ATR-SLT System for SENSEVAL-2 Japanese Translation Task

Tadashi Kumano, Hideki Kashioka and **Hideki Tanaka**
ATR Spoken Language Translation Research Laboratories
2-2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619–0288 JAPAN
{tadashi.kumano, hideki.kashioka, hideki.tanaka}@atr.co.jp

Abstract

We propose a translation selection system based on the vector space model.

When each translation candidate of a word is given as a pair of expressions containing the word and its translation, selecting the translation of the word can be considered equivalent to selecting the expression having the most similar context among candidate expressions. The proposed method expresses the context information in "context vectors" constructed from content words co-occurring with the target word. Context vectors represent detailed information composed of lexical attributes(word forms, semantic codes, etc.) and syntactic relations (syntactic dependency, etc.) of the co-occurring words.

We tested the proposed method with the SENSEVAL-2 Japanese translation task. Precision/recall was 45.8% to the gold standard in the experiment with the evaluation set.

1 Introduction

The SENSEVAL-2 Japanese translation task defines a sense of a Japanese word as an English translation. The same Japanese word in different contexts may have different English translations; therefore, translation ambiguity arises.

Translation Memory (henceforth TM) defining word senses were given to the task participants. Each target word has translation pairs of Japanese and English expressions as word sense candidates[1]. The target word is marked in the Japanese expression, but the corresponding part is unspecified in the English expression. Hence, selecting the most appropriate translation of the target Japanese word in the evaluation expression can be considered to be equivalent to selecting the expression with the most similar context in the TM. This is equivalent to the word sense disambiguation problem in a single language.

[1]Each target word has 21.6 pairs on average.

Generally, word sense disambiguation uses context information, such as the frequency of words that co-occur with the target word. The context information is learned from the correctly-annotated training corpora. However, no training corpus was given for the task and the given TM had shorter contexts because the TM expressions were rather incomplete. Therefore, instead of learning the co-occurring words with the target word from the training corpora, we extract detailed information from the TM expressions as context information. We utilize the information of co-occurring words with the target word (context words) as shown below.

- lexical attributes (word form, part-of-speech, semantic codes on thesaurus, etc.)
- syntactic relations to the target word (dependency relation, etc.)

We employed the vector space model, which is used for text retrieval (Salton and McGill, 1983) to calculate the similarity between the context word information of evaluation expressions and those of the TM. The detailed context information are expressed as "context vectors." We use cosine values between context vectors as a measure of similarity.

In this paper, we will explain first how to construct "context vectors," and then show the accuracy of the selection experiment to the correct data (gold standard).

2 Translation Selection Using Context Vectors

2.1 Context Vectors

2.1.1 Concept

We will explain how to construct a context vector from an expression e_1 with the target word "間 (*aida*; interval)", as an illustration.

Figure 1 shows the expression, which contains the content words "夫婦 (*fuufu*; married couple)", "子供 (*kodomo*; child)", and "産まれ

Table 1: Context Vectors Construction

	Type of syntactic relationship to the target word											
	modifying target word in case relation:				modified by target word in case relation:				target word	...	following words	all context words
	WO	*NO*	*NI*	...	*WO*	*NO*	*NI*	...				
(e_1) "夫婦 の 間 に 子供 が 産まれる (a baby is born to the couple)"	ϕ	*fuufu*	ϕ	ϕ	ϕ	ϕ	*umareru*	ϕ	*aida*	...	*kodomo* *umareru*	*fuufu* *kodomo* *umareru*
(e_2) "仕事 の 間 を ぬって 見舞い に 行く (to visit in hospital at the interval during one's work)"	ϕ	*shigoto*	ϕ	ϕ	*nutte*	ϕ	ϕ	ϕ	*aida*	...	*nutte* *mimai* *iku*	*shigoto* *nutte* *mimai* *iku*

fuufu-no aida-ni kodomo-ga umareru (for e_1); *shigoto-no aida-wo nutte mimai-ni iku* (for e_2)

$= \mathbf{c}_{e_1}$, $= \mathbf{c}_{e_2}$

$\lambda_{modifying\_TW} : \lambda_{modified\_by\_TW} : \lambda_{target} : \cdots : \lambda_{follow} : \lambda_{all}$

The ratio of vector components for each word attribute

る (*umareru*; be born)", and shows that the phrases containing these content words have some syntactic dependencies.

We then prepare a table that enumerates all possible syntactic relations between target word and context words, as in Table 1. For each expression, we then insert corresponding words to the column for each syntactic relation. For example, the row for e_1 of Table 1 can be obtained by the enumeration of expression e_1. If a syntactic relation is applicable to several words, such as the relation "following words" in Table 1, all of them are enumerated in the same column. If no content word comes under the syntactic relation, it is assigned empty (ϕ).

Each row of the table is designated a "context vector" $\mathbf{c}_e$ of a corresponding expression e.

2.1.2 Calculation of Context Vectors

In the preceding section, the table was explained as if it had context words in its elements, but "word attribute vectors" of context words are assigned to them practically. Hence, context vectors are the conjunctions of "word attribute vectors." Each word attribute vector $\mathbf{a}_w$ of a word w expresses lexical attributes of w, such as POS or semantic code. Word attribute vectors have a fixed dimension number, and each element has a non-negative value. The procedure for constructing word attribute vectors will be described below in Section 2.1.3.

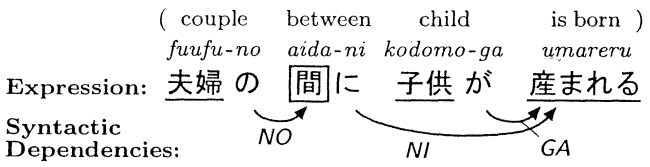

Figure 1: Syntactic Dependencies in Expression e_1

When several context words fall under the same syntactic relation like *kodomo* and *umareru* as we can see in the "following words" relation in Table 1, the word vectors assigned to the relation is calculated by selecting the maximum value for every vector component among values of all words in that relation. The calculation named vecmax is defined as follows:

$$\underset{i=1\ldots m}{\text{vecmax}}\ \mathbf{a}_i = (b_1, b_2, \ldots, b_n),$$

where

$$\begin{cases} \mathbf{a}_i \text{ is a } n\text{-dimensional vector,} \\ a_{ij} \text{ is a } j\text{-th element of vector } \mathbf{a}_i, \text{ and} \\ b_j = \max\limits_{i=1\ldots m} a_{ij}. \end{cases}$$

When joining word attribute vectors into a context vector, each word attribute vector is given a weight in order to get a certain ratio of vector components for each syntactic relation. This is necessary to specify the degree of the contribution to the context vectors according to the type of syntactic relation. For example, assuming that the ratio of the vector components is specified using λ_{syn_rel} (*syn_rel* denotes a specific syntactic relation type) as shown in Table 1, the context vector $\mathbf{c}_{e_1}$ of the expression e_1 will be calculated as follows:

$$\mathbf{c}_{e_1} = \cdots \oplus \lambda_{modifying\_TW} \cdot \frac{\mathbf{a}_{fuufu}}{|\mathbf{a}_{fuufu}|} \oplus \cdots \oplus \lambda_{modified\_by\_TW} \cdot \frac{\mathbf{a}_{umareru}}{|\mathbf{a}_{umareru}|} \oplus \cdots$$

Table 2: Constructing Word Attribute Vectors

	Type of syntactic attribute											
	Emergent Form				Pronunciation				POS			
	夫婦	子供	産まれる	···	*fu-u-fu*	*ko-do-mo*	*u-ma-re-ru*	···	noun	verb	···	
$\mathbf{a}_{fuufu} = ($	η_{e_form}	0	0	0	η_{e_pron}	0	0	0	η_{pos}	0	0	)
$\mathbf{a}_{kodomo} = ($	0	η_{e_form}	0	0	0	η_{e_pron}	0	0	η_{pos}	0	0	)
$\mathbf{a}_{umareru} = ($	0	0	η_{e_form}	0	0	0	η_{e_pron}	0	0	η_{pos}	0	)

	Type of syntactic attribute													
	Semantic Code													
	N86	N85	N74	N72	N5	N4	N3	N2	N1	P26	P17	P16	P1	···
	0	0	$\frac{\eta_{sem}}{\sqrt{7}}$	$\frac{\eta_{sem}}{\sqrt{7}}$	$\frac{\eta_{sem}}{\sqrt{7}}$	$\frac{\eta_{sem}}{\sqrt{7}}$	$\frac{\eta_{sem}}{\sqrt{7}}$	$\frac{\eta_{sem}}{\sqrt{7}}$	$\frac{\eta_{sem}}{\sqrt{7}}$	0	0	0	0	0)
	$\frac{\eta_{sem}}{\sqrt{9}}$	$\frac{\eta_{sem}}{\sqrt{9}}$	$\frac{\eta_{sem}}{\sqrt{9}}$	$\frac{\eta_{sem}}{\sqrt{9}}$	$\frac{\eta_{sem}}{\sqrt{9}}$	$\frac{\eta_{sem}}{\sqrt{9}}$	$\frac{\eta_{sem}}{\sqrt{9}}$	$\frac{\eta_{sem}}{\sqrt{9}}$	$\frac{\eta_{sem}}{\sqrt{9}}$	0	0	0	0	0)
	0	0	0	0	0	0	0	0	0	$\frac{\eta_{sem}}{\sqrt{4}}$	$\frac{\eta_{sem}}{\sqrt{4}}$	$\frac{\eta_{sem}}{\sqrt{4}}$	$\frac{\eta_{sem}}{\sqrt{4}}$	0)

$$\oplus\ \lambda_{target} \cdot \frac{\mathbf{a}_{aida}}{|\mathbf{a}_{aida}|} \oplus \cdots$$

$$\oplus\ \lambda_{follow} \cdot \frac{\underset{i\in\{kodomo,umareru\}}{\text{vecmax}} \mathbf{a}_i}{\left|\underset{i\in\{kodomo,umareru\}}{\text{vecmax}} \mathbf{a}_i\right|}$$

$$\oplus\ \lambda_{all} \cdot \frac{\underset{i\in\{fuufu,kodomo,umareru\}}{\text{vecmax}} \mathbf{a}_i}{\left|\underset{i\in\{fuufu,kodomo,umareru\}}{\text{vecmax}} \mathbf{a}_i\right|}.$$

2.1.3 Word Attribute Vectors

For lexical attributes, we prepare another table similar to that for context words described in the previous section. Table 2 shows that the table enumerates attributes for all words appearing for each lexical attribute. For each word, values are assigned to the column corresponding to the lexical attribute. The value zero is assigned to the column when the lexical attribute is not applicable to the word. In Table 2, the lexical attributes of each context word in expression e_1 are expressed in each row. The row is called "word lexical attributes" $\mathbf{a}_w$ of the corresponding word w.

We employ the semantic codes of a Japanese thesaurus as the semantic attributes. A semantic code may have superordinates because a thesaurus represents semantic relations on the hierarchical tree structure. For example, the word *fuufu* has semantic codes on seven levels, from "Noun 74" on the leaf node to "Noun 1" on the top, in the thesaurus "Nihongo Goi Taikei (Ikehara et al., 1997)" that we used. We treat all semantic codes as semantic attributes of word attribute vectors, and assign values to the corresponding elements equally.

Each lexical attribute of a word attribute vector should be assigned a value, the ratio of component vectors for each word lexical attribute being the specific value η_{word_attr} ($word\_attr$ denotes a specific word attribute type) in Table 2. Semantic attributes may have multiple components to be assigned values, each component should be normalized by the number of the components (See Table 2).

2.2 Translation Selection

To select an appropriate translation for an evaluation expression containing a target Japanese word, we need to compare the context vector of the evaluation expression with the context vectors of all candidate Japanese expressions in the TM. We then choose the candidate whose cosine value to the context vector of the evaluation expression is the maximum.

Each expression should have a unique context vector in order to compare context vectors. But context words, like target words, have ambiguity, and they have several candidates for semantic codes in the thesaurus. It seems unacceptable that the method requires disambiguation of context words before disambiguation of the target word. Therefore, we decided not to disambiguate context words before constructing the context vector. Instead, we construct "context vector candidates" from all combinations of the context word candidates. All combinations of the context vector candidates are used for calculating similarity, and the combination that has the maximum value is selected as the pair of the evaluation and the TM expressions. We can resolve ambiguity of context words when selecting the translation of the target word.

3 Description of Participating System

3.1 Resources, etc.

Our system used the following resources in addition to the given TM and evaluation set.

Table 3: Employed Parameters

word attribute type	ratio
Emergent Word Form	1
Pronunciation	1
Standard Form	4
(standard) Pronunciation	4
Part-Of-Speech	0
Conjugated Form	1
Semantic Code	12

syntactic relation type	ratio
modifying target word (case relation: specific)	3
(case relation: non-specific)	1
modified by target word (case relation: specific)	3
(case relation: non-specific)	1
target word	2
the phrase containing target word	2
preceding target word	1
following target word	1
all content words	2

Japanese Morphological Analyzer:
JUMAN (Kurohashi and Nagao, 1998)

Japanese Syntactic Analyzer:
KNP (Kurohashi, 1998)

Thesaurus:
Nihongo Goi Taikei (Ikehara et al., 1997)

3.2 Parameters

The following parameters have significant effects on the accuracy of our method.

1. The η_{word_attr} ratio of vector components specified for each word attribute when making word attribute vectors (Section 2.1.3)
2. The λ_{syn_rel} ratio of the vector components specified for each syntactic relation when joining word attribute vectors into context vectors (Section 2.1.2)

However, we did not optimize the parameters in our participating system, because of the task specification that no training corpus was given and the time limitations in the course of system development. Parameters were given manually by considering the parameter functions. All of the lexical and syntactic attributes and parameters that represent the ratio between attributes, which our participating system employed, are shown in Table 3.

4 Evaluation

Our participating system marked both the precision and the recall at 45.8% of the correct data (the gold standard) in the evaluation corpus selection. However, our participating system had some serious bugs in the vector normalization process. After correcting the bugs, we made another selection experiment using the same parameters described in Section 3.2. The accuracy of the corrected system was 49.3% (nouns: 50.0%, predicates: 48.5%).

5 Summary

We proposed a translation selection method for the SENSEVAL-2 Japanese translation task. The proposed method calculates the similarity between an evaluation expression containing the target word and Japanese expressions containing the same word in the TM. For calculating similarity, "context vectors" are constructed. Context vectors represent lexical attributes of context words and syntactic relations between context words and the target word. The system employed the proposed method with an accuracy of 49.3% after bug elimination.

Future plans are as follows.

1. To optimize parameters using the gold standard. We would like to use the optimized parameters to study the relation between context information type and accuracy on translation selection. In addition, we will examine whether employed lexical and syntactic attributes are appropriate for the task.
2. To apply the machine learning method to the task, preparing the training corpora. We will make use of the detailed context information proposed, the lexical and syntactic attributes, at machine learning.

References

S. Ikehara, M. Miyazaki, S. Shirai, A. Yokoo, H. Nakaiwa, K. Ogura, Y. Oyama, and Y. Hayashi, editors. 1997. *Nihongo Goi Taikei*, volume 1–5. Iwanami Shoten. (in Japanese).

S. Kurohashi and M. Nagao, 1998. *Japanese Morphological Analysis System JUMAN version 3.61.* Kyoto University. (in Japanese).

S. Kurohashi, 1998. *Japanese Syntactic Analysis System KNP version 2.0 b6 user's manual.* Kyoto University. (in Japanese).

G. Salton and M. J. McGill. 1983. *Introduction to Modern Information Retrieval.* McGraw-Hill.