Augmenting a Bilingual Lexicon with Information for Word Translation Disambiguation

Takashi Tsunakawa
Faculty of Informatics
Shizuoka University
`tuna@inf.shizuoka.ac.jp`

Hiroyuki Kaji
Faculty of Informatics
Shizuoka University
`kaji@inf.shizuoka.ac.jp`

Abstract

We describe a method for augmenting a bilingual lexicon with additional information for selecting an appropriate translation word. For each word in the source language, we calculate a correlation matrix of its association words versus its translation candidates. We estimate the degree of correlation by using comparable corpora based on these assumptions: "parallel word associations" and "one sense per word association." In our word translation disambiguation experiment, the results show that our method achieved 42% recall and 49% precision for Japanese-English newspaper texts, and 45% recall and 76% precision for Chinese-Japanese technical documents.

1 Introduction

The bilingual lexicon, or bilingual dictionary, is a fundamental linguistic resource for multilingual natural language processing (NLP). For each word, multiword, or expression in the source language, the bilingual lexicon provides translation candidates representing the original meaning in the target language.

Selecting the right words for translation is a serious problem in almost all of multilingual NLP. One word in the source language almost always has two or more translation candidates in the target language by looking up them in the bilingual lexicon. Because each translation candidate has a distinct meaning and property, we must be careful in selecting the appropriate translation candidate that has the same sense as the word inputted. This task is often called word translation disambiguation.

In this paper, we describe a method for adding information for word translation disambiguation into the bilingual lexicon. Comparable corpora can be used to determine which word associations suggest which translations of the word (Kaji and Morimoto, 2002). First, we extract word associations in each language corpus and align them by using a bilingual dictionary. Then, we construct a word correlation matrix for each word in the source language. This correlation matrix works as information for word translation disambiguation.

We carried out word translation experiments on two settings: English-to-Japanese and Chinese-to-Japanese. In the experiments, we tested Dice/Jaccard coefficients, pointwise mutual information, log-likelihood ratio, and Student's t-score as the association measures for extracting word associations.

2 Constructing word correlation matrices for word translation disambiguation

2.1 Outline of our method

In this section, we describe the method for calculating a word correlation matrix for each word in the source language. The correlation matrix for a word f consists of its association words and its translation candidates. Among the translation candidates, we choose the most acceptable one that is strongly suggested by its association words occurring around f.

We use two assumptions for this framework:
(i) Parallel word associations:
Translations of words associated with each other in a language are also associated with each other in another language

(Rapp, 1995). For example, two English words "tank" and "soldier" are associated with each other and their Japanese translations "戦車 (*sensha*)" and "兵士 (*heishi*)" are also associated with each other.

(ii) One sense per word association:

A polysemous word exhibits only one sense of a word per word association (Yarowsky, 1993). For example, a polysemous word "tank" exhibits the "military vehicle" sense of a word when it is associated with "soldier," while it exhibits the "container for liquid or gas" sense when it is associated with "gasoline."

Under these assumptions, we determine which of the words associated with an input word suggests which of its translations by aligning word associations by using a bilingual dictionary. Consider the associated English words (tank, soldier) and their Japanese translations (戦車 (*sensha*), 兵士 (*heishi*)). When we translate the word "tank" into Japanese, the associated word "soldier" helps us to translate it into "戦車 (*sensha*)", not to translate it into "タンク (*tanku*)" which means "a storage tank."

This naive method seems to suffer from the following difficulties:

● A disparity in topical coverage between two corpora in two languages
● A shortage in the bilingual dictionary
● The existence of polysemous associated words that cannot determine the correct sense of the input word

For these difficulties, we use the tendency that the two words associated with a third word are likely to suggest the same sense of the third word when they are also associated with each other. For example, consider an English associated word pair (tank, troop). The word "troop" cannot distinguish the different meanings because it can co-occur with the word "tank" in both senses of the word. The third word "soldier," which is associated with both "tank" and "troop," can suggest the translation "戦車 (*sensha*)."

The overview of our method is shown in Figure 1. We first extract associated word pairs in the source and target languages from comparable corpora. Using a bilingual dictionary, we obtain alignments of these word associa-
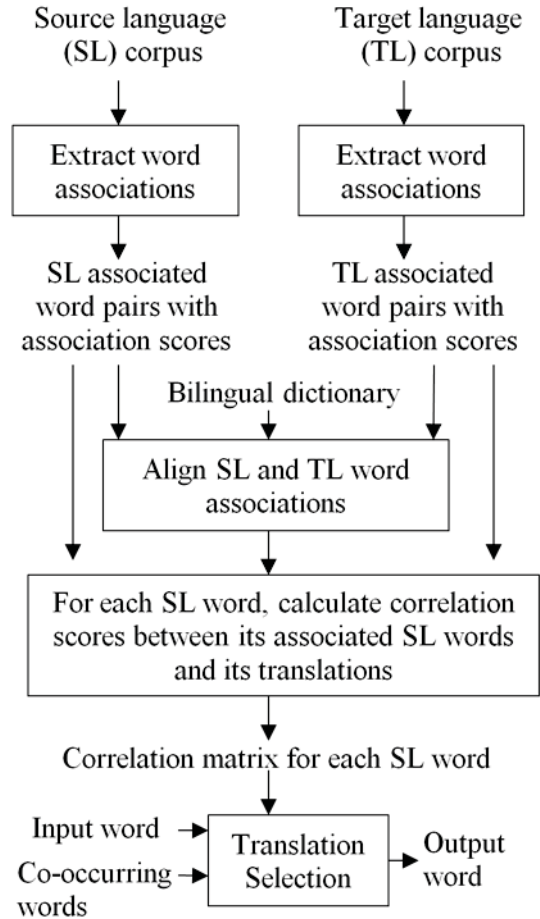

Figure 1. Overview of our method.

tions. Then, we iteratively calculate a correlation matrix for each word in the source language. Finally, we select the translation with the highest correlation from the translation candidates of the input word and the co-occurring words.

For each input word in the source language, we calculate correlation values between their translation candidates and their association words. The algorithm is shown in Figure 2.

In Algorithm 1, the initialization of correlation values is based on word associations, where D is a set of word pairs in the bilingual dictionary, and A_f and A_e are the sets of associated word pairs. First, we retain associated words $f'(i)$ when its translation e' exists and when e' is associated with e. In the iteration, the correlation values of associated words $f'(i)$ that suggest $e(j)$ increase relatively by using association scores $\alpha(f'(i), f)$ and

31

$\alpha(f'(i), f'')$. In our experiments, we set the number of iterations N_r to 10.

2.2 Alternative association measures for extracting word associations

We extract co-occurring word pairs and calculate their *association scores*. In this paper, we focus on some frequently used metrics for finding word associations based on their occurrence/co-occurrence frequencies.

Suppose that words x and y frequently co-occur. Let n_1 and n_2 be the occurrence frequencies of x and y respectively, and let m be the frequency that x and y co-occur between w content words. The parameter w is a window size that adjusts the range of co-occurrences.

Let N and M be the sum of occurrences/co-occurrences of all words/word pairs, respectively. The frequencies are summarized in Table 1.

The word association scores $\alpha(x, y)$ are defined as follows:
- Dice coefficient (Smadja, 1993)

$$\text{Dice}(x, y) = \frac{2m}{n_1 + n_2} \quad (1)$$

- Jaccard coefficient (Smadja et al., 1996)

$$\text{Jaccard}(x, y) = \frac{m}{n_1 + n_2 - m} \quad (2)$$

- Pointwise mutual information (pMI) (Church and Hanks, 1990)

$$\text{pMI}(x, y) = \log_2 \frac{m/M}{(n_1/N)(n_2/N)} \quad (3)$$

- Log-likelihood ratio (LLR) (Dunning, 1993)

$$\begin{aligned} &\text{LLR}(x, y) \\ &= -2\big(\text{logL}(m, n_1, r) \\ &+ \text{logL}(n_2 - m, N - n_1, r) \\ &- \text{logL}(m, n_1, r_1) \\ &- \text{logL}(n_2 - m, N - n_1, r_2)\big); \end{aligned} \quad (4)$$

$$\begin{aligned} &\text{logL}(k, n, r) = \\ &k \log_2 r + (n - k) \log_2(1 - r), \end{aligned} \quad (5)$$

$$r_1 = \frac{m}{n_1}, r_2 = \frac{n_2 - m}{N - n_1}, r = \frac{n_2}{N} \quad (6)$$

- Student's *t*-score (TScore) (Church et al., 1991)

$$\text{TScore}(x, y) = \frac{m - n_1 n_2 / N}{\sqrt{m}} \quad (7)$$

We calculate association scores for all pairs of words when their occurrence frequencies are not less than a threshold T_f and when their

Algorithm 1:
Input:
 f: an input word
 $f'(1), \ldots, f'(I)$: associated words of f
 $e(1), \ldots, e(J)$: translation candidates of f
 N_r: number of iterations
 A bilingual lexicon
 Word association scores α for both languages
Output:
 $C_f = [C_f(f'(i), e(j))]$: a correlation matrix for f

1: **if** $\sum_k \delta_f\big(f'(i), e(k)\big) \neq 0$ **then**

2:     $C_f\big(f'(i), e(j)\big) \leftarrow \dfrac{\delta_f\big(f'(i), e(j)\big)}{\sum_k \delta_f\big(f'(i), e(k)\big)}$

3: **else**

4:     $C_f\big(f'(i), e(j)\big) \leftarrow 0$

5: **end**

6: $i \leftarrow 0$

7: **while** $i < N_r$

8:     $i \leftarrow i + 1$

9:     $C_f\big(f'(i), e(j)\big) \leftarrow \alpha(f'(i), f)$
        $\times \dfrac{\sum_* \alpha(f'(i), f'') \cdot C_f(f'', e(j))}{\max_k \sum_* \alpha(f'(i), f'') \cdot C_f(f'', e(k))}$

10: **end**

$$\begin{pmatrix} \delta_f(f', e) = \\ \begin{cases} 1 & (\exists e': (e, e') \in A_e, (f', e') \in D) \\ 0 & (\text{otherwise}) \end{cases} \end{pmatrix}$$
$$\big(\Sigma_* := \Sigma_{\{f''|(f, f'') \in A_f, (f'(i), f'') \in A_f\}}\big)$$

Figure 2. Algorithm for calculating correlation matrices.

	x occur	x not occur	Total
y occur	m	$n_2 - m$	n_2
y not occur	$n_1 - m$	$M - n_1 - n_2 + m$	$N - n_2$
Total	n_1	$N - n_1$	N

Table 1. Contingency matrix of occurrence frequencies.

<home>	国 (*kuni*) [country]	本塁 (*honrui*) [home base]	家 (*ie*) [house]	自宅 (*jitaku*) [my home]	家庭 (*katei*) [homeplace]	施設 (*shisetsu*) [facilities]
base	0.009907	**0.017649**	0.005495	0.006117	0.005186	0.005597
game	0.043507	**0.048358**	0.025145	0.028208	0.019987	0.023014
Kansas	**0.010514**	0.003786	0.004280	0.007307	0.004320	0.005459
run	0.023468	**0.042035**	0.014430	0.015765	0.012061	0.012986
season	0.044855	**0.050952**	0.025406	0.028506	0.020716	0.023631

Table 2. Word correlation matrix for a word "home,"
as information for word translation disambiguation

threshold T_c.

We handle word pairs whose association scores are not less than a predefined value T_A; some of the thresholds were evaluated in our experiment. The associated word pair sets A_f and A_e in Algorithm 1 includes only word pairs whose scores are not less than T_A in the source and target language, respectively.

3 Word Translation Disambiguation

Consider that a translator changes a word in an input sentence. Usually, two or more translation candidates are enumerated in the bilingual dictionary for a word. The translator should select a translation word that is grammatically/ syntactically correct, semantically equivalent to the input, and pragmatically appropriate.

We assume that the translation word e for an input word f tends to be selected if words occurring around f are strongly correlated with e. Using the correlation matrices, we select e as a translation if the associated word f' occurs around f and the score $C_f(f',e)$ is large. In addition, we take distance between f and f' into account.

We define the score of the translation word $e(f_0)$ for an input word f_0 as follows. Consider an input word f_0 that occurs in the context of "$\ldots f_{-2}\ f_{-1}\ f_0\ f_1\ f_2\ \ldots$" The score for a translation word $e(f_0)$ for an input word f_0 is

$$\text{Score}\big(e(f_0)\big)$$
$$= \sum_{1 \le |p| \le \gamma} \frac{1}{\sqrt{|p|}} C_{f_0}\big(f_p, e(f_0)\big), \quad (8)$$

defined as where p is the relative position of the words surrounding f_0, $C_{f_0}\big(f_p, e(f_0)\big)$ is the value of the correlation matrix for f_0, and γ

A career .284 hitter, Beltran batted .267 in the regular *season*, split between *Kansas* City and Houston, but came alive in the playoffs. He hit .435 in 12 postseason *games*, with six stolen *bases*, eight **home** *runs* and 14 *runs* batted in.

Figure 3. An example of an input word "home"

is the window size for word translation disambiguation.

A simple example is shown in Figure 3. The word "home" in this context means the sense of "home base" used in baseball games, not "house" or "hometown." The surrounding words such as "games," "bases," and "runs" can be clues for indicating the correct sense of the word. By using the correlation matrix (Table 2) and formula (8), we calculate a score for each translation candidate and select the best translation with the largest score. In this case, Score(本塁 (*honrui*)) = 0.1134 was the best score and the correct translation was selected.

4 Experiment

4.1 Experimental settings

We carried out word translation experiments on two different settings. In the first experiment (Experiment A), we used large-scale comparable corpora from English and Japanese newspaper texts. The second experiment (Experiment B) was targeted at translating technical terms by using Chinese and Japanese domain-specific corpora.

We used the following linguistic resources for the experiments:
● Experiment A

- ■ Training comparable corpora
  - – The New York Times texts from English Gigaword Corpus Fourth Edition (LDC2009T13): 1.6 billion words
  - – The Mainichi Shimbun Corpus (2000-2005): 195 million words
- ■ Test corpus
  - – A part of The New York Times (January 2005): 157 paragraphs, 1,420 input words
- ● Experiment B1
  - ■ Training comparable corpus
    - – In-house Chinese-Japanese parallel corpus in the environment domain: 53,027 sentence pairs[1]
  - ■ Test corpus
    - – A part of the training data: 1,443 sentences, 668 input words
- ● Experiment B2
  - ■ Training comparable corpus
    - – In-house Chinese-Japanese parallel corpus in the medical domain: 123,175 sentence pairs
  - ■ Test corpus
    - – A part of the training data: 940 sentences, 3,582 input words
- ● Dictionaries
  - ■ Japanese-English bilingual dictionaries: Total 333,656 term pairs
    - – EDR Electronic Dictionary
    - – Eijiro, Third Edition
    - – EDICT (Breen, 1995)
  - ■ Chinese-English bilingual dictionary
    - – Chinese-English Translation Lexicon Version 3.0 (LDC 2002L27): 54,170 term pairs
    - – Wanfang Data Chinese-English Science/Technology Bilingual Dictionary: 525,259 term pairs

For the Chinese-Japanese translation, we generated a Chinese-Japanese bilingual dictionary by merging Chinese-English and Japanese-English dictionaries. The Chinese-Japanese bilingual dictionary includes every Chinese-Japanese term pair (t_C, t_J) when (t_C, t_E) and (t_J, t_E) were present in the dictionaries for one or more English terms t_E. This merged dictionary contains about two million term pairs. While these Chinese-Japanese term pairs include wrong translations, it was not a serious problem in our experiments because wrong translations were excluded in the procedure of our method.

We applied morphological analysis and part-of-speech tagging by using TreeTagger (Schmid, 1994) for English, JUMAN for Japanese, and mma (Kruengkrai et al., 2009) for Chinese, respectively.

In the test corpus, we manually annotated reference translations for each target word.[2] The parameters we used were as follows:
Experiment A:
   $T_f = 100$ (Japanese), $T_f = 1000$ (English), $T_c = 4$, $w = 30$, $\gamma = 30$.
Experiment B1/B2:
   $T_f = 100$, $T_c = 4$, $w = 10$, $\gamma = 25$.
Some of the parameters were empirically adjusted.

In the experiments, the matrices could be obtained for 9103 English words (A), 674 Chinese words (B1) and 1258 Chinese words (B2), respectively. In average one word had 3.24 (A), 1.15 (B1) and 1.51 (B2) translation candidates[3] by using the best setting. Table 2 is the resulted matrix for the word "home" in the Experiment A.

4.2 Results of English-Japanese word translation

Table 3 shows the results of Experiment A. We classified the translation results for 1,420 target English words into four categories: True, False, R, and M. When the translation was output, the result was True if the output is included in the reference translations, and it was False otherwise. The result was R when all the associated words in the correlation matrix

---

did not occur around the input word. The result was M when no correlation matrix existed for the input word. We did not select a translation output in these cases. The recall and precision are shown in the parentheses below.

- Recall = (True) / (Number of input words)
- Precision = (True) / (True + False)

Among the settings, we obtained the best results, 42% recall and 49% precision, when we used the Jaccard coefficient for association scores and $T_A = 0$, which means all pairs were taken into consideration. Among other settings, the Dice coefficient achieved a comparable performance with Jaccard.

4.3 Results of Chinese-Japanese word translation

Tables 4 and 5 show the results of Experiment B. In each domain, we tested only the settings on Dice, pMI, and LLR with $T_A = 0$.

In the environmental domain, the pointwise mutual information score achieved the best performance, 45% recall and 76% precision. However, the Dice coefficient gave the best recall (55%) for the medical domain. This result indicates that Experiment B1/B2 had higher precision and more words without the correlation matrix than Experiment A had.

4.4 Discussion

As a result, we could generate bilingual lexicons with word translation disambiguation information for 9103 English words and 1932 Chinese words. Although the number of words might be augmented by changing the settings, the size does not seem to be sufficient as bilingual dictionaries. The availability of larger output should be investigated.

The experimental results show that our method selected correct translations for at least half of the input words if a correlation matrix existed and if the associated words co-occur. Among all input words, at least 40% of the input words can be translated. The bilingual dictionaries included 24.4, 38.6, and 52.0 translation candidates for one input word in Experiment A, B1, and B2, respectively. When we select the most frequent word, the precisions were 7%, 1%, and 1%, respectively. Meanwhile, the average numbers of translation

Score	T_A	True	False	R	M
Dice	0	588 (41%/48%)	627	90	115
	0.001	586 (41%/49%)	619	94	121
	0.01	479 (34%/49%)	507	243	191
Jaccard	0	**594** (**42%**/**49%**)	621	90	115
	0.001	584 (41%/49%)	609	105	122
	0.01	348 (25%/48%)	378	374	320
pMI	0	292 (21%/49%)	309	704	115
	1	293 (21%/49%)	308	703	116
LLR	10	530 (37%/42%)	747	28	115
	100	529 (37%/42%)	744	32	115
T-Score	1	486 (34%/38%)	793	26	115
	4	489 (34%/38%)	787	26	118

Table 3. Results of English-Japanese word translation (A)

Score	T_A	True	False	R	M
Dice	0	**1984** (**55%**/69%)	895	82	621
pMI	0	1886 (53%/**70%**)	804	271	621
LLR	0	1652 (46%/57%)	1246	63	621

Table 4. Results of Chinese-Japanese word translation for environmental domain (B1).

Score	T_A	True	False	R	M
Dice	0	277 (41%/69%)	124	14	253
pMI	0	**303** (**45%**/**76%**)	95	17	253
LLR	0	269 (40%/67%)	131	15	253

Table 5. Results of Chinese-Japanese word translation for medical domain (B2).

candidates in the correlation matrices for one input word are shown in Table 6. These indicate that our method effectively removed noisy

translations from the Chinese-Japanese dictionary merged Japanese-English and Chinese-English dictionaries, and that the association scores contributed word translation disambiguation.

Among the settings, the Jaccard/Dice coefficients were proven to be effective, although pointwise mutual information (pMI) was also effective for technical domains and the Chinese-Japanese language pair. Because the Jaccard/Dice coefficients were originally used for measuring the proximity of sets, these might be effective for collecting related words by using the similarity of kinds of co-occurring words. However, pMI tends to emphasize low-frequency words as associated words. The consequence of this tendency might be that low-frequency associated words do not appear around the input word in the newspaper text.[4]

In most metrics for the association score, the lowest threshold value T_A achieved the best performance. This result indicates that the cut-off of associated words by some thresholds was not effective, although it requires more time and memory space to obtain correlation matrices without cut-off. How to optimize other parameters in our method remains unsolved. More words without the correlation matrix were present in Experiment B1/B2 than in Experiment A because the input word was often a technical term that was not in the bilingual dictionary. The better recall and precision of Experiment B1/B2 came from several reasons, including difference of test sets and language pairs. In addition, it might have an impact on this result that the fact that word translation disambiguation of technical terms is easier than word translation disambiguation of common words.

We handled only nouns as input words and associated words in this study. Considering only the co-occurrence in a fixed window would be insufficient to apply this method to the translation of verbs and other parts of speech. In future work, we will consider syntactic co-occurrence, which is obtained by conducting dependency parsing of the results of a sentence. The correlation between associated words and translation candidates also needs to be re-examined. Similarly, we will handle verbs as associated words to the input nouns by using syntactic co-occurrence.

Score	T_A	Exp.A	Exp.B1	Exp.B2
Dice	0	1.83	0.61	0.91
pMI	0	1.27	0.50	0.79
LLR	10/0	1.34	1.48	2.75

Table 6. Average numbers of translation candidates in the correlation matrices for one input word.

5 Related Work

Statistical machine translation (Brown et al., 1990) automatically acquires knowledge for word translation disambiguation from parallel corpora. Word translation disambiguation is based on probabilities calculated from the word alignment, phrase pair extraction, and the language model. However, much broad context/domain information is not considered. Carpuat and Wu (2007) proposed context-dependent phrasal translation lexicons by introducing context-dependent features into statistical machine translation.

Unsupervised methods using dictionaries and corpora were proposed for monolingual WSD (Ide and Veronis, 1998). They used grammatical information including parts-of-speech, syntactically related words, and co-occurring words as the clues for the WSD. Our method uses a part of the clues for bilingual WSD and word translation disambiguation.

Li and Li (2002) constructed a classifier for word translation disambiguation by using a bilingual dictionary with bootstrapping techniques. We also conducted recursive calculation by dealing with the bilingual dictionary as the seeds of the iteration.

Vickrey et al. (2005) introduced a context as a feature for a statistical MT system and they generated word-level translations. How to introduce the word-level translation disambiguation into sentence-level translation is a considerable problem.

---

[4] We limited the maximum number of association words for one word to 400 in descending order of their association scores because of restriction of computational resources. In future work, we may alleviate the drawback of pMI by enlarging or deleting this limitation.

6 Conclusion

In this paper, we described a method for adding information for word translation disambiguation into the bilingual lexicon, by considering the associated words that co-occur with the input word. We based our method on the following two assumptions: "parallel word associations" and "one sense per word association." We aligned word associations by using a bilingual dictionary, and constructed a correlation matrix for each word in the source language for word translation disambiguation. Experiments showed that our method was applicable for both common newspaper text and domain-specific text and for two language pairs. The Jaccard/Dice coefficients were proven to be more effective than the other metrics as word association scores. Future work includes extending our method to handle verbs as input words by introducing syntactic co-occurrence. The comparisons with other disambiguation methods and machine translation systems would strengthen the effectiveness of our method. We consider also evaluations on real NLP tasks including machine translation.

Acknowledgments

References

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79-85.

Carpuat, Marine and Dekai Wu. 2007. Context-dependent phrasal translation lexicons for statistical machine translation. In *Proc. of Machine Translation Summit XI*, pages 73-80.

Church, Kenneth W., William Gale, Patrick Hanks and Donald Hindle. 1991. Using statistics in lexical analysis. *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, pages 115-164.

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

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61-74.

Ide, Nancy and Jean Veronis. 1998. Introduction to the special issue on word sense disambiguation: the state of the art. *Computational Linguistics*, 24(1):1-40.

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

Kruengkrai, Canasai, Kiyotaka Uchimoto, Jun'ichi Kazama, Yiou Wang, Kentaro Torisawa, and Hitoshi Isahara. 2009. An error-driven word-character hybrid model for joint Chinese word segmentation and POS tagging. In *Proc. of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 513-521.

Li, Cong and Hang Li. 2002. Word translation disambiguation using bilingual bootstrapping. In *Proc. of the 40th Annual Meeting of Association for Computational Linguistics*, pages 343-351.

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

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proc. of the 1st International Conference on New Methods in Natural Language Processing.*

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143-177.

Smadja, Frank, Kathleen R. McKeown and Vasileios Hatzivassiloglou. 1996. Translating collocations or bilingual lexicons: A statistical approach. *Computational Linguistics*, 22(1):3-38.

Vickrey, David, Luke Biewald, Marc Teyssier, and Daphne Koller. 2005. Word-sense disambiguation for machine translation. In *Proc. of the Conference on HLT/EMNLP*, pages 771-778.

Yarowsky, David. 1993. One sense per collocation. In *Proc. of ARPA Human Language Technology Workshop*, pages 266-271.