The Use of Abstracted Knowledge from an Automatically Sense-Tagged Corpus for Lexical Transfer Ambiguity Resolution

Hui-Feng Li, Namwon Heo, Kyounghi Moon, Jong-Hyeok Lee
Dept. of Computer Science & Engineering
Pohang University of Science & Technology
San 31 Hyojadong, Nam-Gu, Pohang, 790-784, Korea
E-mail: {hflee, nwheo, khmoon, jhlee}@kle.postech.ac.kr

Abstract

This paper proposes a method for lexical transfer ambiguity resolution using corpus and conceptual information. Previous researches have restricted the use of linguistic knowledge to the lexical level. Since the extracted knowledge is stored in words themselves, these methods require a large amount of space with a low recall rate. On the contrary, we resolve word sense ambiguity by using concept co-occurrence information extracted from an automatically sense-tagged corpus. In one experiment, it achieved, on average, a precision of 82.4% for nominal words, and 83% for verbal words. Considering that the test corpus is completely irrelevant to the learning corpus, this is a promising result.

1 Introduction

In Korean-to-Japanese machine translation (MT), which employs a direct MT strategy, a Korean word with multiple senses may be translated into different Japanese equivalents depending on which sense is used in a given context. In general, it is much more difficult for Korean-to-Japanese MT to resolve such word sense ambiguities than its opposite, Japanese-to-Korean MT. This is because unlike Japanese words whose stem is usually written in Kanji (the Chinese ideographic script), Korean words may appear written in Hangul (the Korean phonetic alphabet) only. Thus, word sense disambiguation (WSD) is essential to lexical transfer in an Korean-to-Japanese MT system.

There is a large body of previous work on word sense disambiguation (Kelly 1975; Church 1990; McRoy 1992; Liu 1993; Tanaka 1994; Yarowsky 1995; Hwee 1996). Early work made use of manually coded knowledge, but it required time-consuming and laborious work for knowledge acquisition (Kelly and Stone 1975). The recent emphases on corpus-based WSD usually fall into two categories: supervised and unsupervised. Yarowsky (1995) used an unsupervised learning procedure. However, his experiment was only restricted to "binary" WSD, a kind of coarse sense distinction (Yarowsky 1995; Hwee and Lee 1996). LEXAS (Hwee and Lee 1996) adopted a supervised learning method using multiple knowledge sources such as a human sense-tagged corpus, the POS of neighboring words, morphological forms, local collocations, unordered sets of surrounding words, and verb-object syntactic relations. Since the extracted knowledge in these methods is stored in words themselves, they require a large amount of space to store the knowledge, and they have lower applicability.

In this paper, a corpus-based WSD method is proposed. Unlike previous work (Yarowsky 1995; Hwee 1996) that restricted the use of linguistic knowledge to the lexical level, i.e., surface words, we rely on concept co-occurrence information (CCI) extracted from a sense-tagged corpus. The sense-tagged corpus is automatically constructed from a Japanese raw corpus by an existing Japanese-to-Korean MT system called COBALT-J/K (Collocation-Based Language Translator) (Park 1997). The concept co-occurrence information consists of two parts: local syntactic patterns (LSPs) and unordered co-occuring words (UCWs) encoded with the concept codes of the Kadokawa thesaurus (Ohno and Hamanishi 1991). To improve the recall rate and also to save storage space while keeping high precision, the extracted conceptual information is type-abstracted into higher levels. A WSD algorithm is executed at four levels for nominal words and two levels for verbal words, each level uses different knowledge like selectional restrictions of verbs, concept co-occurrence information (LSPs and UCWs), and heuristics. In one experiment, our method achieved, on the average, a precision of 82.4% for nouns, and 83% for verbs.

The method proposed consists of a learning phase and an application phase as shown in Figure 1. The outline of the rest of this paper is as follows: Section 2 describes how to extract concept co-occurrence information from a sense-tagged corpus through partial parsing and statistical processing. Section 3 presents an algorithm for WSD using multiple knowledge sources. In section 4, some

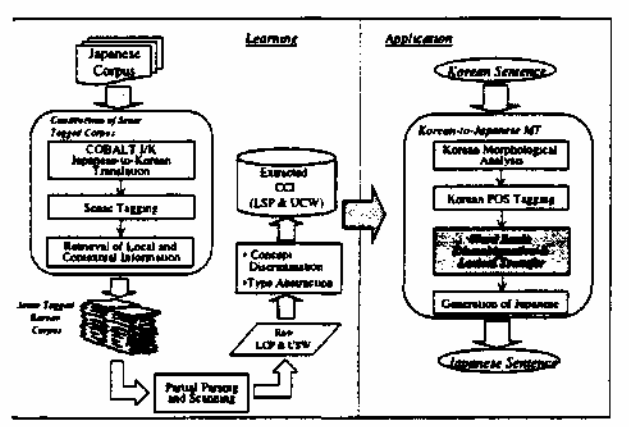

Figure 1. System Architecture

experimental results are given showing that the proposed method may be useful for WSD in real texts. Concluding remarks will be given in section 5. In this paper, Yale Romanization is used for representing Korean expressions.

2 Extraction of Conceptual Co-occurrence Information

2.1 Automatic Construction of the Sense-Tagged Corpus

For automatic construction of the sense-tagged corpus, we used a Japanese-to-Korean MT system called COBALT-J/K (Park 1997). COBALT-J/K is a high-quality practical MT system developed by POSTECH Pohang University of Science and Technology) in 1996. I: has been used successfully in full operation at the Poaang Iron and Steel Company, Korea, to translate patent materials on iron and steel subjects. In a transfer dictionary of COBALT-J/K, nominal and verbal words are annotated with concept codes of the Kadokawa thesaurus (Ohno and Hamanishi 1981), which has a 4-level hierarchy of about 1,100 semantic classes, as shown in Figure 2. Concept nodes in level L_1, L_{10} and L_{100} are further divided into 10 subclasses. For each Japanese word, which

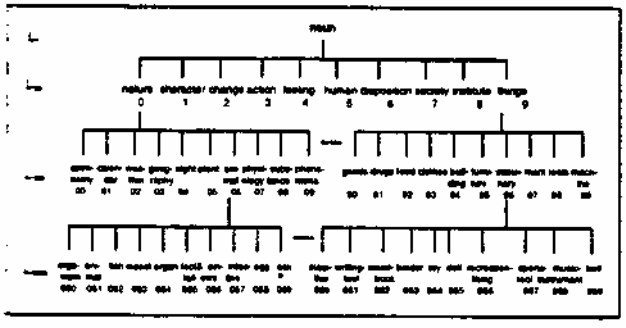

Figure 2. Conceptual Hierarchy of Kadokawa Thesaurus

may be translated into different Korean equivalents, it has several collocation patterns in the transfer dictionary, one for each of the different Korean translations, which specifies a surrounding context necessary for lexical transfer. We made a slight modification of COBALT-J/K so that it can produce Korean translations from Japanese texts with all nominal and verbal words tagged with the most specific codes at level L_{1000} of the Kadokawa thesaurus. As a result, we have obtained a Korean sense-tagged corpus.

2.2 Extraction of LSP and UCW

In English, local collocations can be defined by word order to be common expressions containing the word to be disambiguated (Hwee and Lee 1996). Unlike English, Korean has almost no syntactic constraints on word order as long as a verb appears in the final position. The variable word order often results in discontinuous constituents. Consequently, we defined 10 local syntactic patterns (LSPs) for homographs using syntactically related words in a sentence as shown in Table 1. Examples of LSPs for the homograph *nwun, cito, pay,* and *ssu-ta* are shown in Table 1. The concept codes beginning with 'v' are for verbal words, and those with 'n' are for nominal words.

Frequently co-occurring words in a sentence, which have no syntactic relations with homographs, are retrieved as the unordered co-occurring words (UCWs). Examples of UCWs for *nwun* with the sense 'eye' are *koyangi* (n061: cat), *mosup* (n110, n620: appearance), and *salam* (n507: human); and examples of UCWs for *nwun* with the sense 'snow' are *yengha* (n126: below zero), *san* (n032: mountain), and *salam* (n507: human).

2.3 Discrimination of Co-occurring Concept Types

In the extracted LSPs and UCWs, however, the same concepts may appear for determining different meanings of a word. We can observe from Table 1 and from the UCWs for *nwun* listed above, that there exist the same co-occurring concepts in the LSPs and UCWs, such as ***khu-ta*** and ***salam***, for the different meanings of a word. To select the most probable concept types, which frequently co-occur with a sense of a homograph, Shannon's entropy model (Shannon 1951) is adopted to define the noise of a concept type in discriminating the homograph (Lee 1997). Let S_i be the i^{th} word sense of a homograph W, C_k the concept type of its co-occurring word, $p(C_k,S_1)$ the probability of concept C_k to co-occur with a word sense S_i in a sentence, and n the number of word senses of W. The noise that is generated by concept type C_k is defined by Formula 1. The smaller the noise of C_k is, the more contribution C_k has on deciding the word sense S_i. Using this definition, we define the discrimination value DS_k of concept type C_k for word sense S_i by Formula 2.

$$noise_k = -\sum_{i=1}^{n} \frac{p(C_k \mid S_i)}{\sum_{j=1}^{n} p(C_k \mid S_j)} \log_2 \frac{p(C_k \mid S_i)}{\sum_{j=1}^{n} p(C_k \mid S_j)} \qquad (1)$$

Table 1. Local Collocation Patterns (LSPs) and UCW

Pattern	Structure of collocation	Homograph	LSPs
$type_1$	**noun** + noun	*nwun (n024:snow)*	songi (n114:flake), pola(n024:)
$type_2$	noun + *uy* + **noun**	*nwun (n613:eye)*	caki (n501:oneself),
$type_3$	**noun** + *other particles* + noun	*nwun (n613:eye)*	kwi (n613:ear)
$type_4$	**noun** + *lo/ulo* + verb	*kutwu(n340:speech)*	selmyoung-hata (explain)
$type_5$	**noun** + *ey* + verb	*nwun (n613:eye)*	poi-ta (v312:seeable)
$type_6$	**noun** + *eygey* + verb	*chelswu(n500:human)*	cwu-ta (v370:give)
$type_7$	**noun** + *eyse* + verb	*cito (n824:map)*	chac-ta (v426:search)
$type_8$	**noun** + *ul/lul* + verb	*pay (pear:057)*	mek-ta (v354:eat)
$type_9$	**noun** + *i/ka* + verb	*nwun (n613:eye)*	penttuk-ita(v312:), **khu-ta** (v243:big)
		nwun (n053:bud)	**khu-ta** (v243:big)
$type_{10}$	verb + *relativizer* + **noun**	*nwun (n613:eye)*	chonghyeltoyn (v076:blood-shot)
$type_{11}$	noun + *i/ka* + noun + *ul/lul* + noun + {*ey/eyse/lo*} + **verb**	*ssu-ta (v339:write)*	ku-nun (n500:he) chak-ul (n847:book)
$type_{12}$	noun + *i/ka* + noun + {*ey/eyse/lo*} + **verb**	*ssu-ta (v144:taste bitter)*	mas-i (n144:favor, taste)

$$DS_k = \frac{log_2 n - noise_k}{log_2 n} \quad (2)$$

The value of $noise_k$ will be between 0 and $log_2 n$. If the discrimination value DS_k of concept type C_k is larger than 0.7, the concept type is selected as useful information for determining word sense S_j; if it is not, then we discard it. If the DS_k values of concept type C_k are the same for word sense S_i and S_j, then this means the concept type Ck has no contribution to the discrimination of word sense of W. In this case, we just disregard code C_k although the DS_k values for both senses may be over 0.7.

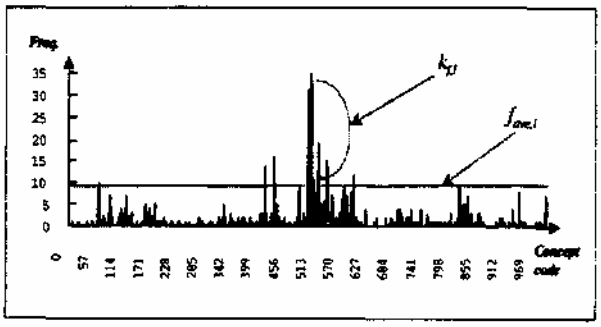

Figure 3. Histogram for Concept Type Frequency.

2.4 Concept Type Generalization for an Abstracted CCI

For the purpose of type generalization of words in LSP and UCW, the Kadokawa thesaurus is used. In this paper, the symbol L_{1000} indicates the lowest-level semantic classes, L_{100} the next higher level, and so on. All words in LSPs and UCWs are annotated with the concept type of the three-digit form at level L_{1000} in the Kadokawa thesaurus. For each word sense of a homograph W, the frequency of concept codes in LSPs and UCWs shows a very different distribution so that the distribution of concept codes may be used together with their frequency as a clue to the word sense disambiguation of the homograph. Table 2 shows the concept types that can co-occur with the homograph *nwun* (eye) in the form of LSP $type_2$, and their frequencies.

For a homograph W, conceptual frequency patterns (CFPs), i.e., ({<$C_1 f_1$>, <$C_2 f_2$>, ...,<$C_k f_k$>}, $type_i$, $W(S_i)$), are constructed for each type of LSP, where f_i is the frequency (number of appearances) of concept C_i appearing in the corpus, $type_i$ is an LSP type, and $W(S_i)$ is a homograph W with the sense S_i. CFPs for UCWs can have the form of ({<$C_1 f_1$>, <$C_2 f_2$>, ...,<$C_k f_k$>), UCW, $W(S_i)$). To perform type abstraction, we refer to Smadja's work (Smadja 1990, 1993), and define the standard deviation σ_ℓ of the code frequency at level ℓ (denoted as L_ℓ) and $k_{f,\ell}$ (the strength of code frequency f at L_ℓ and represents the amount of standard deviation above the average frequency $f_{ave.\ell}$). In the formulas, $f_{k,\ell}$ denotes the frequency of concept code C_k of the Kadokawa thesaurus at L_ℓ, and n_ℓ the number of concept codes at L_ℓ.

$$\sigma_\ell = \sqrt{\frac{\sum_{k=1}^{n_\ell} (f_{k,\ell} - f_{ave,\ell})^2}{n_\ell - 1}} \quad (3)$$

$$k_{f_{k,\ell},\ell} = \frac{f_{k,\ell} - f_{ave,\ell}}{\sigma_\ell} \quad (4)$$

The standard deviation $\sigma_{/}$ at $L_{/}$ characterizes the shape of the distribution of code frequencies. If $\sigma_{/}$ is small, then the shape of the histogram for concepts and their frequencies will tend to be flat, which means that the concept codes have a low discrimination power between themselves. If $\sigma_{/}$ is large, there is one or more codes that tend to be peaks in the histogram, and its corresponding concept codes are likely to be typical concepts that can restrict the sense of homograph *W*.

The filter in our system selects the concept codes that have a variation larger than threshold $\sigma_{0,/}$ and pulls out the concept codes that have a strength of frequency larger than threshold $k_{0,/}$ If the value of the $\sigma_{/}$ is small, than it can be assumed that there is no peak frequency of the code for the pattern. The concept codes for a local syntactic pattern or co-occurring words that are produced by the filter should represent the concept types of extracted words that appear most frequently with sense Si of the word *W*.

Through experiments, the threshold of the standard deviation $\sigma_{0.1000}$ and $\sigma_{0.100}$ are fixed at 1.5 and 4, and the strength of frequency $k_{0.1000}$ *and* $k_{0.100}$ are fixed at 6.0 and 3.0. The lower the value of threshold that $\sigma_{0,/}$ and $k_{0,/}$ are assigned, the more concept codes can be extracted as conceptual patterns from the CFPs. A balance is maintained between extracting conceptual codes at low levels of the conceptual hierarchy for the specific usage of concept type and extracting general concept types at higher levels for enhancing overall system performance. These values may be variable in different applications.

In Table 2, since the value of strength $k_{11.1000}$ for a code with a frequency of 11 is 6.02, and the value of $k_{0.1000}$ is set at 6.0, the concept codes with frequencies of more than 11 are selected as generalized concept types at L_{1000}. After abstraction at L_{1000}, the system performs generalization at L_{100}. It removes selected frequencies and codes, such as code 504 and its frequency in Table 2, and sums

Table 2. Concept Types and Frequencies in CFP

$(\{<C_i, f_i>\}, type_2, nwun(\text{eye}))$

Code	Freq.	Code	Freq.	Code	Freq.
28	12	61	7	103	4
107	8	121	7	126	4
143	8	160	5	179	7
277	4	320	8	331	6
416	7	419	12	433	4
501	13	503	10	504	11
505	6	507	12	508	27
513	5	530	6	538	16
552	4	557	7	573	5
709	5	718	5	719	4
733	5	819	4	834	4
966	4	987	9	other	210

* 'other': indicates the total codes with frequency of less than 4

up the frequencies of the remaining concept codes to form the CFPs with code of a higher level of conceptual hierarchy. After processing, the system selects the most promising codes and stores conceptual patterns *($\{C_1, C_2, C_3, ...\}$, $type_i$, $W(S_i)$)* as a knowledge source for WSD of real texts. The type abstracted LSP for $type_2$ of the noun *nwun*(eye) is ({n028, n419, n501, n504, n507, n508, n538,.., n50,...}, $type_2$, *nwun*(eye)) in this case.

2.5 Compensation for the Lost Senses

While automatically constructing a sense-tagged corpus through COBALT-J/K, some senses of a Korean word cannot be produced from their corresponding Japanese equivalents. For example, as shown in Figure 4, the sense S_3 (bud) of the Korean word *nwun* cannot be generated from its corresponding Japanese translations although the Japanese word *me-2* has the same sense as 'bud'. This is because a Japanese word may be translated into two or more target words with the same meaning, but

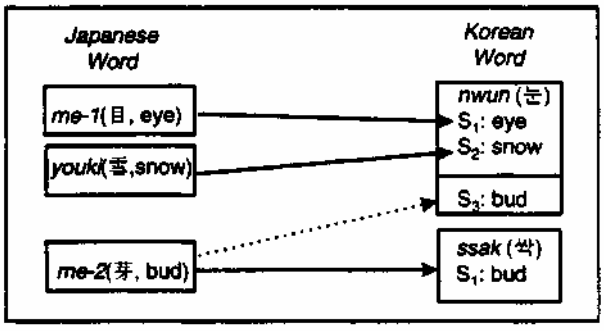

Figure 4. Japanese Words and Korean Translations

out of them only one translation equivalent is defined in our bilingual dictionary. Such a word sense like S_3 of *nwun* is called a "lost sense" in this paper. Since the lost sense makes it difficult to collect all LSPs necessary for the word, we made up for its LSPs by manually adding the LSPs of another word with a similar sense.

3 Algorithm for Word Sense Disambiguation Using CCI

For a given homograph *W,* the algorithm below describes the overall processing flow of CCI-based word sense disambiguation. Similarity calculations for word sense disambiguation using extracted knowledge are defined by Formulas 5, 6, 7 and 8, where *S(W)* denotes a set of word senses of the homograph *W, SR(V)* a selectional restriction (SR) of a verb *V* ($type_{11}$ and $type_{12}$ in LSP) that takes the word *W* as its argument, *LSP(W)* its abstracted local syntactic patterns, and *UCW(W)* its abstracted concept types of unordered co-occurring words. In the formulas, C_i and P_j are concept types and S_k indicates a word sense of *W.* Formula 5 is used to calculate concept similarity between senses of a homograph and concept types in verb selectional restriction. Formulas 6 and 7 are used to cal-

culate the concept similarity between senses of word *W* and concept types in LSP and UCW respectively. $Csim(C_i, P_j)$ in Formula 8 is used to compute the concept similarity between C_i and P_j, where $MSCA(C_i, P_j)$ is the most specific common ancestor of concept types C_i and P_j, and *weight* is a weighting factor reflecting that C_i as a descendant of P_j is preferable to other cases. In Figure 5, the concept similarity values between C_i and P_1, P_2, and P_3 are larger than 0.3, but less than 0.3 between C_i and P_4 or other nodes. Therefore, the threshold *T* is set to 0.3 in our experiment considering the property of the thesaurus hierarchy.

Algorithm (CCI-based Word Sense Disambiguation of Verbs and Nouns)
For a given ambiguous word W,

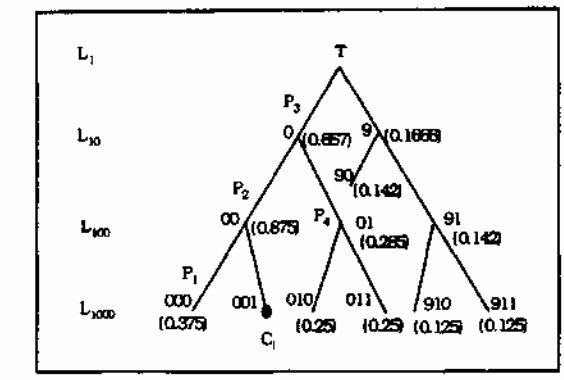

Figure 5. Concept Type Similarity Calculation

Step 1. If W is a verb, then calculate *Score=Vsim(S(W), SR(V))*. If *Score* is greater than threshold *T,* then set the sense of *W* to the S_i with the maximum *Csim* value and exit. Otherwise, set the verb sense to the default sense, which appears most frequently in a corpus, and exit.

Step 2. If W is a noun, then consider executing Steps 2. 3 and 4 if necessary. At first calculate *Score=Vsim(S(W), SR(V))* and determine S_i by it if *Score* is greater than threshold *T,* else go to Step 3.

Step 3. If *Score=Lsim(S(W), LSP(W))* is greater than *T*, then decide S_k by it, else continue to Step 4.

Step 4. *If Score=Ssim(S(W), USW(W))* is greater than *T*, then decide S_k by it, else set S_n to the default sense (the most frequently appearing sense in a corpus).

$$Vsim(S(W),SR(V)) = \max_{i}(Csim(C_i,P_j)) \quad (5)$$
$$1 \le i \le n, 1 \le j \le m, C_i \in S(W), P_j \in SR(V)$$

$$Lsim(S(W),LSP(W)) = \max_{k}(Csim(C_i,P_{k,j,l})) \quad (6)$$
$$1 \le i,k \le n; 1 \le j \le 10; 1 \le l \le w; P_{k,j,l} \in LSP_j(S_k)$$

$$Ssim_t(S(W),UCW(W)) = \max_{k}(Csim(C_i,P_{k,j})) \quad (7)$$
$$1 \le i,k \le n, 1 \le j \le r, P_{k,j} \in UCW(S_k)$$

$$Csim(C_i,P_j) = \frac{2*level(MSCA(C_i,P_j))}{level(C_i)+level(P_j)} * weight \quad (8)$$

4 Experimental Evaluation

Table 3. Experimental Results of Noun Sense Disambiguation

Homograph	Sense	No.	*SR*		*LSP*		*UCW*		*Freq.*		Total No.	
			O	X	O	X	O	X	O	X	O	X
Pwuca	father & child	40			4	6	16	2	12		32	8
	rich man	12					4	3		5	4	8
Kancang	liver	33			3		13	3	14		30	3
	soy sauce	16	1		2		10			3	13	3
Kasa	housework	25			6		12	1	6		24	1
	words of song	27					22			5	22	5
Kwutwu	shoes	39	9		3		16	2	9		37	2
	word of mouth	10					5	2		3	5	5
Nwun	eye	41	6		8		3		24		41	0
	snow	9	4		1		2	1		1	7	2
Yongki	courage	31			1	2	17		11		29	2
	container	19	1		6		1	3		8	8	11
Kyengpi	expenses	38	3		13		7	6	9		32	6
	defense	13	2			2	6			3	8	5
Kyeongki	times	27	1		7		8	1	10		26	1
	match	24		1			15	3		5	15	9
Average Precision		404	96 %		84.3 %		85.3 %		74.2 %		82.4 %	

Table 4. Experimental Results of Verb Sense Disambiguation

Homograph	Sense	No.	Manually-coded *SR*		Auto-detected *SR*	
			O	X	O	X
nayli-ta	get off (bus)	14	14	0	14	0
	draw (a decision)	21	11	10	12	9
	fall (snow)	15	10	5	10	5
seywu-ta	make (a plan,)	27	27	0	27	0
	build	23	14	9	17	6
ssu-ta	use	23	23	0	22	1
	write	17	10	7	14	3
	put on (hat)	10	4	6	7	3
taywu-ta	burn	25	24	1	24	1
	give some one a ride	25	19	6	19	6
Average Precision		200	78%		83%	

For experimental evaluation, eight nouns and four verbs that are ambiguous were selected together with a total of 604 test sentences in which one of the homographs appears. The test sentences were randomly selected from KIBS (Korean Information Base System, '94-'97), a large-scale corpus. Out of several senses of each homograph, we considered only two or three senses that are most frequently used in the corpus. For each sense of a homograph, the number of its appearances in the test sentences is shown in Table 3 and Table 4. The mark 'O' indicates whether the word sense disambiguation was correctly performed by the system, and 'X' indicates that it was not. Symbols *SR* (selectional restriction), *LSP, UCW,* and *Freq.* in the tables indicate used knowledge by the steps.

The experiment achieved a 14.6% improvement of the baseline for nouns (with an average precision of 82.4%). Here, the baseline is the case where the most frequently used sense is always taken as the sense of the homographs. The experimental results also show that the selecoonal restriction of verbs has low applicability with the highest precision, and that the co-occurrence surrounding words and local collocation patterns have higher applicability and precision than the default word senses, which appear most frequently in a corpus.

Table 4 shows the experimental results for verb sense disambiguation using manually-coded selectional restriction (SR) and automatically constructed SR in **CCI.** The CCI-based WSD achieved a 5% improvement over the use of manually-coded knowledge, which is a promising result considering that the test corpus is completely irrelevant to the learning corpus and that the knowledge is extracted without laborious work by a human.

5 Conclusion

To resolve the lexical transfer ambiguity in Korean-to-Japanese machine translation, this paper proposes a word sense disambiguation method using corpus and conceptual information. Unlike most of previous work that has restricted the use of their linguistic knowledge to the lexical level only, we rely on a concept-level knowledge, called concept co-occurrence information, that is extracted from a sense-tagged corpus. Local syntactic patterns and unordered co-occurring information are extracted from a large-scale sense-tagged corpus, which is automatically constructed by an existing high-quality Japanese-to-Korean MT system. The extracted local and co-occurrence information is type-abstracted using the Kadokawa thesaurus, which enables the method to be more robust to the data sparseness problem and also to deterioration caused by domain shifts. WSD using concept co-occurrence information for both ambiguous nouns and verbs showed on average 82.4% and 83% accuracy respectively. We will do further research on how to automatically or semi-automatically compensate for the lost senses.

References

Kenneth Ward Church (1990). "Word Association Norms, Mutual Information, and Lexicography." Computational Linguistics, Vol. 16, No. 1.

Tou Ng Hwee and Hian Beng Lee (1996). "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: an Exemplar-Based Approach." In proceedings of 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, California, pp.40-47.

Edward Kelly and Phillip Stone (1975). Computer Recognition of English Word Senses, North-Holland, Amsterdam.

Rey-Long Liu, Von-Wun Soo (1993). "An Empirical Study on thematic Knowledge acquisition Based on Syntactic Clues and Heuristics." In proceedings of 31th Annual Meeting of the Association for Computer Linguistics, Ohio State University, Columbus, pp.243-250.

Susan W. McRoy (1992). "Using Multiple Knowledge Sources for Word Sense Discrimination." Computational Linguistics, Vol.18, No.l, pp. 1-30.

S. Ohno and M. Hamanishi (1981). New Synonym Dictionary, Kadokawa Shoten, Tokyo. (Written in Japanese).

Chul-Jae Park, Jong-Hyeok Lee, Geunbae Lee and K. Kakechi (1997). "Collocation-Based Transfer Method in Japanese-Korean Machine Translation." Transactions of Information Processing Society of Japan, Vol.38, No.4, pp.707-718. (Written in Japanese).

C. E. Shannon (1951). "Prediction and Entropy of Printed English." Bell System Technical Journal, pp.50-65.

Frank Smadja (1990). "Automatically Extracting and Representing Collocations for Language Generation." In proceedings of 28th Annual Meeting of the Association for Computational Linguistics, Pittsburgh, Pennsylvania, pp. 252-259.

Frank Smadja (1993). "Retrieving Collocations from Text: Xtract." Computational Linguistics, Vol.19, No.l, pp.143-177.

Hideki Tanaka (1994). "Verbal Case Frame Acquisition from a Bilingual Corpus: Gradual Knowledge Acquisition." In proceedings of the 15th International Conference on Computational Linguistics (COLING'94), Kyoto, Japan, pp. 727-731.

David Yarowsky (1995). "Unsupervised Word Sense Disambiguation Rivaling Supervised Methods." In proceedings of 33rd Annual Meeting of the Association for Computational Linguistics, Cambridge, Massachusetts, pp.189-196.