An Associative Concept Dictionary for Verbs and its Application to Elliptical Word Estimation

Takehiro Teraoka[†], Jun Okamoto[‡], Shun Ishizaki[†]

[†]Graduate School of Media and Governance, Keio University
[‡]Keio Research Institute at SFC, Keio University
[†‡]5322, Endo, Fujisawa-shi, Kanagawa, 252-8520 Japan
{teraoka, juno, ishizaki}@sfc.keio.ac.jp

Abstract

Natural language processing technology has developed remarkably, but it is still difficult for computers to understand contextual meanings as humans do. The purpose of our work has been to construct an associative concept dictionary for Japanese verbs and make computers understand contextual meanings with a high degree of accuracy. We constructed an automatic system that can be used to estimate elliptical words. We present the result of comparing words that were estimated both by our proposed system (VNACD) and three baseline systems (VACD, NACD, and CF). We then calculated the mean reciprocal rank (MRR), top N accuracy (top 1, top 5, and top 10), and the mean average precision (MAP). Finally, we showed the effectiveness of our method for which both an associative concept dictionary for verbs (Verb-ACD) and one for nouns (Noun-ACD) were used. From the results, we conclude that both the Verb-ACD and the Noun-ACD play a key role in estimating elliptical words.

1. Introduction

Recently, research on natural language processing has been evolving at a rapid pace, but it is still difficult for computers to understand contextual meanings well. A major reason for this difficulty is that while high accuracy of semantic analysis is required, computers still have only limited information about linguistic concepts such as the parts of speech, grammar, and surface meanings. In contrast, whenever human beings speak or write they use vast amounts of complicated knowledge related to natural language expressions. Thus, to improve their ability to understand contextual meaning, computers need to be systemized with this kind of background knowledge.

One way to improve natural language processing is to put such background knowledge onto a computer operating system. To do this, we constructed an associative concept dictionary for nouns (hereinafter referred to as Noun-ACD) by using the results of large-scale online association experiments (Okamoto and Ishizaki, 2001a; Okamoto and Ishizaki, 2001b). It contains quantitative distances between stimulus words and associated words as well as a hierarchy of word meanings. The stimulus words are basic nouns obtained from Japanese elementary school textbooks. The participants were asked to give associated words from the stimulus words with the following seven semantic relations: 'Hypernym', 'Hyponym', 'Part/material', 'Attribute', 'Synonym', 'Action', and 'Situation'. We applied the Noun-ACD to word sense disambiguation (Okamoto et al., 2008) and extracted sentences that were needed to summarize a text (Okamoto and Ishizaki, 2003).

In our daily conversations and in various verbal contexts, verbs play an extremely important role in enabling us to understand meanings. Therefore, systemizing the background knowledge that lies behind verbs is essential.

We conducted association experiments where the stimulus words were verbs with semantic relations that corresponded to deep cases, not simply surface ones. We tried to systemize the background knowledge related to such words by constructing an associative concept dictionary for verbs. One of the features we incorporated into it was a prototype system that could be used to estimate elliptical words in a sentence. In the Japanese language, more so than in other languages, words that have a meaning that is obvious to both reader and listener are often omitted in speaking and writing. Supplying the missing information is therefore crucial in natural language processing. A particularly notable feature of the Japanese language is that there are an especially large number of elliptical words in texts and conversations. This topic has been researched widely through such devices as a probabilistic model (Seki et al., 2002) and case frames (Kawahara and Kurohashi, 2004). Our method differs from these previous studies in that we use associative information from the ACD. We compared the words estimated with the certainty factor used in this system to those words extracted by baseline systems to show its effectiveness at predicting elliptical words.

2. ACD for Verbs

2.1. Association Experiment

To collect associative information on verbs, we conducted large-scale association experiments on the web where the stimulus words were basic verbs with semantic relations corresponding to deep cases. These verbs were from Japanese elementary school textbooks, and we prioritized 200 of them that were entry words in a dictionary of basic Japanese (Morita, 1989). For association purposes, we used the semantic relations shown in Table 1; 'Agent', 'Object', 'Source', 'Goal', 'Duration', 'Location', 'Tool', 'Aspect', 'Reason', and 'Purpose'.

2.2. Dictionary Construction

By using the linear programming method, we calculated distances between stimulus words and associated ones in the same way as used in the Noun-ACD (Okamoto and Ishizaki, 2001a). From the experiments, we obtained three

Semantic relation	Content
Agent	Subject of an action
Object	Object of an action
Source	Source of an action
Goal	Goal or end of an action
Duration	Time or term of an action
Location	Location or space during an action
Tool	Tool or material of an action
Aspect	Aspect, degree or frequency of an action
Reason	Reason or cause of an action
Purpose	Purpose of an action

Table 1: Semantic relations used in experiments

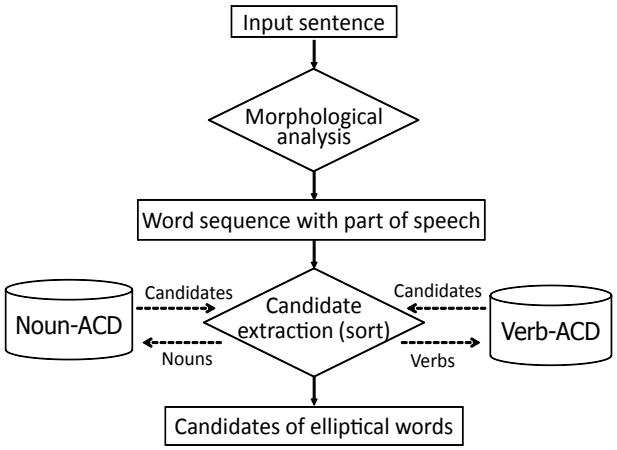

Figure 1: System outline

parameters: the frequency of an associated word F, the average of the associated word order S, and the response time to generate an association T. We assumed a linear equation using these three parameters. Two boundary conditions were provided such that one was for the shortest distance and the other for a comparatively long distance. The optimum coefficients of the linear equation were obtained by using the Simplex Method. These two parameters, F and S, were found to be significant for calculating the distances but the third parameter, T, was not. The distance $D(x, y)$ between the stimulus word x and the associated word y is expressed with the following formulas:

$$D(x,y) = \frac{7}{10}F(x,y) + \frac{1}{3}S(x,y) \quad (1)$$

$$where\, F(x,y) = \frac{N}{n+\delta}, \quad (2)$$

$$\delta = \frac{N}{10} - 1(N \geq 10), \quad (3)$$

$$S(x,y) = \frac{1}{n}\sum_{i=1}^{n} s_i. \quad (4)$$

Let N denote the number of participants in the experiments, and n denote the number of the experiments participants who replied with the associated word, y, to the stimulus word x. Let δ denote a factor introduced to limit the maximum value of F to 10, and let s denote the order each participant gave the associated word. By using three elements, the stimulus verbs, the associated words, and respective distances, we constructed an associative concept dictionary for verbs (hereinafter referred to as Verb-ACD).

In the Verb-ACD, each semantic relation of two words is expressed by each distance where the smaller the distance is, the closer two words are. For example, when a stimulus verb is the Japanese word, *aruku* 'walk' and the semantic relation is 'Source', one of the associated words is *ie* 'home' of which the distance is 1.38. Meanwhile, the distance between 'walk' and *kaisha* 'office' is 9.92. The number of the stimulus verbs in the Verb-ACD is 239 and each stimulus verb has data answered by approximately 40 participants. Currently, the total number of the associated words is about 97,000. In addition, when all of overlapping words are eliminated, the number of the associated words is about 24,000.

3. Application to Elliptical Word Estimation

We constructed a prototype system to estimate elliptical words in a Japanese sentence with both the Verb-ACD and the Noun-ACD (Teraoka et al., 2009). We developed the system and used it to analyze sentences from weblogs. Basically, a notable feature of the Japanese language is that there are many ellipses of the subject and the object in texts and conversations (Nariyama, 2009). Therefore, complementing these words is significant in NLP. This topic related to resolving the zero pronoun has been researched through such devices as a probabilistic model (Seki et al., 2002) and case frames (Kawahara and Kurohashi, 2004).

3.1. System Outline

The outline of the system is shown in Figure 1. First, the system analyzes a sentence morphologically which includes an elliptical word. Associated words from a predicate verb in the sentence are then extracted from the Verb-ACD. If some of these associated words are related to nouns in the sentence or their hypernyms as stimulus words or associated words in the Noun-ACD, these words are selected preferentially as candidates for the elliptical words. Finally, the candidates are sorted by the certainty factors (Section 3.4) and the system outputs words from the sorted list.

3.2. Candidate Extraction

In the Verb-ACD, each entry word (i.e., stimulus verb) is shown as a basic form, so the system first needs to get verbs in an input sentence with a morphological analysis. After conducting a morphological analysis, the system obtains the basic form of a predicate verb and extracts all associated words with a semantic relation specified in the experiments from it. Here we show an example sentence in Japanese:

> *saifu-wo ie-ni wasureta-node tomodachi-kara (gap) karita*
>
> 'I left my wallet at home, so I borrowed (gap) from my friend'.

For the system to ascertain the object of the past-tense verb *karita* 'borrowed', it must extract all associated words from the present-tense verb *kariru* 'borrow' with the semantic relation of the object as the elliptical word candidates.

3.3. Candidate Selection

The system uses information from stimulus nouns and associated words in the Noun-ACD to select any plausible elliptical words from the candidates described in Section 3.2. First, the system examines all the nouns in an input sentence and the hypernyms that are included in the stimulus nouns or associated words in the Noun-ACD. If a noun is found among the stimulus nouns, the system extracts all the associated words that have a semantic relation to the stimulus noun in the Noun-ACD. Among them, the system is used to select elliptical words that overlap the candidates described in Section 3.2 and calculate the certainty factors (Section 3.4). If a noun in the input sentence is found among the associated words in the Noun-ACD, the system extracts all the stimulus words in the Noun-ACD that have a semantic relation with the verb. It then chooses the words among them that are the same as the elliptical candidates.
When there is a proper noun in the sentence, the system searches for stimulus-word hyponyms that include the proper noun. By tracing the links from the stimulus word, it also obtains associative information about the proper noun. In this way, it selects possible elliptical words from the candidates.

3.4. Certainty Factor for Candidates

Finally, the candidates are sorted by the certainty factors and the system outputs these words. The certainty factor of an elliptical word is defined as the product of two distances, L_V and L_N. The former distance L_V is calculated from a predicate verb in the sentence to the elliptical word. The latter distance L_N is basically from a noun in the sentence to the elliptical word. In addition, to prevent from discarding the elliptical candidates which were suitable for final candidates but were not overlapped between the Verb-ACD and the Noun-ACD, the system calculates their certainty factors by regarding the average of each the overlapped word distance as their distances. That is, these distances is L_N in that case. The certainty factor C is expressed as follows:

$$C = \frac{c_j}{\sum_{j=1}^{n} c_j} \quad (5)$$

$$0 < C \leq 1 \quad (6)$$

$$c_j = \frac{1}{L_V L_N} \ (L_V L_N \neq 0) \quad (7)$$

$$0 < c_j \leq 1 \quad (8)$$

The minimum value of these distances is set to 1.0 (Okamoto and Ishizaki, 2001a; Teraoka et al., 2008), that is, the product of these distances L_V L_N must be more than or equal to 1.0. Thus, the range of c_j is shown with Eq. (8) and it divides its by its summation to get the certainty factor C. Therefore, the range of C is also shown with Eq. (6).
Given that the input sentence is the example shown in Section 3.2, the system extracted some candidates for the elliptical words: *kane* 'money' , *te* 'hand', and so on. The certainty factor of 'money' is 0.125, and that of 'hand' is 0.023. It is understood intuitively that someone who does not have a wallet borrows money. On the other hand, the expression 'borrow someone's hand' is a Japanese metaphor which means 'ask for someone's help', but it is not suitable because of 'wallet'. These relations show how the certainty factor represents the strength of estimation.

4. System Evaluation

We carried out some experiments in order to measure the accuracy of the Verb-ACD and our proposed system to estimate elliptical words. To do this, we prepared Japanese test sentences from the web and baseline systems that consisted of the Verb-ACD, the Noun-ACD and the Japanese case frame dictionary. In this section, we describe the details of the experiments and the results.

4.1. Baseline Systems

To investigate the performance of our system and the effectiveness of the Verb-ACD, we used three baseline systems with the following three dictionaries; the Verb-ACD, the Noun-ACD, and the case frames (hereinafter referred as to CF) (Kawahara and Kurohashi, 2005; Kawahara and Kurohashi, 2006). As previously explained, our proposed method needs both the Verb-ACD and the Noun-ACD which were used to extract elliptical word candidates and select output words from those candidates, respectively. We showed the feature of our system by comparing it to two baselines where each ACD were used.
The CF was automatically constructed from about 160,000,000 Japanese web sentences. It contains a great deal of information related to numerous types of verbs and some surface cases related to them (i.e., Japanese case frames; *'ga '*, *'wo '*, *'ni '*, *'de '*, *'kara '*, *'yori '*, *'made '*, etc.) Furthermore, each case frame has many elements (e.g., nouns) that are included in the case and their frequencies that express the number of sentences involving each case and elements. Japanese case frames *'ga '* and *'wo '* respectively correspond to nominative and accusative, so according to circumstances, their elements have semantic relations of the Verb-ACD (e.g., 'Agent' and 'Object'). We therefore regarded the CF as the database that partially contains deep case information of each verb and set the other baseline to it.
The outline of these baseline is as follows;

Baseline 1 (CF). The CF has many verbal types per verb, so the baseline1 system should be decided first. After doing a morphological analysis, this system searches for verbs that have pairs of case frames and nouns in the input sentence. The system then searches the decided verbal types for all elements with case frames that corresponded generally to the following semantic relations shown in Table 2. The system regards all of the obtained elements as elliptical word candidates and outputs them in order of their frequencies.

Baseline 2 (NACD). the baseline2 system can use the Noun-ACD although this system must relay to nouns in the input sentence. After doing the morphological analysis, the system extracts the associated words and the stimulus nouns for all nouns in the input sentence with the semantic relations used in Section 3.1: 'Part/Material' and 'Location' (Okamoto and Ishizaki, 2001a). In the same way as that of the baseline1, the

Case Frame	Semantic Relation
ga	Agent
wo	Object
kara , yori	Source
made , he	Goal
de	Location, Tool

Table 2: Case frames corresponding to semantic relations

system outputs candidates sorted by the distances between the nouns and the candidates.

Baseline 3 (VACD). As our proposed system in Section 3.1, the baseline3 system does the same processing from inputting sentences to extracting elliptical word candidates. After morphological analysis to an input sentence, this system can be used to search for the associated words of the verb with the semantic relation directed by the experimenter and extract all those as elliptical candidates from the Verb-ACD. However, this cannot predict which words would be dropped instead of using only the Verb-ACD. When this system outputs candidate words, they are sorted by the distances between the verbs and the candidates.

4.2. Test Sentences and Correct Answers

From approximately 104,000 Japanese sentences on weblogs, we chose 126 test sentences that were covered by all the systems in our method. This was because the number of entry verbs (i.e., stimulus verbs) in the Verb-ACD was significantly less than that of the CF and the number of sentences which contained stimulus nouns of the Noun-ACD and stimulus verbs of the Verb-ACD was also too. Each test sentence had gaps (i.e., ellipses) which has one of the following semantic relations to the verb: 'Object', 'Source', 'Goal', 'Location', and 'Tool'. In particular, the Japanese language has many ellipses of 'Agent', but we did not use it because *watashi* 'I' is usually omitted and it was not clear enough to identify the omitted agent from only a sentence. Therefore, we used 126 test sentences and made the systems estimate the elliptical words in them.

To set answers for the test sentences, we conducted association experiments where the stimulus was each test sentence with the semantic relation and five participants predicted the elliptical words for each test sentence. Our purpose was to investigate whether our proposed system would estimate words as humans did, so all the words that a majority of the experiment participants (i.e., more than three participants) associated were regarded as correct answers. Then, the 10 test sentences had no correct answer which more than three participants answered, and we set the other 116 test sentences for the evaluation.

4.3. Results

To compare our system to the three baselines, we calculated the mean reciprocal rank (MRR) as shown in Eq. (9). The MRR is frequently used for information retrieval (IR)and the average of the reciprocal ranks r_k. The reciprocal rank is the inverse of the rank of the first correct answer. We calculated three cases, top 1 ($1/r_k = 1.0$), top 5 ($1/r_k \geq 0.2$), and top 10 ($1/r_k \geq 0.1$), and tested a series of changes in the MRR value.

$$MRR = \frac{1}{T}\sum_{k=1}^{T}\frac{1}{r_k} \tag{9}$$

The result of each system is shown in Table 3. The statistical difference was determined by the sign test. Our system (**VNACD**) provided the highest MRR value of all in each case: top 1, top 5, and top 10. The value of the VACD was close to the VNACD. Each value in the three cases shows that these two systems could be used to output a correct answer that is at least within the high ranks. There was a significant difference among them on top 1 from the table ($p<0.05$). In addition, between each case of the VACD and the other baseline systems, there were significant differences ($p<0.01$). On the other hand, MRR values of the NACD and the CF mean that these baseline systems output correct answers in comparatively low ranks. From each relation, it seemed reasonable to suppose that both the VNACD and the VACD systems based on the Verb-ACD were more effective in extracting correct answers within high ranks than the others. In particular, judging from each MRR of the VNACD, it successfully estimated one of the correct answers within the second rank.

As shown in Table 4, we then calculated the top N accuracy (N=1, 5, 10). Each rate means whether the first correct answer was in rank of top 1, top 5, and top 10, respectively. In other words, each is an accuracy of the first correct answer in the top N. As shown on the table, each system has a higher level of accuracy in the top 10 than that within the top 1. The VNACD and the VACD systems had a high possibility of containing at least one correct answer while the other baseline systems did not so. Regarding the relation among the former systems, the VNACD had higher accuracy on top 1 than that of the VACD and there was a significant difference ($p<0.05$). However, on the top 10, the VACD had a higher one than that of the VNACD and there was a significant difference ($p<0.05$). This was because, compared to the VACD, the VNACD had more correct answers in high ranks (i.e., around the top 1) by selecting high certainty factors calculated with the distance information in the Noun-ACD, and fewer in low ranks. This is the adverse effect caused by using the Noun-ACD. In fact, the tendency is shown clearly in top 1 and top 10 among these systems. As previously mentioned, our purpose was to improve the ability of the system to estimate the elliptical words as humans did, so we could ignore the relation of these accuracies in top 10. Hence, our system had a higher possibility of being used to take advantage of the Noun-ACD to select more suitable elliptical words than the baseline system VACD did.

Next, to investigate in more depth whether the VACD could use Noun-ACD in sorting the ellipsis word candidates more effectually than the VACD, we calculated the mean average precision (MAP) in Eq. (10). Let Q denote the number of test sentences and l denote what number test sentence is retrieved. The MAP is used for IE similarly to the MRR

	CF	NACD	VACD	VNACD
MRR (top 1)	0.155	0.086	0.440** ++	0.534** ++ †
MRR (top 5)	0.226	0.157	0.554** ++	0.611** ++
MRR (top 10)	0.241+	0.164	0.570** ++	0.617** ++

Table 3: Results with Mean Reciprocal Rank (MRR). The asterisks, +, and †indicate statistical significance over CF, NACD, and VACD, respectively. (* + †p<0.05, ** ++ ††p<0.01)

	CF	NACD	VACD	VNACD
Accuracy (top 1)	0.155	0.086	0.440** ++	0.534** ++ †
Accuracy (top 5)	0.353	0.284	0.741** ++	0.733** ++
Accuracy (top 10)	0.474+	0.345	0.853** ++ ‡	0.741** ++

Table 4: Top N Accuracy. The asterisks, +, †, and ‡indicate statistical significance over CF, NACD, VACD, and VNACD, respectively. (* + ††‡p<0.05, ** ++ ††‡‡p<0.01)

and the average of the average precisions (AP) in Eq. (11). Let P, m, and R denote the precision of each test sentence, the number of ranks, and the number of the correct answers of each test sentence, respectively.

$$MAP = \frac{1}{Q}\sum_{l=1}^{Q} AP_l \quad (10)$$

$$AP = \frac{l}{R}\sum_{m=1}^{R} P_m \quad (11)$$

As shown in Table 5, our system provided the highest MAP values in all systems. Furthermore, there was a statistical significance over the baseline system VACD (p<0.05). It means that the VNACD had more correct answers within the rank, which was the number of the correct answers in each sentence, than the VACD. Thus, it was clear that the Noun-ACD played a key role in sorting the elliptical word candidates.

In our daily life, we associate only an elliptical word that is the most suitable contextually and understand contextual meanings under ordinary circumstances by using contextual information: nouns, verbs, and so on. Here, each test sentence is almost a simple sentence, so, instead of setting only one answer, we regarded words that the majority of the experiment participants predicted as correct answers. Given the features of the correct answers used in this evaluation, these answers seemed to be based on human associations and part of the background knowledge that humans use whenever they speak or write. Thus, the most effective system VNACD estimated more elliptical words as well as humans do than all baseline did.

Therefore, we concluded that our proposed system VNACD was more effective than all baseline systems and both the Verb-ACD and the Noun-ACD played the key role in estimating elliptical words and sorting words, respectively.

5. Conclusion

We used our Verb-ACD and a previously developed Noun-ACD in an application that estimates elliptical words. We evaluated our system with the MRR (top 1, top 5, and top 10), top N accuracy (top 1, top 5, and top 10), and the MAP. The evaluation was done against baseline systems. Our system was found to have a higher MRR (top 1), accuracy (top 1), and MAP than the values of all baselines. We therefore conclude that the Verb-ACD and the Noun-ACD can be applied to estimate elliptical words and sort these words. Additionally, the Verb-ACD and the Noun-ACD represents the background knowledge of verbs that humans use whenever they speak or write. Our future work is to reduce the number of unsuitable words output and to improve our system.

6. Acknowledgments

This work has been partially supported by the Graduate School Doctorate Student Aid Program 2009, Keio University. The authors gratefully acknowledge the helpful discussions with Dr. Ryuichiro Higashinaka on various points in the paper.

7. References

Daisuke Kawahara and Sadao Kurohashi. 2004. Zero pronoun resolution based on automatically constructed case frames and structural preference of antecedents. In *Proceedings of the 1st International Joint Conference on Natural Language Processing (IJCNLP-04)*, pages 334–341.

Daisuke Kawahara and Sadao Kurohashi. 2005. Gradual fertilization of case frames. *Journal of Natural Language Processing*, 12(2):109–131. (in Japanese).

Daisuke Kawahara and Sadao Kurohashi. 2006. Case frame compilation from the web using high-performance computing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, pages 1344–1347.

Yoshiyuki Morita. 1989. *A Dictionary of Basic Japanese*. Kadokawa Gakugei Shuppan Publishing, Tokyo. (in Japanese).

Shigeko Nariyama. 2009. *How can we know who did what to whom in Japanese? [The Grammar of Omission: Less is More]*. Meijishoin, Tokyo.

Jun Okamoto and Shun Ishizaki. 2001a. Assosiative concept dictionary construction and its comparison with electronic concept dictionaries. In *Proceedings of the*

	CF	NACD	VACD	VNACD
MAP	0.150	0.092	0.401** ++	0.470** ++ †

Table 5: Results with Mean Average Precision (MAP). The asterisks, +, and †indicate statistical significance over CF, NACD, and VACD, respectively. (* + †p<0.05, ** ++ ††p<0.01).

Conference of the Pacific Association for Computational Linguistics(PACLING2001), pages 214–220.

Jun Okamoto and Shun Ishizaki. 2001b. Construction of associative concept dictionary with distance information, and comparison with electronic concept dictionary. *Journal of Natural Language Processing*, 8(4):37–54. (in Japanese).

Jun Okamoto and Shun Ishizaki. 2003. Evaluation of extraction method of important sentence based on associative concept dictionary with distance information between concepts. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics(PACLING2003)*, pages 315–323.

Jun Okamoto, Kiyoko Uchiyama, and Shun Ishizaki. 2008. A contextual dynamic network model for wsd using associative concept dictionary. In *Proceedings of the 6th International Language Resources and Evaluation Conference (LREC2008)*, pages 1595–1599.

Kazuhiro Seki, Atushi Fujii, and Tetsuya Ishikawa. 2002. Japanese zero pronoun resolution using a probabilistic model. *Journal of Natural Language Processing*, 9(3):63–85. (in Japanese).

Takehiro Teraoka, Jun Okamoto, and Shun Ishizaki. 2008. Construction of associative concept dictionary for verbs. In *Proceedings of Forum on Information Technology 2008 (FIT2008)*, volume 2, pages 229–230. (in Japanese).

Takehiro Teraoka, Jun Okamoto, and Shun Ishizaki. 2009. Estimating elliptical words with associative information from verbs. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics(PACLING2009)*, pages 60–65.