Extraction and Verification of KO-OU Expressions from Large Corpora

Atsuko kida[†] **, Eiko Yamamoto**[‡] **, Kyoko Kanzaki**[‡] **, and Hitoshi Isahara**[‡]

[†] The Institute of Behavioral Sciences
2-9 Honmura-cho, Ichigaya, Shinjuku-ku,
Tokyo, 162-0845, Japan
akida@ibs.or.jp

[‡] Communications Research Laboratory
3-5 Hikari-dai, Seika-cho, Souraku-gun,
Kyoto, 619-0289, Japan
{eiko,kanzaki,isahara}@crl.go.jp

Abstract

In the Japanese language, as a predicate is placed at the end of a sentence, the content of a sentence cannot be inferred until reaching the end. However, when the content is complicated and the sentence is long, people want to know at an earlier stage in the sentence whether the content is negative, affirmative, or interrogative. In Japanese, the grammatical form called the KO-OU relation exists. The KO-OU relation is a kind of concord. If a KO element appears, then an OU element appears in the latter part of a sentence. It is being pointed out that the KO-OU relation gives advance notice to the element that appears in the latter part of a sentence. In this paper, we present the method of extracting automatically the KO-OU expression data from large-scale electronic corpus and verify the usefulness of the KO-OU expression data.

1 Introduction

The Japanese language has a grammatical form called the KO-OU relation. The KO-OU relation is a kind of concord, also referring to a sort of bound relation that a KO element appearing in a sentence is followed by an OU element in the latter part of the same sentence. On the contrary, the cooccurrence relation refers to two words appearing in the same sentence.

Because Japanese predicates are usually located at the end of sentences, the contents of Japanese sentences cannot be decided until reaching the end. Furthermore, in Japanese, it is hard to comprehend the meaning of the sentence without reading through the entire sentence. The KO-OU relation is the grammatical form which can be helpful for understanding the sentence meaning at the early stage. While in archaic Japanese, KAKARI-MUSUBI, which had morphemic KO-OU relation between KAKARI-JOSI[1] and the conjugation at the end of a sentence, had been used. KAKARI-MUSUBI gave advance notice to the elements that would appear toward the end of a sentence due to KAKARI-JOSI. Today, KAKARI-MUSUBI has dropped out of use. However, the KO-OU relation such as "*sika-nai* (only)" or "*kessite-nai* (never)" is present. In this research, we have attempted to collect such elements to extract KO-OU expression data. In this paper, the main points of argument are as follows:
(1) Method of extracting automatically the KO-OU expression data.
(2) What the KO-OU expression data can be used for.

2 The Previous Works and How to Position this Study

(Ohno, 1993) pointed out that there were expressions that try to give advance notice to whether a sentence is affirmative, negative, or interrogative at the early stage of a language expression which continues timewise. It suggested that there were certain adverbs that have replaced KAKARI-JOSI in the archaic Japanese words.

(Masuoka, 1991) described the KO-OU relation of sentence elements. According to Masuoka, some sentences have the KO-OU expressions as shown in Table 1.

However, this has the following weaknesses. The KO and OU elements in a KO-OU relation are placed together in the same category, and there is

[1] A Japanese particle.

no description as to the OU element. Furthermore, only a limited number of elements are listed. And the objectivity of the KO and OU elements is not guaranteed.

The KO-OU expression data is useful as basic data to dissolve ambiguity in parsing and to decide on the modification relation. However, first of all, it is necessary for the data to have a certain length for being useful basic data. Secondly, it also needs to be objective. Therefore, we have attempted to extract KO-OU relations automatically from large-scale corpus.

Table 1 Masuoka's KO-OU expression data

KO element	OU element
Nee, oi	*te-kudasai, naa*
tabun, doumo	*daro-u, rasii, you-da*
kessite, kanarazu-si-mo	*nai*

3 Assumed Usage of KO-OU Expression Data

3.1 To Dissolve Ambiguity

The KO-OU expression data is useful for dissolving ambiguity of parsing. Furthermore, it is useful for deciding the modification relation (Figure 1).

3.2 Gradual Understanding

Using the KO-OU expression data will enable the reader to guess the end expression midway through a sentence. This is because as the KO elements appear it is possible to predict the appearance of the OU elements (Figure 2). It can be used as a basic data for understanding sentences. In addition, this technology can be used to guess the point in the minutes of a meeting at which the speakers change.

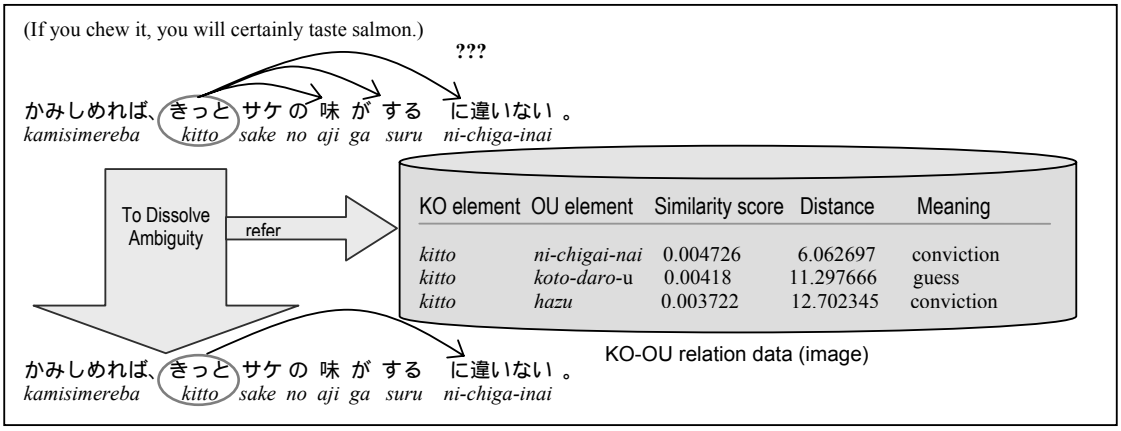

Figure 1 To Dissolve Ambiguity

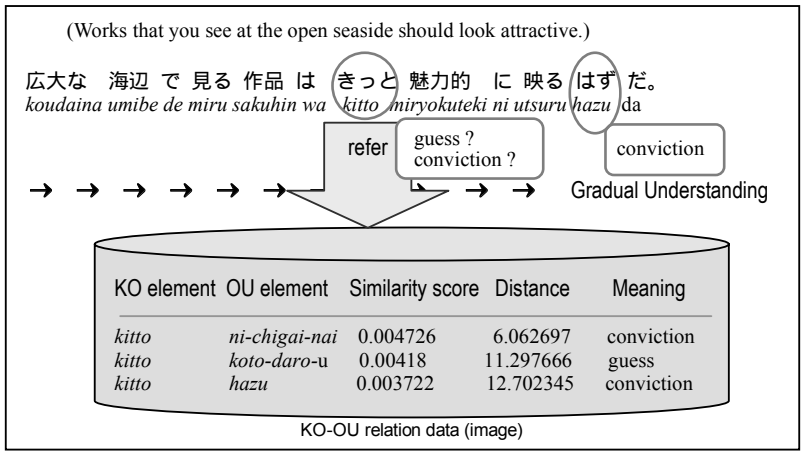

Figure 2 Gradual Understanding

4 Extraction of KO-OU Expression Data

4.1 Method

(Yamamoto and Umemura, 2002) considered the estimation of the one-to-many relation between entities in corpus. They carried out experiments on extracting one-to-many relation of phenomena from corpus using complementary similarity measure (CSM) which can cope very well with inclusion relation of appearance patterns. The KO-OU relation in this research can be regarded as a type of one-to-many relation.

4.2 Data Used

In this paper, we dealt with what is called FUKU-JOSI[2], KAKARI-JOSI, and some adverbs shown below. We proceeded on the assumption that these are the KO elements in the KO-OU relation. For our research, we used newspaper articles from the Mainichi Shimbun, Nihon Keizai Shimbun, and Yomiuri Shimbun issued between 1991 and 2000.

[Target words]
koso, *sika*, *sae*, *ha*, *mo*, *bakari*, *nomi*, *sura*, *nara*, *kurai*, *dake*, *nannte*, *kessite*, *osoraku*, *tabun*, *zehi*, *marude*, *mosi*, *kitto*

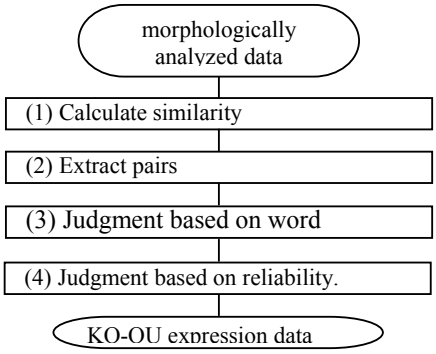

Figure 3 Process flow

4.3 Process Flow

Process flow is shown in Figure 3.
(1) We calculated the similarity measure using CSM for newspaper articles data that had been morphologically analyzed with ChaSen[3].
(2) We extracted pairs containing the target words from the results of similarity measure calculation.
(3) Out of the pairs in (2), we extracted words that appeared in the order of KO and OU elements. (We judge the pairs based on this word order.)
(4) We carried out judgment based on reliability.

As a result of this process, we obtained 14 pairs of data which had "*kesshite*" as KO element, 16 which had "*sae*," and 23 which had "*wa*." Data of approximately 20 pairs was obtained per target word.

5 Verification of KO-OU Expression Data

5.1 Necessity to Give Meaning/Information

If the KO-OU expression data is used for gradual understanding of sentences, it was necessary for the data to be given meaning/information. When the KO element appears, it will be possible to sufficiently grasp or guess the contents of a sentence by referring the KO-OU expression data (Figure2). However it is difficult to give meaning/information using the data obtained from the process in Chapter 4 because the data is broken down into each morpheme by the morphological analysis, and each element is too short.

In Japanese sentences, there are many cases in which continuation of a particle and an auxiliary verb builds a predicate. This continuation plays an important role in determining the event of the sentence. Particles and auxiliary verbs are functional words. Therefore, it is not possible to determine the meaning of some of the particles and auxiliary verbs when they appeared independently. Furthermore, there are some cases in which they change their meaning when paired with another word.

Table 2 shows the OU elements obtained pursuant to the procedure in Chapter 4 for KO element "*kitto*". "*Da*" listed in Table 2 has an assertive meaning when used in a sentence like "*kyou wa ame* ***da*** . (It is raining today.)" On the other hand, it has an inferential meaning in the context of "*asu wa hareru* ***daro****-u* . (It should be fine tomorrow.)" In addition, although "*nai*" is a negative auxiliary verb, when it is paired as in "*ka-mo-shire-nai* (may be)" and "*chigai-nai* (must be)," the negative meaning disappears. And the overall pairing stands for guess and conviction.

[2] A Japanese particle.
[3] Morphological Analyzer ChaSen. See http://chasen.aist-nara.ac.jp/.

Table 2 KO-OU expression data

KO element	OU element	KO element	OU element
Kitto	*u* (auxiliary)	*kitto*	*yo* (particle)
kitto	*da* (auxiliary)	*kitto*	*chigai* (noun)
kitto	*to* (particle)	*kitto*	*ka* (particle)
kitto	*omou* (verb)	*kitto*	*Ne* (particle)
kitto	*nai* (auxiliary)	*kitto*	*you* (noun)
kitto	*hazu* (noun)	:	:

5.2 Verification of OU Element Using "*Kitto*"

In this section, we carry out an analytical example using OU element for KO element "*kitto* (certainly)." We can classify the OU elements obtained from the procedure in Chapter 4, as follows:
(a) It can be an OU element by itself,
(b) It can become an OU element when paired with others,
(c) It does not have the possibility of becoming an OU element.

Words of (c) were not found in the OU elements obtained for KO element "*kitto.*" In the following, we will describe the details on (a) and (b).

(a) OU element by itself

Out of the OU elements for KO element "*kitto*" in Table 2, "*hazu*" can be an OU element by itself.

[1] *koudaina umibe de miru sakuhin wa* ***kitto*** *miryokuteki ni utsuru* ***hazu*** *da* .
(Works that you see at the open seaside should look attractive.)

This is the only sentence with an independent OU element for "*kitto*" in the data obtained from the process in Chapter 4. The same can be said of data for KO elements other than "*kitto*." Because of morphological analysis, the row of letters has been shortened. As a result, there are few elements that can be regarded as an OU element by itself. And just looking at this element does not determine the meaning.

(b) OU element when paired with others

When "chigai" is paired with "*ni*" and "*nai*" to make "*ni-chigai-nai* (must be)," it becomes an OU element. Similarly, pairing "*da*" with "*u*" results in an OU element "*daro-u* (perhaps)." "*Da*" is the original form of "*daro*" and becomes "*daro-u*" when paired with "*u*."

[2] ***kitto*** *kintyou suru* ***daro-u*** .
(It is certain that one will be nervous.)
[3] *kamisimereba ,* ***kitto*** *sake no aji ga suru* ***ni-chiga-inai*** .
(If you chew it, you will certainly taste salmon.)

If we look over the entire pairing shown above, we can give meaning to such guess and conviction.

6 Questions for the Future

As we described in Chapter 5, it is necessary to pair multiple elements before giving meaning/information. We currently persuade the issue of automatic generation of pairing multiple elements. Now, we are carrying out experiments on calculating the similarity measure of pairing of elements. These will give us pairing of automatically generated elements and the similarity measure of the pairings. This should be useful data for resolving ambiguity (Figure 1).

7 Conclusion

This paper presented the process of extracting KO-OU expression data using CSM and the usefulness of the extracted KO-OU expression data. We are planning to report on the findings of experiments on automatic generation of OU elements pairings.

Acknowledgments To compile this paper, we used newspaper articles from The Mainichi Newspapers, The Yomiuri Shimbun, and Nihon Keizai Shimbun.
We would like to sincerely thank Dr. M. Utiyama of the Communications Research Laboratory for allowing us to use a KWIC tool "tea[4]."

References

A.Kida, E.Yamamoto and H.Isahara. 2002. Analysis of expression which projects the following elements beforehand. IPSJ SIG Notes NL-152, pp.137-143.

A.Kida, E.Yamamoto, K.Kanzaki and H.Isahara. 2003. The key on the syntax which brings forth a concord relation. Proceedings of the 9th Annual Meeting of the Association for NLP. pp.23-26.

T.Masuoka. 1991. Grammar of modality. Kurosio-syuppan.

S.Ohno. 1993. Research of a KAKARI-MUSUBI. Iwanami-Shoten.

E.Yamamoto and K.Umemura. 2002. A similarity Measure for Estimation of One-to-Many Relationship in Corpus. Journal of Natural Lamguage Processing. Vol.9 No.2. pp.45-75.

[4] See http://www2.crl.go.jp/jt/a132/members/mutiyama/software.html.