Example-based Inference of Unknown Word Category by a Surrounding POS Sequence

Tomotaka Fukuoka
Graduate School of Bionics, Computer and Media Sciences, Tokyo University of Technology
g2110045e1@gss.teu.ac.jp

Shun Hattori
School of Computer Science, Tokyo University of Technology
hattori@cs.teu.ac.jp

Chiaki Kubomura
General Department of Aesthetics, Yamano College of Aesthetics
ckubomura@yamano.ac.jp

Hiroyuki Kameda
School of Computer Science, Tokyo University of Technology
kameda@cs.teu.ac.jp

Abstract

In this paper, we evaluate the accuracy of our example-based method to infer a grammatical category of unknown words in Japanese input sentences. We focus on the similarity of a sequence of POS and other surficial information of words surrounding unknown words in sentences. We change the length of POS sequence around unknown words and the number of similar example sentences to infer word category, to study the effects of these factors on accuracy and computational time. We compare our proposed example-based inference of unknown word category with the other inference methods by rearranging these factors. As a result, we got conclusion that we should use the nearest POS sequence as a query to retrieve example sentences similar to an input sentence and also the most similar one to get more correct accuracy of unknown word POS inference.

Key Words- Part-Of-Speech; POS sequence; Similarity; Category Inference; Unknown Word;

1. Introduction

Recent developments of Information Technology enables us to talk with computers as new partners, who have much amounts of information and knowledge. In chatting, promoting products on the Web, and nursing/caring, there are many cases that human communicate with a computer.

But there are many defects in the communication with computers, such as inferior response and process, comparing the communication between human and a computer with that between human and human. One critical reason is a lack of "smoothness" in dialogue between human and computer. In this paper, we define the smoothness as how speedy communication goes on in accordance with speaker's intention. The lack of smoothness is remarkable when an "unknown word", which a computer doesn't have as its knowledge in a database, appears in an input sentence. For example, asking the meaning of the unknown word or changing a topic of conversation might be expected to solve this problem. But these strategies possibly make lack of smoothness because of interferences by their frequently asking and changing.

Another strategies to solve this problem is to automatically infer information about the unknown word, e.g., its POS, hypernym, and meaning. By this method, any sentence could be manipulated as if with no unknown words.

We adopt a method not to use example sentences with the relevant unknown word contained in a given Japanese input sentence, but to use the example sentences with similar pattern to the input sentence with the unknown word. The similarity between an input sentence and an example sentence is evaluated based on POS sequence, the numbers of words in the input sentence and/or the example sentence, and so on. By using the evaluation, similar example sentences are extracted from a database of our system.

Based on the above approach, this paper proposes a method to infer POS of an unknown word in an input sentence by using a query of some POS sequence surrounding the unknown word in the sentence and some of its similar example sentences. We have already known that a number of example sentences in the database is not important for an accuracy of unknown word POS inference by the latest study [1]. To validate the accuracy of our proposed method, we compare it with the other methods to use a POS sequence by n-gram. We show the dependence of the accuracy and computational time on the length of POS sequence around the unknown word in the sentence and the number of similar example sentences, by comparing our method with the n-gram method of Japanese word category inference.

2. Proposed method

Our method infers the POS of unknown word in an input sentence by using its similar example sentences based on their POS sequence and surficial information set in Japanese. The method would be a part of meaning inference method for unknown words. Our system needs the following circumstance and operation steps, and presupposes that a Japanese morphological analysis tool, MeCab, can correctly detect unknown words in an input sentence.

2.1. Operation Steps

We show the operation steps of our system to infer the POS of an unknown word in an input sentence as follows.

Query Generation and Example Retrieval: generates a POS sequence query from an input sentence with an unknown word, and retrieves example sentences from the example database.

Similarity Calculation and Example Ranking: calculates the degree of similarity between an input sentence and each example sentence based on their surficial information set, and ranks example sentences by the similarity.

POS Inference: infers the POS of an unknown word in an input sentence by using similar example sentences.

2.2. Database

Our system has a word database and an example database. We adopt an IPA's (Information-technology Promotion Agency [2]) dictionary, ipadic 2.7.0 as the word database. The example database has 1279 example sentences that are extracted from a corpus of Japanese dialogues on the Web [3]. These example sentences are made by separating the Japanese dialogues by a point. The kinds of POS are dependent on ipadic 2.7.0. The system fundamentally uses 13 kinds of POS, or 22 kinds of POS by separating Particles into 10 subcategories (shown in Table 1).

Table 1. Kinds of POS in ipadic 2.7.0

noun	verb	auxiliary verb	adjective
adverb	conjunction	interjection	prefix
adnominal	symbol	filler	other
case marker	**adverbial marker**	**final particle**	**dependency marker**
conjunctive particle	**parallel marker**	**adverblize**	**adnominalize**
adverbial makrer/parallel marker/final particle			**special**

2.3. Query generation

To retrieve the example sentences similar to an input sentence with an unknown word from the example database, our system uses a POS sequence query. We preselect a number n which means how many words are included in the POS sequence around the unknown word. The system makes POS sequence as a query by using at most n POS followed by the unknown word or following it. Figure 1 shows an instance. In this case, the system makes a POS sequence query by using 2 forward POS and 2 backward POS. Then the system searches for example sentences similar to the input sentence, which match the query's POS sequence, i.e., the forward POS sequence followed by any 1 POS (corresponding to the unknown word) followed by the backward POS sequence. When a forward POS sequence is shorter than the given number n, the query is shorter than our expected one.

2.4. Similarity calculation

The degree w of similarity between an input sentence and an example sentence retrieved from the example database, is calculated by the Dice coefficient:

$$w = 2 \times \frac{|X \cap Y|}{|X| + |Y|}$$

where X and Y stand for a surficial information set of words in a part of the input sentence and each example sentence corresponding to the query's POS sequence, respectively. And more than 2 words with quite the same surficial information in a sentence are considered as different elements.

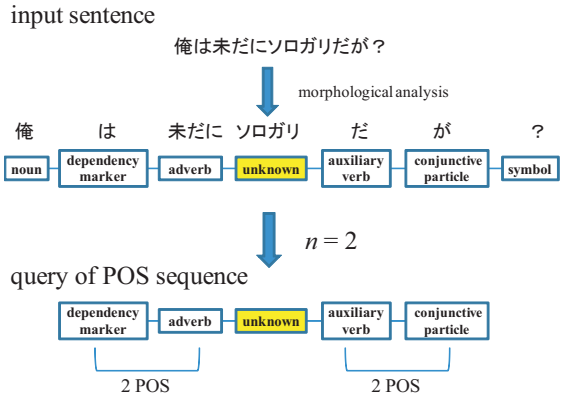

Figure 1. An instance in query generation when n = 2.

2.5. POS Inference

Our system ranks retrieved example sentences in descending order of their similarity (with an input sentence), and then selects some top k examples of the ranking. The number k of similar examples for POS inference is preselected. The system uses the selected similar examples to infer the unknown word's POS. These similar examples have such information as a POS (of the word) at the position of the unknown word in the POS sequence query.

Our system selects a POS at random from among similar example's POS at the position of the unknown word in the POS sequence query, and regards it as the unknown word's POS. If the similarity of some example sentences is quite the same, the system gives more priority to one with higher frequency. In addition, if the frequency as well as the similarity of some examples is quite the same, the system selects one recorded more forward in the example database.

3. POS Inference by n-gram POS sequence

This section explains a POS inference method based on n-gram POS sequence. The method uses either forward or backward POS sequence adjoining an unknown word in an input sentence. The number n of n-gram specifies how many words the POS sequence has.

By using the example database, the n-gram system makes a database of pairs of forward/backward POS sequences and their frequency for each of 22 POS kinds. The number of words in a POS sequence varies dependent on n. If n equals to 1, the system uses just forward or backward POS around the unknown word. If n equals to 10, the system uses a sequence of at most 10 forward or backward POS adjoining the unknown word.

4. Experiment

We evaluate the effects of the length of POS sequence query, n, and the parameter of similarity-based ranking, k, on the accuracy of POS inference by executing our method with several input sentences including an unknown word.

We prepared 135 input sentences that include one unknown word from the Web. The POS kinds of unknown words with the input sentence have 92 nouns, 22 verbs, 10 adjectives, 10 adverbs, and 1 interjection. The correct POS answer of an unknown word in an input sentence is judged by authors.

MeCab's unknown word inference method is judging any unknown word as a noun. Therefore, MeCab's unknown word inference accuracy is 68% in this case.

4.1. Evaluation of proposed method

We evaluate the accuracy and operation time of POS inference affected by both the length of POS sequence query, n, and the parameter of similarity-based ranking, k.

Figure 2 shows the 3D graph that has the length of POS sequence query, n ($1 \leq n \leq 10$), as x-axis, the parameter of similarity-based ranking, k ($1 \leq k \leq 10$), as y-axis, and POS inference accuracy as z-axis. It show that the more the length n of POS sequence query and the parameter k of similarity-based ranking are, the less the accuracy is. These means the best performance is when n = 1 and k = 1.

Figure 3 shows the 3D graph that has the length of POS sequence query, n ($1 \leq n \leq 10$), x-axis, as the parameter of similarity-based ranking, k ($1 \leq k \leq 10$), as y-axis, and the operation time of POS inference as z-axis.

We could know the parameter of similarity-based ranking, k, affects little the average of operation time. And the more the length n of POS sequence query is, the less the average operation time is because the number of example sentences for Step 1 and 2 is reduced by n.

These results mean that to improve the POS inference accuracy and operation time, we had better make the length n and the parameter k smaller because a large number n makes the POS inference accuracy too low and a large number k makes the result unstable by corpora.

4.2. Comparison

We validate our proposed method by comparing

a) our method with k = 1,
b) our method where k is set to the number of all retrieved example sentences (in Step 1),
c) a method by n-gram forward POS sequence,
d) a method by n-gram backward POS sequence.

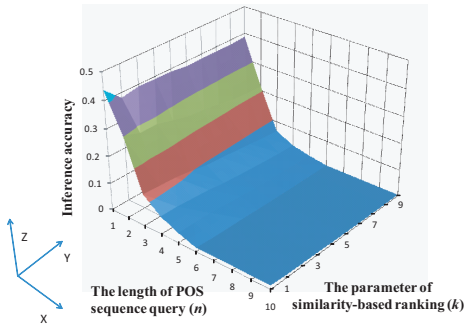

Figure 2. Inference accuracy at n-k.

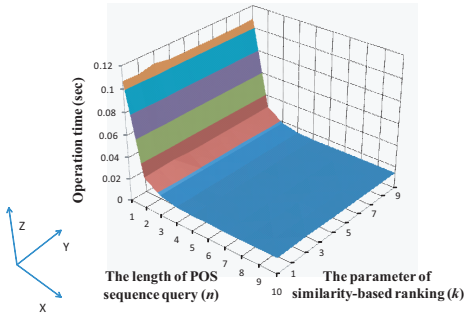

Figure 3. Operation time at n-k.

Figure 4 shows the graph that has the length of POS sequence query, n ($1 \leq n \leq 10$), as x-axis, and the POS inference accuracy by each method as y-axis. And moreover, to compare them in more detail, we gather the results for each POS kind of unknown words as follows. Figures 5, 6, 7, and 8 show the result about the POS inference accuracy for Noun, Verb, Adverb, and Adjective, respectively. Figure 4 shows our proposed method gives the best performance when k = 1. And also it shows the validity of similarity-based example ranking because our method with k = 1 is superior to our method where k is set to the number of all retrieved example sentences. Figure 5 shows the most effective method for Noun inference is not our method but (d) n-gram backward method where n equals to 2 or 3. Figure 6 or 7 shows the most effective method for Verb or Adverb inference is executed by (a) our method with k = 1. Meanwhile, Figure 8 shows that we should not use our method for Adjective inference.

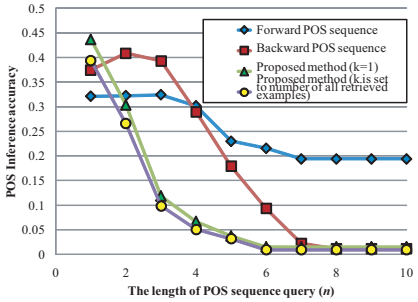

Figure 4. Comparison of accuracy.

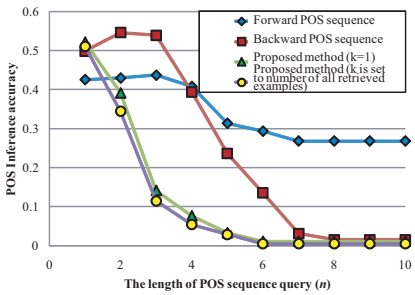

Figure 5. Comparison of accuracy (noun).

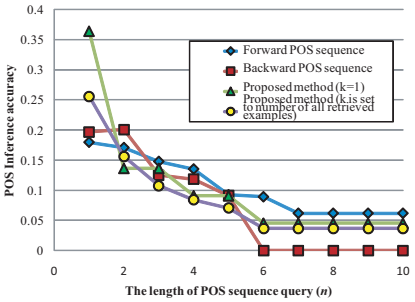

Figure 6. Comparison of accuracy (verb).

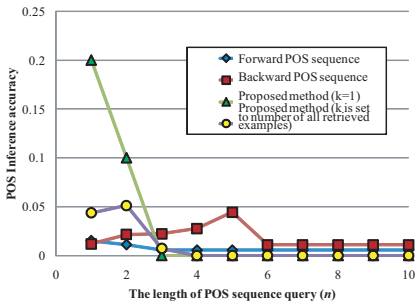

Figure 7. Comparison of accuracy (adverb).

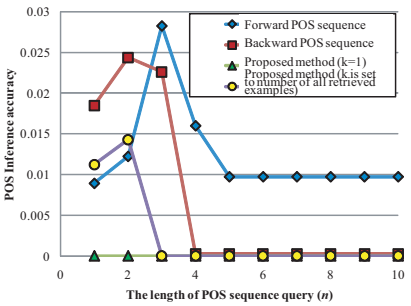

Figure 8. Comparison of accuracy (adjective).

4.3. Discussion

When we set the length of POS sequence, n, so large, it makes the number of retrieved example sentences from the example database too small and they tend to have few POS sequence to give a correct answer. So, we should set n to be the smallest number, i.e., n = 1, even if it causes that they have noisy example sentences. But the length n can be a small number (not just 1) to improve the accuracy of POS inference by the method (c) or (d). It means POS inference methods by using POS sequences get better performance when the length of a POS sequence followed by and/or following an unknown word in a Japanese sentence is set to be small. That is because the best method for POS inference is dependent on POS kinds of unknown words. We should use only the backward POS sequence are Nouns, while we should use not the forward/backward but the surrounding POS sequence when they are Verbs or Adverbs.

To make the parameter of similarity-based ranking, k, large is not much effective for the accuracy of our POS inference method because Figures 2, 4, 5, 6, 7, and 8 show that the more k is, the less the inference accuracy is. We should set k to be 1. Because the dependence of the operation time on k is low and the decline of the inference accuracy on k is gentle, there possibly exists more effective number of k than 1.

5. Conclusion

This paper has proposed a method for POS inference that uses a POS sequence around an unknown word in an input sentence as a query to search an example database and one or more similar example sentences. Our method is executed based on an assumption that words with a same POS have similar POS sequences around them. We have evaluated how long the length of POS sequence queries, n, is effective for POS inference, and compared our method with the other POS inference methods. As a result, we have found the most effective number of POS sequence around unknown words, i.e., n = 1. But also we have a problem, the average accuracy of our POS inference method is not enough high, because our similarity-based filtering of similar example sentences might not work so well. We should invent more effective method with more effective similarity-based ranking to filter only correct information from retrieved example sentences.

References

[1] Fukuoka, T., Hattori, S., Kubomura, C., and Kameda, H.: Studies of Example-based Inference of Unknown Word Category by Query Relaxation and Reinforcement of Part-of-Speech Sequence, HAI'2010, 3D-2 (2010).

[2] Information-Technology Promotion Agency. http://www.ipa.go.jp/.

[3] Research on Understanding and Generating Dialogue by Integrated Processing of Speech, Language and Concept. Spoken dialogue corpus. http://winnie.kuis.kyoto-u.ac.jp/taiwa-corpus/.