Automatic Construction of Japanese WordNet

Hiroyuki Kaji and Mariko Watanabe

Department of Computer Science, Shizuoka University
3-5-1 Johoku, Hamamatsu-shi 432-8011, Japan
kaji@inf.shizuoka.ac.jp

Abstract

Although WordNets have been developed for a number of languages, no attempts to construct a Japanese WordNet have been known to exist. Taking this into account, we launched a project to automatically translate the Princeton WordNet into Japanese by a method of unsupervised word-sense disambiguation using bilingual comparable corpora. The method we propose aligns English word associations with those in Japanese and iteratively calculates a correlation matrix of Japanese translations of an English word versus its associated words. It then determines the Japanese translation for the English word in a *synset* by calculating scores for translation candidates according to the correlation matrix and the associated words appearing in the gloss appended to the *synset*. This method is not robust because a gloss only contains a few associated words. To overcome this difficulty, we extended the method so that it retrieves texts by using the gloss as a query and uses the retrieved texts as well as the gloss to calculate scores for translation candidates. A preliminary experiment using *Wall Street Journal* and *Nihon Keizai Shimbun* corpora demonstrated that the proposed method is promising for constructing a Japanese WordNet.

1 Introduction

The WordNet (Miller, 1990) has become the *de facto* standard lexical database. A family of WordNets has been developed for a number of European languages as well as a few Asian languages and these will be linked to form a multilingual lexical database. However, no attempts to develop a WordNet for the Japanese language have been reported. Although a few lexical databases including the EDR concept dictionary (Yokoi, 1995) are available for Japanese, their architecture is different from that of the WordNet. Taking into account the importance of linking lexical databases across languages, we launched a project to develop a Japanese WordNet.

A number of automatic methods were proposed for the EuroWordNet project (Vossen, 1998) to help to develop WordNets for new languages (Farreres, et al., 1998). However, because of the immature technologies and the limited availability of language resources, WordNets have mainly been developed manually. Current corpus-based natural language processing technologies have made a great deal of progress. Consequently, we took the approach of automatically translating the Princeton WordNet into Japanese. This paper describes the method we propose along with a preliminary experiment, which has proved that our approach is promising for constructing a Japanese WordNet.

2 Approach

A WordNet consists of *synsets*, i.e., synonym sets, which define word senses, and inter-*synset* relations such as hypernyms, hyponyms, and others. Therefore, translating a WordNet means translating *synsets*. Our problem is how to determine appropriate translations for a word depending on the *synsets* to which it belongs. We focused on the gloss appended to each *synset*, which provides clues to determine the translations of the words belonging to the *synset*. We adopted a method of unsupervised word-sense disambiguation using bilingual comparable corpora (Kaji and Morimoto, 2002; 2005) to translate words belonging to a *synset*. This method, in which each sense of a first-language word is defined by a set consisting of second-language translations, can directly be applied to translation tasks. We assumed that translation candidates for a word would represent different senses.

Our method calculates the correlations between Japanese translations of an English word and its associated words by aligning English word associations with those in Japanese, where English and Japanese word associations are extracted from the corpora of both languages (See solid arrows in Fig. 1). Given an English word in a *synset*, it calculates the score for each of its Japanese translation candidates according to the gloss appended to the *synset*; the score is defined as the sum of correlations between the translation candidate and the associated words appearing in the gloss (See dotted arrows in Fig. 1).

It should be noted that the proposed method does not require parallel corpora, whose availability is limited. The method can be applied to a pair of weakly comparable corpora in English and Japanese, e.g., the *Wall Street Journal* and the *Nihon Keizai Shimbun*. In addition, it is fully unsupervised; it does not require any manual input such as sense-tagging on the training corpora. It thus has the potential for allowing a large number of English *synsets* to be automatically translated into Japanese.

However, the method has inherent weaknesses in that it lacks robustness because a gloss only contains a few associated words. To overcome this difficulty, we extended it so that it retrieves texts by using a gloss as a query and uses associated words contained in the retrieved texts to determine the translations for the words in the *synset* to which the gloss is appended (See dashed arrows in Fig. 1).

3 Correlation Matrix of Translations versus Associated Words

3.1 Outline

The proposed method is based on the assumption that the translations of associated words are also associated (Rapp, 1995). The alignment of word associations across languages can reveal which associated word of a target word suggests which of its translations. For example, the alignment of an English word association (*tank*, *soldier*) with its Japanese counterpart (戦車<*SENSHA*>, 兵士<*HEISHI*>) reveals that, for the target word "*tank*," its associated word "*soldier*" suggests its translation "戦車<*SENSHA*>." Naive word-association alignment methods, however, are not effective when using non-parallel bilingual corpora. They suffer

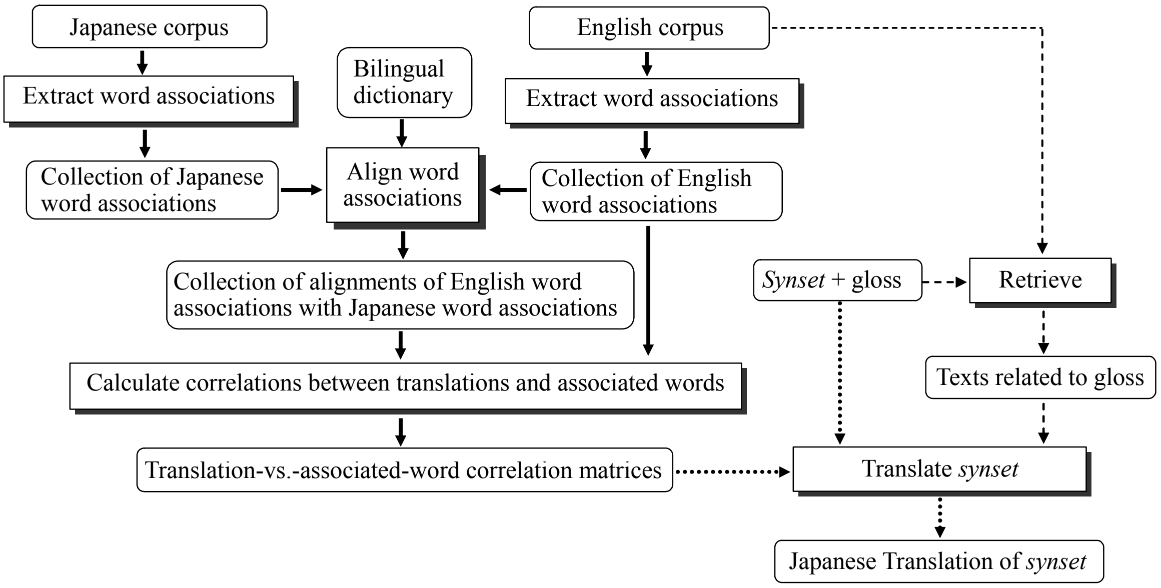
Japanese corpus
English corpus
Bilingual dictionary
Extract word associations
Extract word associations
Collection of Japanese word associations
Align word associations
Collection of English word associations
Collection of alignments of English word associations with Japanese word associations
Synset + gloss
Retrieve
Calculate correlations between translations and associated words
Texts related to gloss
Translation-vs.-associated-word correlation matrices
Translate *synset*
Japanese Translation of *synset*

Fig. 1: Overview of proposed method for *synset* translation

from failed alignments due to topical-coverage disparity between the corpora of two languages as well as ambiguity in alignment. To overcome these difficulties, our method defines a correlation between a translation and an associated word by using correlations between the translation and other associated words.

The method consists of the following steps (See Fig. 1). First, word associations are extracted from the corpus of both languages by setting a threshold for mutual information between words. Second, word associations are aligned across languages with the assistance of a bilingual dictionary. Third, pairwise correlation between the Japanese translations of an English word and its associated words is calculated iteratively.

3.2 Extraction of word associations

The mutual information, $MI(x, x')$, of a pair of words x and x' is defined by

$$MI(x,x') = log\frac{Pr(x,x')}{Pr(x)\cdot Pr(x')},$$

where $Pr(x)$ is the occurrence probability for x, and $Pr(x, x')$ is the co-occurrence probability for x and x' (Church and Hanks, 1990). The occurrence and co-occurrence probabilities are estimated by counting the occurrence and co-occurrence frequencies in a corpus. A medium-sized window is used to count co-occurrence frequencies; in the experiments described in the following sections, the window covered 12 content words on each side of the target word.

A word association is a pair of words with mutual information larger than a threshold, θ. Every pair of words, (x, x'), such that $MI(x, x')> θ$ is extracted from the corpora of both languages. In the experiment described in the following sections, threshold θ was set at zero.

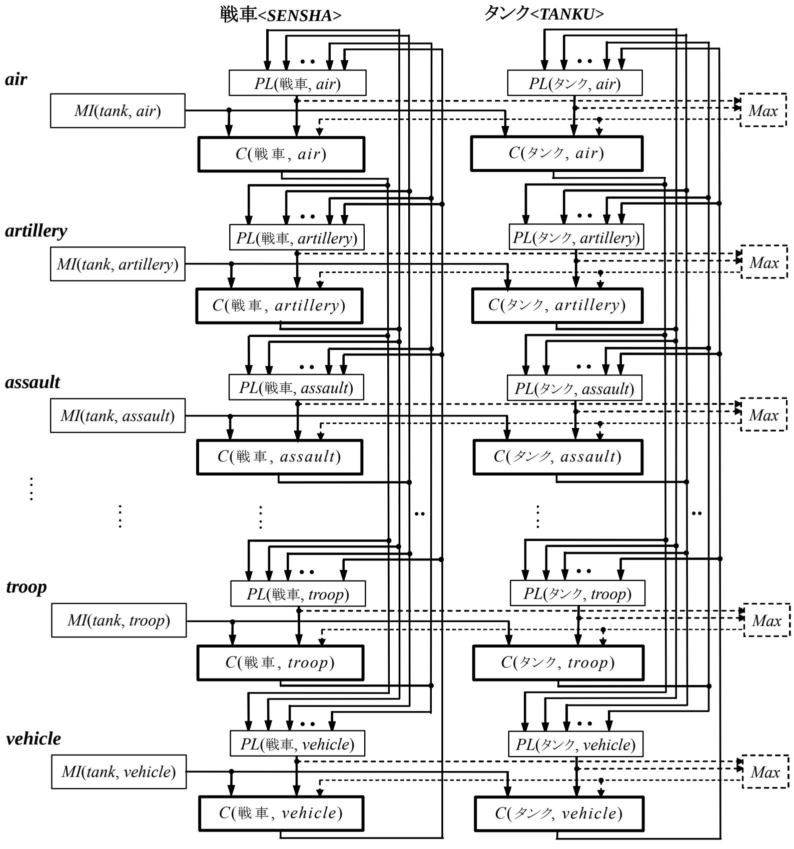

Fig. 2: Recursively defined correlations between translations and associated words

3.3 Calculation of correlations between translations and associated words

The correlations between translations of a target word and its associated words are defined recursively (Kaji, 2005). Figure 2 illustrates an example. For simplicity, it has been assumed that the target word "*tank*" has a set of translations {戦車<*SENSHA*>, タンク<*TANKU*>} and a set of associated words {*air*, *artillery*, *assault*, …, *troop*, *vehicle*}. In the following, focus has been placed on the associated word "*troop*."

The correlation, *C*, between each of the translations for "*tank*" and the associated word "*troop*" is defined as the product of the mutual information between "*tank*" and "*troop*" and a (normalized) plausibility that "troop" suggests the translation. That is,

- $C(\text{戦車} < SENSHA >, troop)$
  $$= MI(tank, troop) \cdot \frac{PL(\text{戦車}, troop)}{\max\limits_{y \in \{\text{戦車}, \text{タンク}\}} PL(y, troop)}$$

- $C(\text{タンク} < TANKU >, troop)$
  $$= MI(tank, troop) \cdot \frac{PL(\text{タンク}, troop)}{\max\limits_{y \in \{\text{戦車}, \text{タンク}\}} PL(y, troop)}$$

The plausibility, *PL*, that the associated word "*troop*" suggests each of the translations for "*tank*" is defined as follows:

- *PL*(戦車<*SENSHA*>, *troop*)
  $$\begin{aligned} &= w(\text{戦車}, tank, troop, air) \cdot C(\text{戦車}, air) \\ &+ w(\text{戦車}, tank, troop, artillery) \cdot C(\text{戦車}, artillery) \\ &+ w(\text{戦車}, tank, troop, assault) \cdot \mathrm{C}(\text{戦車}, assault) \\ &\quad \dots \\ &+ w(\text{戦車}, tank, troop, vehicle) \cdot \mathrm{C}(\text{戦車}, vehicle) \end{aligned}$$
  - *w*(戦車, *tank*, *troop*, *x'*) = 1+α --- *x'* is associated with both "*tank*" and "*troop*," and at least one translation of *x'* is associated with both "戦車" and a translation of "*troop*."
  - *w*(戦車, *tank*, *troop*, *x'*) = 1 --- *x'* is associated with both "*tank*" and "*troop*," but none of the translations of *x'* are associated with either "戦車" or a translation of "*troop*."
  - *w*(戦車, *tank*, *troop*, *x'*) = 0 --- otherwise.

- *PL*(タンク<*TANKU*>, *troop*)
  $$\begin{aligned} &= w(\text{タンク}, tank, troop, air) \cdot C(\text{タンク}, air) \\ &+ w(\text{タンク}, tank, troop, artillery) \cdot C(\text{タンク}, artillery) \\ &+ w(\text{タンク}, tank, troop, assault) \cdot \mathrm{C}(\text{タンク}, assault) \\ &\quad \dots \\ &+ w(\text{タンク}, tank, troop, vehicle) \cdot \mathrm{C}(\text{タンク}, vehicle) \end{aligned}$$
  - *w*(タンク, *tank*, *troop*, *x'*) = 1+α --- *x'* is associated with both "*tank*" and "*troop*," and at least one translation of *x'* is associated with both "タンク" and a translation of "*troop*."
  - *w*(タンク, *tank*, *troop*, *x'*) = 1 --- *x'* is associated with both "*tank*" and "*troop*," but none of the translations of *x'* are associated with either "タンク" or a translation of "*troop*."
  - *w*(タンク, *tank*, *troop*, *x'*) = 0 --- otherwise.

The correlations are calculated iteratively with initial values:

C(戦車<*SENSHA*>, *air*) = *C*(タンク<*TANKU*>, *air*)
= *MI*(*tank*, *air*),

C(戦車, *artillery*) = *C*(タンク, *artillery*)
= *MI*(*tank*, *artillery*),

C(戦車, *assault*) = *C*(タンク, *assault*) = *MI*(*tank*, *assault*),

…,

C(戦車, *troop*) = *C*(タンク, *troop*) = *MI*(*tank*, *troop*), and

C(戦車, *vehicle*) = *C*(タンク, *vehicle*) = *MI*(*tank*, *vehicle*).

Note that *C*(戦車<*SENSHA*>, *troop*) probably becomes larger than *C*(タンク<*TANKU*>, *troop*), because:

(1) *PL*(戦車<*SENSHA*>, *troop*) naturally has a larger number of terms weighted with (1+α) than *PL*(タンク<*TANKU*>, *troop*), and

(2) Most of *C*(戦車<*SENSHA*>, *x'*) weighted with (1+α) or 1 probably become larger than *C*(タンク<*TANKU*>, *x'*).

It has experimentally been proved that the iterative algorithm works stably for a rather wide range of values for parameter α and the correlations converge rapidly (Kaji and Morimoto, 2005). Table 1 shows a correlation matrix of translations versus associated words calculated for the target word "*tank*" by using a *Wall Street Journal* corpus compiled for one and a half years and a *Nihon Keizai Shimbun* corpus compiled for one year. The matrix reveals that associated words such as "*artillery*," "*assault*," and "*battle*" suggest a translation, "戦車<*SENSHA*>," while other associated words such as "*air*," "*car*," and "*explosion*" suggest another translation, "タンク<*TANKU*>."

4 *Synset* Translation Using Glosses

4.1 Method

For a translation candidate of a word in a *synset*, a score is defined as the sum of the correlations between the translation candidate and associated words appearing in the gloss appended to the *synset*. That is,

Table 1: Correlation matrix of translations versus associated words for "*tank*"

	戦車<*SENSHA*>	タンク<*TANKU*>
air	0.530	0.778
artillery	4.044	0.050
assault	1.730	0.047
battle	1.068	0.024
Bosnian	2.169	0.389
car	0.054	0.822
explosion	0.071	1.817
force	1.065	0.029
fuel	0.118	2.954
gallon	0.546	2.006
gas	0.027	1.077
gasoline	0.081	2.192
helicopter	2.044	0.069
leak	1.266	3.564
liquid	0.193	2.253
marine	2.158	2.708
military	1.171	0.035
Pentagon	1.111	0.020
Serb	2.422	0.034
soldier	2.004	0.028
troop	2.387	0.061
vehicle	0.066	1.305

Note: Associated words have only been shown when they are relevant to examples presented in the paper.

$$S(x,i,y(x,j)) = \sum_{x'\in G(x,i)} C(y(x,j),x'),$$

where *S*(*x*, *i*, *y*(*x*, *j*)) denotes the score for the *j*-th Japanese translation candidate *y*(*x*, *j*) of English word *x* in its *i*-th *synse*t, and *G*(*x*, *i*) denotes a set consisting of words appearing in the gloss appended to the *i*-th *synset*. The translation candidate maximizing the score is selected for the word in the *synset*.

Let us look at the following example.

The first *synset* to which "*tank*" belongs is {*tank*, *army tank*, *armored combat vehicle*}, and the gloss appended to it is "*an enclosed armored military vehicle; has a cannon and moves on caterpillar treads*." The underlined words are found in the correlation matrix shown in Table 1. Therefore,

S(*tank*, 1, 戦車<*SENSHA*>)
= *C*(戦車, *military*) + *C*(戦車, *vehicle*)
= 1.171 + 0.066 = 1.237 and
S(*tank*, 1, タンク<*TANKU*>)
= *C*(タンク, *military*) + *C*(タンク, *vehicle*)
= 0.035 + 1.305 = 1.340

As a result, "*tank*" in the *synset* {*tank*, *army tank*, *armored combat vehicle*} is incorrectly translated into "タンク<*TANKU*>."

The second *synset* to which "*tank*" belongs is {*tank*, *storage tank*}, and the gloss appended to it is "*a large (usually metallic) container for holding, gases, or liquids*." Therefore,

S(*tank*, 2, 戦車<*SENSHA*>)
=*C*(戦車, *gas*) + *C*(戦車, *liquid*)
= 0.027 + 0.193 = 0.220 and
S(*tank*, 2, タンク<*TANKU*>)
= *C*(タンク, *gas*) + *C*(タンク, *liquid*)
= 1.077 + 2.253 = 3.330

As a result, "*tank*" in the *synset* {*tank*, *storage tank*} is correctly translated into "タンク<*TANKU*>."

The fifth *synset* to which "*tank*" belongs is {*cooler*, *tank*}, and the gloss appended to it is "*a cell for violent prisoners*." Therefore,

S(*tank*, 5, 戦車) = 0 and *S*(*tank*, 5, タンク) = 0

As a result, "*tank*" in the *synset* {*cooler*, *tank*} cannot be translated. This is reasonable since it seems that "*tank*" never represented this sense in the *Wall Street Journal*.

4.2 Preliminary experiment

A preliminary experiment was done by using a *Wall Street Journal* corpus (July 1994 to December 1995, 189 Mbytes), a *Nihon Keizai Shimbun* corpus (December 1993 to November 1994, 275 Mbytes), and the EDR English-Japanese Dictionary including 633,000 pairs of 269,000 English nouns and 276,000 Japanese nouns. A number of typical polysemous words were translated by using the glosses and the translation-vs.-associated-word correlation matrices. Table 2 lists the results.

The experimental results reveal that the method is not robust; glosses are very short, and a few associated words determine the translation of a word in a *synset*. In addition, it suffers from the following limitations.

First, its resolution is low. For example, it can distinguish between the first sense of "*bill*" (politics) and the second and third senses (economy), but it fails to distinguish between the second and third senses. This is because the method relies on "topically" associated words. This problem may be overcome by combining the translation results for all the words belonging to a *synset*; although "*bill*" in the third *synset* {*bill*, *note*, *government note*, *bank bill*, *banker's bill*, *bank note*, *Federal Reserve note*, *greenback*} is incorrectly translated into "請求書<*SEIKYUUSHO*>," other words belonging to the *synset* are not translated into "請求書<*SEIKYUUSHO*>."

Second, the method often fails to translate a *synset* whose gloss explains the sense mainly using verbs, because word associations are limited to those between nouns. For example, "*trial*" in a *synset* {*test*, *trial*, *run*}, to which the gloss "*the act of testing something*" is appended, cannot be translated. However, it would successfully have been translated into "試験<*SHIKEN*>" if the verb "*test*" appearing in the gloss had been regarded as being the same as the associated noun "*test*."

Third, the method sometimes does erroneous translations due to discrepancies between glosses and general texts. For example, "*plant*" in the second *synset* {*plant*, *flora*, *plant life*} is erroneously translated into "設備<*SETSUBI*>" due to the associated word "power" appearing in the gloss "*a living organism lacking the power of locomotion*." Note that "*plant*" representing the flora sense scarcely co-occurs with "*power*" in general texts. This type of error is serious, although it occurs rather infrequently.

5 Using Texts Retrieved with Glosses

5.1 Method

A score for each translation candidate of a word in a *synset* is defined as a weighted sum of the correlations between the translation candidate and associated words appearing in the texts retrieved by using the gloss appended to the *synset* as a query. That is,

$$S'(x,i,y(x,j)) = \sum_{k} \sum_{x'\in T(x,i,k)} \frac{1}{\sqrt{d(x',x,i,k)}} \cdot C(y(x,j),x'),$$

where *S'*(*x*, *i*, *y*(*x*, *j*)) denotes the score for the *j*-th Japanese translation candidate *y*(*x*, *j*) of English word *x* in its *i*-th *synse*t, *T*(*x*, *i, k*) denotes a set consisting of words appearing in the *k*-th text retrieved with the gloss appended to the *i*-th *synset*, and *d*(*x'*, *x*, *i*, *k*) denotes the distance between *x* and *x'* in the text. Texts are passages consisting of three consecutive sentences, and texts including *x* in the middle sentence are retrieved in descending order of similarity to the gloss.

The weight given to correlations is the same as that in the original word-sense disambiguation method; associated words occurring near the target word are more reliable than those occurring far from it. The range of three sentences is narrower than the range of contexts looked up with the original word-sense disambiguation method, i.e., 60 words before and after the target word. This results from considering the difference in the present task to the ordinary word-sense disambiguation task; many texts can be used to translate one word in a *synset*.

5.2 Example

First, assume that the following two texts have been retrieved with the gloss "*an enclosed armored military vehicle; has a cannon and moves on caterpillar treads*" appended to the first *synset* to which "*tank*" belongs. Note that the underlined words can be found in the correlation matrix in Table 1.

Table 2: Results for *synset* translation using glosses

Test word (Number of *synsets*)	#*	*Synset*	Gloss**	Translation results	Correct/ incorrect
bill (10)	1	*bill, measure*	a statute in draft before it becomes law	議案<*GIAN*>	Correct
	2	*bill, account, invoice*	a statement of money owed for goods or services	請求書 <*SEIKYUUSHO*>	Correct
	3	*bill, note, government note, bank bill, banker's bill, bank note, ...*	a piece of paper money (especially one issued by a central bank)	請求書 <*SEIKYUUSHO*>	Incorrect
	4	*bill*	the entertainment offered at a public presentation	議案<*GIAN*>	Incorrect
	8	*poster, placard, notice, bill, card*	a sign posted in a public place as an advertisement	議案<*GIAN*>	Incorrect
ceiling (4)	1	*ceiling*	the overhead inside lining of a room	天井<*TENJOU*>	Correct
	3	*ceiling, cap*	an upper limit on what is allowed	上限<*JOUGEN*>	Correct
chair (4)	1	*chair*	a seat for one person, with a support for the back	いす<*ISU*>	Correct
	2	*professorship, chair*	the position of professor	講座<*KOUZA*>	Correct
	3	*president, chairman, chairwoman, chair, chairperson*	the officer who presides at the meetings of an organization	会長<*KAICHOU*>	Correct
duty (3)	3	*duty, tariff*	a government tax on imports or exports	関税<*KANZEI*>	Correct
plant (4)	1	*plant, works, industrial plant*	buildings for carrying on industrial labor	設備<*SETSUBI* >	Correct
	2	*plant, flora, plant life*	a living organism lacking the power of locomotion	設備<*SETSUBI* >	Incorrect
rock (5)	1	*rock, stone*	a lump of hard consolidated mineral matter	揺れ<*YURE*>	Incorrect
	2	*rock, stone*	material consisting of the aggregate of minerals like those making up the Earth's crust	揺れ<*YURE*>	Incorrect
	4	*rock 'n' roll, rock and roll, rock, rock music*	a type of dance music originating in the 1950s; a blend of rhythm-and-blues with country-and-western	ロック<*ROKKU*>	Correct
tank (5)	1	*tank, army tank, armored combat vehicle*	an enclosed armored military vehicle; has a cannon and moves on caterpillar treads	タンク<*TANKU*>	Incorrect
	2	*tank, storage tank*	a large (usually metallic) container for holding gases or liquids	タンク<*TANKU*>	Correct
	4	*tank car, tank*	a freight car that transports liquids or gases in bulk	タンク<*TANKU*>	Correct
trial (7)	1	*trial*	(law) legal proceedings consisting of the judicial examination of issues by a competent tribunal	裁判<*SAIBAN*>	Correct
	4	*trial*	(law) the determination of a person's innocence or guilt by due process of law	審理<*SHINRI*>	Correct

* Sense numbers given in WordNet; *synsets* have only been listed when translations were given to the test words in them by the proposed method.

** Underlined words are associated words included in the correlation matrix of translations versus associated words for the test word.

(a) *Turkey, the only country to recognize the self-styled Turkish Republic of Northern Cyprus, has been spending more than $500 million a year in the impoverished north of the island, where gross domestic product was only $706 million last year. Turkey has added 30 new battle* **tanks** *in the past 18 months, giving it a fleet in Cyprus of 265 vehicles. That's 15 more tanks than were in the force with which the Bosnian Serbs nearly conquered Bosnia.*

(b) *It goes out on six-month deployments, trains with militaries from other countries and, if necessary, responds to crises. The MEU's main ground combat element is the Battalion Landing Team,* *which includes Humvees, artillery, amphibious assault vehicles, M1A1* ***tanks****, combat engineers and reconnaissance Marines. The MEU also contains an air combat wing made up of 31 helicopters and six Harriers, and support elements.*

Then,

$$
\begin{aligned}
& S'(tank,1,\text{戦車}<SENSHA>) \\
& = \{C(\text{戦車},battle)/\sqrt{1} + C(\text{戦車},vehicle)/\sqrt{14} + \cdots\} \\
& + \{C(\text{戦車},vehicle)/\sqrt{2} + C(\text{戦車},assault)/\sqrt{3} + \cdots\} \\
& = \{1.068/\sqrt{1} + 0.066/\sqrt{14} + \cdots\} + \{0.066/\sqrt{2} + 1.730/\sqrt{3} + \cdots\} \\
& = 6.891
\end{aligned}
$$

$$
\begin{aligned}
& S'(tank,1,\text{タンク}<TANKU>) \\
& = \{C(\text{タンク},battle)/\sqrt{1} + C(\text{タンク},vehicle)/\sqrt{14} + \cdots\} \\
& + \{C(\text{タンク},vehicle)/\sqrt{2} + C(\text{タンク},assault)/\sqrt{3} + \cdots\} \\
& = \{0.024/\sqrt{1} + 1.305/\sqrt{14} + \cdots\} + \{1.305/\sqrt{2} + 0.047/\sqrt{3} + \cdots\} \\
& = 2.901
\end{aligned}
$$

As a result, "*tank*" in the *synset* {*tank*, *army tank*, *armored combat vehicle*} is correctly translated into "戦車<*SENSHA*>." This demonstrates that the translation accuracy could be improved by using texts retrieved with glosses.

Second, assume that the following text has been retrieved with the gloss "*the act of testing something*" appended to the second *synset* to which "*trial*" belongs.

(c) *Back in 1985, Mr. Millenson, an entrepreneur fresh from the Pollinex turnaround, and his wife, Wendy Strongin, a doctor, came up with a kit for home-testing of HIV. Clinical* ***trials*** *were held. Market research showed millions who otherwise resist testing would use the home kit. Yet the FDA's David Kessler refused to accept an application, instead setting down as policy that his agency would never approve home-testing for HIV.*

Then,

$$
\begin{aligned}
& S'(trial,2,\text{審理}<SHINRI>) \\
& = C(\text{審理},HIV)/\sqrt{2} + C(\text{審理},FDA)/\sqrt{18} \\
& + C(\text{審理},application)/\sqrt{25} + \cdots \\
& = 1.743/\sqrt{2} + 1.296/\sqrt{18} + 0.276/\sqrt{25} + \cdots \\
& = 1.593
\end{aligned}
$$

$$
\begin{aligned}
& S'(trial,2,\text{試験}<SHIKEN>) \\
& = C(\text{試験},HIV)/\sqrt{2} + C(\text{試験},FDA)/\sqrt{18} \\
& + C(\text{試験},application)/\sqrt{25} + \cdots \\
& = 2.163/\sqrt{2} + 1.613/\sqrt{18} + 0.635/\sqrt{25} + \cdots \\
& = 2.037
\end{aligned}
$$

$$
\begin{aligned}
& S'(trial,2,\text{裁判}<SAIBAN>) \\
& = C(\text{裁判},HIV)/\sqrt{2} + C(\text{裁判},FDA)/\sqrt{18} \\
& + C(\text{裁判},application)/\sqrt{25} + \cdots \\
& = 1.641/\sqrt{2} + 1.340/\sqrt{18} + 0.058/\sqrt{25} + \cdots \\
& = 1.488
\end{aligned}
$$

As a result, "*trial*" in the *synset* {*test*, *trial*, *run*} is correctly translated into "試験<*SHIKEN*>." This demonstrates that the applicability could be improved by using texts retrieved with glosses.

We are planning to do a large-scale experiment to confirm the effectiveness of using texts retrieved with glosses. We also need to clarify the relation between the number of retrieved texts and the accuracy of translation, the optimum range for texts to be looked up, and other factors.

5 Conclusion

We proposed automatically translating WordNet *synsets* into Japanese by means of an unsupervised method of word-sense disambiguation using bilingual comparable corpora. The method we proposed calculates the correlation matrix of translations of a word versus its associated words. It then determines a translation for the word in each *synset* according to the associated words appearing in the gloss appended to the *synset*, as well as the texts retrieved by using the gloss as a query. A preliminary experiment proved that the proposed approach is promising for constructing a Japanese WordNet. However, many issues remain to be pursued, which include using hyponyms, hypernyms, and example sentences provided by the Princeton WordNet, and combining the results obtained from corpora of multiple domains.

References

Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16 (1): 22- 29.

Farreres, Xavier, German Rigau, and Horacio Rodriguez. 1998. Using WordNet for building WordNets. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*.

Kaji, Hiroyuki. 2005. Adapting a bilingual dictionary to domains. *IEICE Transactions on Information and Systems*, E88-D (2): 302- 312.

Kaji, Hiroyuki and Yasutsugu Morimoto. 2002. Unsupervised word sense disambiguation using bilingual comparable corpora. In *Proceedings of the 19th International Conference on Computational Linguistics*, pp. 411- 417.

Kaji, Hiroyuki and Yasutsugu Morimoto. 2005. Unsupervised word-sense disambiguation using bilingual comparable corpora. *IEICE Transactions on Information and Systems*, E88-D (2): 289-301.

Miller, George A. 1990. WordNet: An on-line lexical database. *International Journal of Lexicography*, 3 (4): 235- 312.

Rapp, Reinhard. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 320-322.

Vossen, Piek (ed.) 1998. *EuroWordNet: A multilingual database with lexical semantic networks,* Kluwer Academic Publishers, Dordrecht.

Yokoi, Toshio. 1995. The EDR electronic dictionary. *Communications of the ACM*, 38 (11): 42- 44.