Automated Acquisition of Case Frames with Case Order

Kiyonori Ohtake, Masahiko Nedu, Shigeru Masuyama, and
Department of Knowledge-based Info. Eng., Toyohashi University of Technology
1-1 Hibarigaoka, Tempaku-cho, Toyohashi 441-8580, Japan
{otake,nedu,masuyama}@smlab.tutkie.tut.ac.jp

Kazuhide Yamamoto
ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan
yamamoto@itl.atr.co.jp

Abstract

This paper proposes a case transition network model to provide a framework for representing case order information in addition to a Japanese case frame. The model is regarded as an extension of *bi*-gram model employing a case element as a unit. A preliminary investigation of the model leads us to the conclusions that the transition network has sufficient capacity to acquire case frames with case order.

1 Introduction

Case frames are useful for syntactic and semantic disambiguation. However, case frames which have been collected are insufficient for natural language processing systems because most of the past works collecting case frames were performed manually(e.g. (IPA, 1987)).

We propose a model which considers the order of case element in a simple sentence and a method acquiring verbal case frames statistically. Here, a case element is composed of the semantic features and the case marker, and we call it a simple sentence if a sentence has only one verb and there has been no noun after the position of the verb.

Although a number of studies have been done on automatic acquisition of Japanese verbal case frames(Mine et al., 1997; Oishi and Matsumoto, 1995), there is no consideration on the case orders in the case frames, because changing a case order does not affect the meaning expressed by the case frame.

Although Japanese is considered to be a free word-ordered language, the most general order is as follows: "a time ingredient – a place ingredient – a nominative – a dative – an accusative – a verb(Teramura et al., 1987)", and this fact is confirmed statistically in (Institute, 1964). However, the investigation(Institute, 1964) of word orders ignored difference of verbs, and did not mention that whether patterns of word orders differ depending on a verb or not. If patterns of word orders differ depending on a verb, word order information must be useful on detailing a verbal semantic classification.

The acquisition of case frames with case order has a great significance from the practical point of view. We hope that case frames with case order will be applied to:

- acquisition of information, such as stress, that will contribute to a pragmatic context analysis,
- detailing a verbal semantic classification, and
- generation of sentences with natural case order.

2 Case transition network

We propose the case transition network as a framework for representing case order in addition to a case frame. We consider a method of acquiring case frames statistically from instances on a monolingual corpus.

We expand the *bi*-gram model employing a case element as a unit to reflect the word order precisely. The method proposed in this paper acquires surface verbal case frames by learning from a monolingual corpus on the case transition network.

The case transition network is roughly illustrated in Fig. 1. Outline of the learning on the model is as follows.

1. The model scans a case element from the beginning of a sentence.
2. The model transits a state to another by the appearance of a case element. Then the weight on the arc is calculated.

3. Finally, transition reaches to the terminal state by the appearance of a verb, and learning for a sentence is completed.

A path from the starting state to the terminal state, which consists of arcs having non-zero value, represents a case frame.

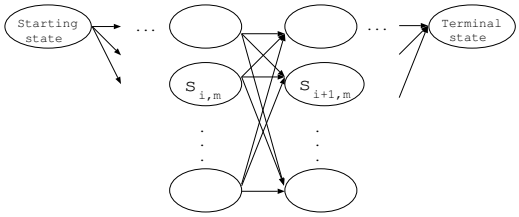

Fig. 1: Overview of the case transition network model

The case transition network model is formally defined as follows. When a verb v, and a number of case elements n are given, the case transition network $N(v,n) = (S, A, w)$ is composed of three components, set S of states, set A of arcs, and weight w of arcs.

State: There are three kinds of states: starting, intermediate, and terminal states. A starting state corresponds to the beginning of a simple sentence, an intermediate state corresponds to a case element, and a terminal state corresponds to a verb. Each case transition network has exactly one starting and one terminal state. Intermediate states are distinguished by the appearance position, and are divided into layers. Each layer consists of all case elements. An intermediate state is represented by $s_{i,m}(1 \leq i \leq n)$, where i denotes the position of a case element in a simple sentence, and m denotes a case element.

Arc: An arc represents the relation between two states. The transition, from $s_{i,k}$ to $s_{i+1,l}$, is represented by arc $(s_{i,k}, s_{i+1,l})$. An arc $(s_{i,k}, s_{i+1,l})$ has a weight $w(s_{i,k}, s_{i+1,l})$.

The weight on arcs: The weight $w(s_{i,k}, s_{i+1,l})$ is given by the conditional probability $p(s_{i+1,l}|s_{i,k})$, where $p(s_{i,k}, s_{i+1,l})$ corresponds to the transitional probability from $s_{i,k}$ to $s_{i+1,l}$ and $p(s_{i,k})$ corresponds to the appearance probability of $s_{i,k}$. Then the $p(s_{i+1,l}|s_{i,k})$ is given by formula (1).

$$w(s_{i,k}, s_{i+1,l}) = p(s_{i+1,l}|s_{i,k}) = \frac{p(s_{i,k}, s_{i+1,l})}{p(s_{i,k})} \quad (1)$$

A sequence of case elements are represented by a path from the starting state to the terminal state. We call it learning to give the weights to the arcs.

The presuming probability in appearance for a sequence of case elements by a case transition network is given by the product of the weights on arcs at the path. For the definition of the case transition network, the network may have a case frame which does not appear in the corpus for learning. We expect that when the network has a case frame which does not appear in the corpus but has a high appearance probability, the case frame will be practically useful.

3 Experiments

To examine the appropriateness of the case transition network, we carried out two experiments. Firstly, to examine how much information of a sequence of case elements is preserved by the case transition network, we executed experiments of comparing the frequency of a sequence of case elements with the presumed probability by the case transition network. This is because, the case transition network based on a bi-gram model, therefore the network does not handle long distance dependencies of cases. Secondly, we investigated whether the case order information by a case transition network is proper or not.

We used articles in The Nihon Keizai Shimbun, a Japanese daily newspaper for business, as a corpus. We obtained 557,048 sentences as a consequence of extracting simple sentences from the corpus.

We must assign semantic features to a noun. We adopted eighteen semantic features mentioned in IPAL verb dictionary(IPA, 1987) which is collected manually by IPA(The Information-technology Promotion Agency, Japan). We constructed a dictionary to assign semantic features to nouns by allocating semantic features to categories of the 'Kadokawa Ruigo Shinjiten' (Kadokawa New Thesaurus)(Oono and Hamanishi, 1981), and we revised a part of the dictionary. We also added some nouns to the dictionary.

We picked up nine postpositional particles: "*ha*, *ga*, *wo*, *ni*, *kara*, *he*, *to*, *yori*, *de*" as case markers. Note that, although '*ha*' is not a case marker, we handle, in this paper, the '*ha*' for

special postpositional particle as a case marker.

The method of extracting a sequence of case order for a verb is given as follows.

1. We obtain simple sentences in advance where a verb is designated from the corpus analyzed by morphological analyzer JUMAN[1].
2. The simple sentence is parsed by KNP[1].
3. We mark the case elements which consist of a noun and a case marker in the parsed sentence. If there is a suffix, the noun and the suffix are put together as a noun.
4. We assign semantic features to each noun in case elements by exact matching. If we can not assign semantic features to a noun by exact matching, we try the longest-first method from behind of the noun. Finally, when we can not assign semantic features to a noun, then we mark the noun as impossible assignment. It is possible that a noun is assigned several semantic features.
5. We output the sequence of case elements.

When a noun is assigned several semantic features in a sequences of case elements, sequences of case elements in compliance with the number of semantic features are produced.

We choose the verbs, "*hiraku*(to open), *hajimeru*(to begin), *matomeru*(to organize), *deru*(to exit), *hairu*(to enter)", which have been ranked higher in the corpus as a subject of our investigation.

3.1 Investigation of probability presumed

We experimented on condition that a verb is '*hiraku*(to open)', and the number of case elements in a simple sentence is equal to three. Fig. 2 demonstrates the relevance between presumed sequences of case elements by the network and frequencies of sequences of case elements in the learning corpus. In Fig. 2, the horizontal axis indicates the frequency of a sequence of case elements, and the vertical axis indicates the probability of appearance for a sequence of case elements presumed by the case transition network.

We also investigated the other verbs. Though the frequency and presumed appearance probability differ, the result shows that the presumed

[1]http://pine.kuee.kyoto-u.ac.jp/nl-resource/

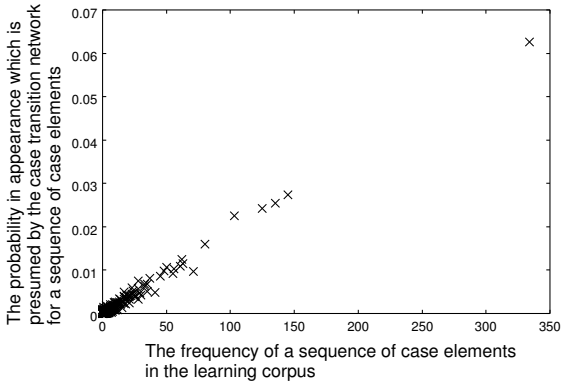

Fig. 2: The relation between presumed sequences of case elements by the network and the sequences in the corpus for the verb, '*hiraku*'

appearance probability becomes higher in proportion to the frequency of sequences by case elements in the corpus in a manner similar to the case of Fig. 2.

3.2 Investigation of case orders

To evaluate how case orders are preserved by the case transition network, we examined sentences having three case elements.

We defined preservation ratio to evaluate it as follows. Firstly, when a number of case elements is three, a sequence α_i of case elements, and sequences obtained by changing case order of α_i, consist of six sequences, then we let $\{\alpha_{i1}, ..., \alpha_{i5}\}$ stand for the sequences except α_i. In addition, we defined a function g:

$$g(\alpha_i) \equiv \begin{cases} 1, & \forall j,\ p(\alpha_i) > p(\alpha_{ij}), \\ & \quad s.t.\ f(\alpha_i) > f(\alpha_{ij}), \\ 0, & \text{otherwise.} \end{cases}$$

Here, $f(\alpha_i)$ denotes the frequency of α_i at the corpus, and $p(\alpha_i)$ denotes the presuming probability by the case transition network of α_i. Let Z be a set of sequences obtained by learning. Then the preservation ratio of case orders is defined by the following formula.

$$\frac{\sum_{\forall \alpha_i \in Z} g(\alpha_i)}{|Z|} \times 100(\%) \qquad (2)$$

The results on investigation of five verbs are shown in Table 1.

4 Discussions

Fig. 2 shows that the case transition network presumes a sequence of case elements in pro-

Table 1: The preservation ratio of case orders

the verb	(A) (%)	(B)
hiraku(to open)	95.0	2143
hajimeru(to begin)	98.7	1304
matomeru(to organize)	97.2	205
deru(to exit)	97.6	336
hairu(to enter)	95.7	419
average	96.8	881.4

(A): the preservation ratio of case order
(B): the number of simple sentences

portion to the frequency of the sequence of case elements in the corpus for learning. We can conclude by Table 1 that the transition network is a model which sufficiently preserves the case order information. These results lead us to the conclusion that the transition network has a sufficient capacity to acquire case frames with case order.

When the number of case elements in a simple sentence is not less than three, the case transition network can cope with the problem of data sparseness by definition. However, the verification of the capacity to cope with the data sparseness is difficult owing to the ambiguity in the dictionary of semantic features.

In this model, we can assign several semantic features to each noun. Consequently, several semantic features on a noun demand the several learning sequences of case elements. Hence the number of learning sequences is very large compared with the number of simple sentences. For example, if a verb is '*hiraku*(to open)', and the number of case elements in a simple sentence is three, then 5,256 learning sequences of case elements are obtained from 2,143 simple sentences. It is easy for us to imagine that the above fact will do harm to the learning process of the case transition network. For the reasons mentioned above, employing a semantic feature dictionary with high quality will be needed to improve the case transition network. How to tackle these problems remains for future work.

Unfortunately, we could not examine the ability of the case transition network for detailing verbal semantic classification. However, the verb '*hiraku*' has two examples "N1 *ga* N2 *de* N3 *wo*" and "N1 *ga* N2 *wo* N3 *de*" as semantically different examples in the IPAL verb dictionary(IPA, 1987). The fact encourages us to believe the possibility of verbal semantic classification by the case transition network.

5 Conclusion

We proposed the case transition network model for acquiring case frames automatically. The case transition network is remarkable as it preserves case order on case frames. The following conclusions were derived from our experimental results and discussions.

1. The proposed case transition network presumes a sequence of case elements in proportion to the frequency of case elements sequence frequency in a learning corpus.
2. The network has a sufficient ability to keep case order.
3. Employing a semantic feature dictionary with high quality will be required for improving the case transition network for practical use.

Acknowledgment

The authors are heartily grateful to Kadokawa Publishing Co. who provided 'Kadokawa Ruigo Shinjiten' (Oono and Hamanishi, 1981) in a machine-readable form.

References

The National Language Research Institute, 1964. *Vocabulary and Chinese Characters in Ninety Magazines of Today (3)–Analysis–*, pages 171–239. SHUEI SHUPPAN, Japan. (in Japanese).

IPA, 1987. *IPA Lexicon of the Japanese Language for computers IPAL (Basic Verbs).* (in Japanese).

T. Mine, M. Higashi, and M.Amamiya. 1997. Case frame acquisition and verb sense disambiguation on a large scale electronic dictionary. In *Proc. of NLPRS '97*, pages 221–226.

A. Oishi and Y. Matsumoto. 1995. Lexical knowledge acquisition for Japanese verbs based on surface case pattern analysis. *Trans. of IPSJ*, 36(11):2597–2610. (in Japanese).

S. Oono and M. Hamanishi. 1981. *Kadokawa Ruigo Shinjiten (Kadokawa New Thesaurus).* Kadokawa Publishing Co. (in Japanese).

H. Teramura, Y. Suzuki, N. Noda, and Makoto Yazawa, editors. 1987. *Case Study Nihon Bunpou(Case Studies of Japanese Grammar).* O-U-FU-U. (in Japanese).