SET-A by using (H2).

2.3.1 Selection of relational phrase

The word which directly precedes a relational phrase must be a noun, because this phrase acts as postpositional particle. Moreover, the word which directly follows a relational phrase must be a noun, a verb, or a comma. In Japanese documents, most nouns are written with KANJI characters. Further, the first character of a verb is generally a KANJI character. These characteristics imply that if a sequence α is a relational phrase, a sequence of the form "C α" or "α C" (where C is a HIRAGANA character) rarely appears in SET-A. By applying this heuristics to all sequences in SET-A, it can be judged whether a sequence is a relational phrase or not.

Note that it is not necessary to search "C α" or "α C" in the whole corpus, but only in SET-A, because these sequences are a σ-sequence.

2.3.2 Selection of auxiliary statement phrase

In most cases, the character which directly follows an auxiliary statement phrase is a period or a KANJI character, because this kind of phrase is either placed at the end of a sentence or a modifier which precedes a noun. This means that if a sequence α is an auxiliary statement phrase, a sequence of the form "α C" (where C is a HIRAGANA character) rarely appears in SET-A.

Since an auxiliary statement phrase acts as auxiliary verb, the word which directly precedes an auxiliary statement phrase is generally a verb, a adjective, or adjectival verb. These words (verbs, adjectives, and adjectival verbs) inflect, and each inflected form ends in a HIRAGANA character. For example, MIZEN form [3] of the Japanese verb "行く" ends in a HIRAGANA character "か" or "こ" (Fig.4).

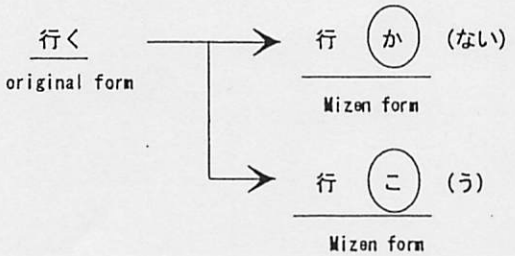

Fig.4 Inflection of "行く"

Each auxiliary statement phrase requires a specific inflectional ending. Thus, not all HIRAGANA characters can be appeared in "C α" above. For example, the auxiliary statement phrase "なければならない" requires a MIZEN form. HIRAGANA characters which can directly precede the auxiliary statement phrase "なければならない" are listed below.

か	こ	が	ご	さ	そ	た	と	ば	ぼ	ま
も	ら	ろ	わ	お	い	き	ぎ	じ	ち	に
ひ	び	み	り	え	け	げ	せ	ぜ	て	で
ね	へ	べ	め	れ	こ	し	さ	せ	ろ	

[3] In Japanese, there are six inflection forms(i.e. MIZEN, RENYOU, SHUUSHI, RENTAI, KATEI, and MEIREI form).

This set of HIRAGANA characters is named MIZEN_H_SET. Similarly, RENYOU_H_SET, SHUUSHI_H_SET, RENTAI_H_SET, KATEI_H_SET, and MEIREI_H_SET are built up.

The following operation is preformed to extract auxiliary statement phrases from the whole sequences in SET-A.

First, pick up a sequence of the form "C α" from SET-A. If C belongs to X_H_SET, increment X counter by 1 (X is one of MIZEN, RENYOU, SHUUSHI, RENTAI, and MEIREI).

If the α above is an auxiliary statement phrase, the number of its occurrences should be nearly equal to the value of one of the X counters.

By applying these rules to α in SET-A, it is judged whether that α can be an auxiliary statement phrase or not.

The sequences which not only belong to SET-A but also satisfy the restrictions described in this section form a set called SET-B.

2.4 Selection by the connection between characters

SET-B still contains many phrases which cannot be regarded as simple auxiliary phrases. For example, "を招いた" is a σ-sequence and often locates between 2 KANJI characters. However, it is hard to regard the phrase "を招いた" as a simple auxiliary phrase.

In this section, we explain the method to remove sequences which cannot be regarded as a simple auxiliary phrase from SET-B, by using (H3).

The heuristics (H3) is based on the same idea explained in [4]. Both methods use the fact that if a sequence represents an idiomatic expression, such sequences must appear repeatedly in a corpus. In addition to this idea, the method proposed in this paper adopts the heuristics (H1) and (H2), as this method is specialized in extracting simple auxiliary phrases.

In view of (H3), it is clear that if a phrase γ

$$\gamma = a_1 a_2 \cdots a_n$$

(where a_i is a character)

is a simple auxiliary phrase, the connection between a_1 and a_2 is strong. So is a_{n-1} and a_n.

length	number of different sequence	total number of sequences	most frequent sequence
4	161026	585379	している (4765)
5	195762	423820	されている (893)
6	196833	310422	になっている (491)
7	174064	226526	なければならな (346)
8	142082	165801	なければならない (327)
9	110113	120632	ことを明らかにした (161)
10	82730	87369	ることを明らかにした (76)
11	60799	62925	いることを明らかにした (30)
12	43927	44909	ていることを明らかにした (27)
13	31333	31782	での主なやりとりは次の通り (23)
14	22183	22387	当たるものであってはならない (5)

Table.1 Extraction by character types

Suppose

$$\gamma' = b_1 a_2 \cdots a_n$$

and

$$\gamma'' = a_1 a_2 \cdots a_{n-1} b_n$$

(where $a_1 \neq b_1, a_n \neq b_n, b_1, b_n$ are characters)

If γ is a simple auxiliary phrase, there are few γ' and γ'' in SET-A. This is the third criterion of the selection.

In the above example "を招いた", both of "彼が彼女を招いた" and "彼女を彼が招いた" are grammatical and equally plausible. Thus, the character which directly precedes "招いた" will not be restricted to one. So "を招いた" cannot be regarded as a simple auxiliary phrase.

The sequences which not only belong to SET-B but also satisfy the restriction described in this section form a set called SET-C.

2.5 Removal of redundant phrases

There are still some junk sequences in SET-C. Most of them are subsequences of other sequences in SET-C.

For example, if the simple auxiliary phrase "から見ると" exists in SET-C, the junk sequence "ら見ると" would also exist in SET-C, because the character which directly precede the sequence "ら見ると" is mostly the character "か" which is a member of MIZEN_H_SET. Thus, the junk sequence "ら見ると" is mistaken as the auxiliary statement phrase connected to the MIZEN form.

The method to remove these junk sequences from SET-C is very simple. If α and "C α" are found in SET-C, we remove α from SET-C. Note that this operation is done increasing order of the length. As a result, we acquire simple auxiliary phrases required in this paper.

3 Experiment

We experimented with the proposed method. As a corpus, we made use of ASAHI-SHINBUN articles for one month. The total amount of the text is about 9 Mbtye.

First, we produced SET-A with this corpus by the operation mentioned in section 2.2. Table.1 shows the number of different sequences and total number of sequences for each length $N \geq 4$. We ignored sequences whose length is greater than 14 because even number of occurrence of this most frequent sequence was too small.

Next, from sequences which appear more than 10 times in SET-A, we removed junk sequences by the operations which were explained in section 2.3 , 2.4 and 2.5. Table.2 shows the number of sequences in the sets (SET-B, SET-C) and the number lastly acquired phrases

As a result, 296 phrases were acquired. The following shows the detail. In the Appendix-1, some of the phrases acquired by this experiment are shown.

length	freq. > 10	SET-B	SET-C	acquired phrases
4	7930	2174	92	83
5	4835	1624	78	63
6	2300	881	83	60
7	881	350	56	36
8	326	159	38	33
9	110	51	15	10
10	48	23	10	4
11	21	13	7	4
12	7	5	3	3
13	2	0	0	0
total	16460	5280	382	296

Table.2 Removing junk sequences

```
+-- Relational Phrase                115
+-- Auxiliary Statement Phrase        87
|
+----- Wrong                         94
```

The causes of wrong extractions are listed below.

(C1) The sequence was an actual word. For example, "とりあえず", "いわゆる" , "おおむね"...etc. (49)

(C2) As a true simple auxiliary phrase had not been extracted, its partial phrase was left in SET-C. (14)
In the next section, we discuss this cause in detail.

(C3) The sequence contained a demonstrative pronoun (in Japanese, demonstrative pronouns consist of HIRAGANA characters). For example, "これに対して", "これまで通り", "それなのに" ... etc. (12)

(C4) It was hard to regard the phrase as one word. For example, "は何なのか", "よう強く" ... etc. (10)

(C5) The sequence contained a suffix of nouns (in Japanese, suffies generally consist of HIRAGANA characters). For example, "さんによると", "たちにとって"...etc. (3)

4 Discussion

The errors caused by (C2) are serious because (C2) implies not only that wrong phrases are extracted but also that true simple auxiliary phrases are not extracted. Hereafter we will talk about the error of this type in detail.

The main cause is that we set too strong constraint for the connection between characters in the simple auxiliary phrase. In the experiment, a sequence γ

$$\gamma = a_1 a_2 \cdots a_n$$
(where a_i is a character)

cannot be regarded as a simple auxiliary phrase if (1) the number of occurrence of γ' exceeds 1/10 of that of γ, or (2) the number of occurrence of γ'' exceeds 1/10 of that of γ.

We show two remarkable cases. The first case is the *interchangeablitly* of Japanese words. For example, the phrase "てください" is a simple auxiliary phrase. But it is also appears in the form of "でください" when some kinds of verbs directly precede this phrase. This means "て" and "で" can be used in the almost same context. The second case is the possibility of insertion of one-character-word "も". This word, different from other words, can be inserted even in a simple auxiliary phrase. For example,

にかかわらず ==> にもかかわらず

To refine the proposed method, the following ideas would be helpful.

(A1) Use of a dictionary

Most of the wrong extractions can be prevented by using a dictionary, especially for the case of (C1). By analyzing the extracted phrases with a dictionary, wrong phrases caused by (c3),(C4) and (C5) can be removed.

(A2) Use of another heuristics

Adding the following heuristics would be useful.

1. The first word in a relational phrase is a postpositional particle.
2. The last word in the auxiliary statement phrase is in SHUUSHI form, or the word is the ending postpositional particle (*"SHUU-JOSHI"*).

(A3) Refinement of the rules

By relaxing the too string restriction of the occurrence, failures caused by (C2) would be reduced.

The proposed method can be used not only for full automatic simple auxiliary phrase extraction but also for computer-aided dictionary compilation.

If we admit a consecutive KANJI characters in a σ-sequence, it is possible to get greater variety of simple auxiliary phrases.

As the proposed method is very simple, it can be easily applied a large scale experiment. In addition, this method can be performed even without a dictionary, a grammar, and syntactic parser, which are very expensive resources. In real, our experiment was done by AWK and SORT which are UNIX tools.

We believe that this method can be used to extract frequent used expressions from a corpus. To build a practical natural language processing system(e.g., machine translation system), it is necessary to gather frequent used expressions[5][6]. This method is very helpful to gather these expressions.

5 Conclusion

A new method was proposed to extract simple auxiliary phrases automatically from a Japanese corpus. The point of this method is to utilize the three heuristics (H1–H3) concerning auxiliary phrases. These heuristics are mainly based on the character type in Japanese. Thus, the proposed method can be implemented without a dictionary, a grammar and a syntactic parser. In addition, the heuristic (H1) is derived from the characteristics of Japanese sentence. By using (H1), this method is efficient. Through the experiment, it was shown that the proposed method is useful to pick up simple auxiliary phrases from a corpus. We provided some ideas(e.g., use of a dictionary) to make the method more efficient.

In the future, we would like to apply the proposed method to extract complex auxiliary phrases and the frequent used expressions.

Acknowledgments

We wish to thank N.Takahashi, who is a member of ETL Natural Language Section, for his helpful comments on this paper.

References

[1] Shudo,K. et al.: "On the Idiomatic Expressions in Japanese Language", (in Japanese) *IPSJ-SIG-L*, 66-1, pp.1-7, (1988).

[2] Matsumoto,Y. : "Approaches to Robust Natural Language Processing" (in Japanese), *IPSJ*, Vol.33, No.7 pp.757-767,(1992).

[3] Inoue,N. : "Automatic noun classification by using Japanese-English word pairs", *Proc. Annual Meeting of the Association for Computational Linguistics*, pp.201-208 (1991).

[4] Nagao,M. and Mori,S. : "A method of n-gram statistics for large text data of Japanese, and the automatic extraction of words and phrases" (in Japanese), *IPSJ-SIG-NL*, 96-1, pp.1-8, (1993).

[5] Katoh,N. and Aizawa,T. : "Extraction and machine translation of sentences with fixed patterns for AP wire service news stories" (in Japanese), *IPSJ-SIG-NL*, 93-1, pp.7-14, (1993).

[6] Furuse,O. and Iida,H. : "Translation by Cooperation Between Transfer and Analysis" (in Japanese), *IPSJ-SIG-NL*, 87-4, pp.27-34, (1992).

Appendix-1

ていかなければならない，ないことを明らかにした，ことが明らかになった，
ていることがわかった，なければならなかった，ことができなかった，
ことを明らかにした，ていかねばならない，ではないでしょうか，なければなりません，
があるとしている，ことになりそうだ，ことになるだろう，ことは許されない，
たのをきっかけに，ていることに対し，ているのだろうか，ということだろう，
というものだった，との考えを示した，とは思わなかった，とみて調べている，
わけにはいかない，ことが望ましい，ことにしている，ことにしました，
ことのないよう，ことはなかった，こともあり得る，ているのに対し，てしまいました，
としか思えない，としては初めて，と見られている，ものとみられる，
をはじめとする，を示すとともに，かもしれない，からといって，ことでしょう，
ことなどから，ことについて，ことになれば，ことになろう，ことによって，
ことに対して，ことを踏まえ，ていくだろう，てもらいたい，になりやすい，
のままにして，は考えにくい，ばかりだった，ませんでした，を抑えるため，
を浴びている，からすれば，からだろう，からみると，が続く中で，ことだろう，
ことながら，ことに加え，だけでなく，ていきたい，ている中で，ておきたい，
にかんがみ，にすぎない，ものとして，をめぐって，とともに，ない限り，
における，について，につれて，によって，によると，によれば，に関して，
に際して，に次いで，に対して，に値する，に伴って，に面した，の中から，
を通して，を通じて　…… etc

Fig.5 A part of simple auxiliary phrases