Statistical Technique for Estimating Word Correspondence for Bilingual Dictionary Development

Nuttaya Somboonphol[1] Virach Sornlertlamvanich[2]

[1] Department of Computer Engineering, Faculty of Engineering
King Mongkut's University of Technology Thonburi
[2] National Electronics and Computer Technology Center
112 Paholyothin Rd., Klong 1 , Klong Luang , Pathumthani 12120 Thailand
Tel 0-2564-6900 Fax 0-2564-6901-5
e-mail: s_nuttaya@yahoo.com, virach@nectec.or.th

Abstract

In this paper we propose a method for estimating word correspondences from the bilingual texts by using of linguistic information and statistical techniques, aiming at automatically extracting a bilingual dictionary from the parallel texts. We introduce a statistical technique for estimating word correspondences using the estimation functions proposed by Gale. The Gale's method have been studied and enhanced by our new criteria to improve the coverage and precision in word-pairs extraction. Extracted word correspondence will improve the accuracy of alignment by combining in the bilingual dictionary.

1. Introduction

A common use of aligned texts is the mostly used to create lexical resources such as bilingual dictionaries and parallel grammars. This is usually done in two steps. First the text alignment is extended to word alignment. Then some criterion such as frequency is used to select aligned pairs for which there is enough evidence to include them in the bilingual dictionary.

Parallel texts or bilingual texts are useful resources for acquiring a variety of linguistic knowledge, especially for Machine Translation systems which inherently require customizations. Bilingual dictionaries are, needless to say, the most basic and powerful knowledge source for improving and customizing translation systems.

Statistics-based processing has proven to be very powerful for aligning sentences and words in parallel corpora (Brown, 1991; Gale, 1993; Chen, 1993). Kupiec proposes an algorithm for finding noun phases in bilingual corpora (Kupiec, 1993). In this algorithm, noun-phases candidates are extracted from tagged and aligned parallel texts using a noun phase recognizer and calculated based on EM algorithm. Yamamoto (1993) proposes a method for generating a translation dictionary from Japanese/English parallel texts. In this method, English and Japanese compound noun phases are extracted from parallel texts and searched by matching their possible translations generated by the existing translation dictionary. Utsuro, Matsumoto and Nagao (1994) propose a unified framework for bilingual text matching by combining existing hand-written bilingual dictionaries and statistical. The used word correspondence information, which is available in hand-written bilingual dictionaries but estimated with statistical base-on Gale and Church (1991) and Kay and Röscheisen (1993) method.

2. Our approach and framework

In this paper, we utilize both linguistic and statistical information to estimate word correspondences which are not included in the seed English-Thai bilingual dictionary. Our goal is to develop a robust method enables highly accurate extraction of translation pairs or corresponding words from a relatively small amount of bilingual texts. We got a number of

new word correspondences by applying the Gale's method that used for estimating words which are not included in dictionary (Utsuro et al. (1994)).

The overall framework of our method is depicted in figure 1. In the first step, an English-Thai dictionary (*Lexitron: a corpus-based Thai-English dictionary developed by NECTEC*) of about 30,000 entries is consulted, to list up the unregistered words. English sentences are tagged by *Brill's POS tagger* (Brill, 1992) and Thai sentences are tagged by *SWATH*, a Thai POS tagger (Charoenpornsawat, 1998). *SWATH* use POS trigram model for word segmentation and POS tagging. We collected only content words according to the POS tagged information. From the list of word candidates, we estimated word correspondences by a statistical approach and extracted them under a threshold. In the course of estimation, we adopted a simple co-occurrence-frequency-based techniques in Gale and Church (1991).

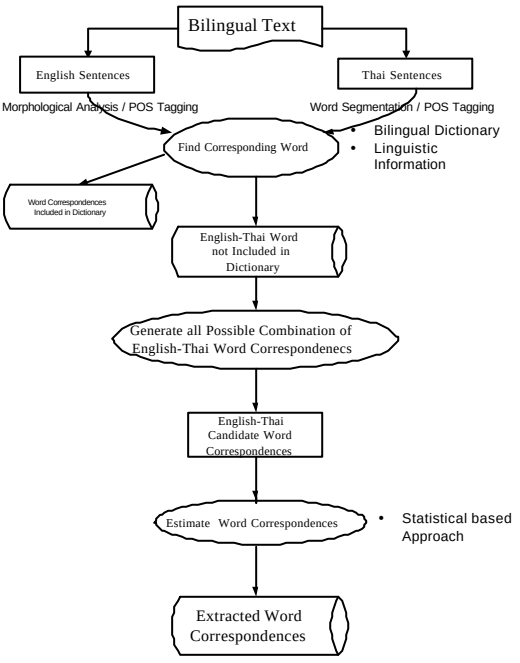

Figure 1. The Framework of Estimating Word Correspondence.

3. ESTIMATING WORD CORRESPONDENCES

We decided to adopt techniques as simple as possible, statistical techniques for estimating word correspondences are the estimation functions of Gale's applying to the aligned bilingual text.

Let w_s and w_t be words in the texts S and T, we define the following frequencies:

$freq(w_s, w_t)$ = (frequency of w_s and w_t's co- occurring in sentence bead)
$freq(w_s)$ = (frequency of w_s)
$freq(w_t)$ = (frequency of w_t)
N = (total number of sentence beads)

Then, estimation functions of Gale's is given as below.

Gale's Method
Let a ~ d be as follows:

$$a = freq(w_s, w_t)$$

$$b = freq(w_s) - freq(w_s, w_t)$$

$$c = freq(w_t) - freq(w_s, w_t)$$

$$d = N - a - b - c$$

The validity of word correspondence w_s and w_t is estimated by the following value:
$h_g(w_s, w_t)$

$$= \frac{(ad-bc)^2}{(a+b)\,(a+c)\,(b+d)\,(c+d)}$$

$$= \frac{(ad-bc)^2}{freq(w_s)\,freq(,w_t)(N-freq(w_s))(N-freq(,w_t))}$$

Let w_s be a word in the text S and w_t, $w\not{c}_t$ be words in the text T. Suppose that the word correspondence of w_s and w_t exits in the bilingual dictionary, while the correspondence word of w_s and $w\not{c}_t$ does not. Then the problem is to estimate the validity of word correspondence of w_s and $w\not{c}_t$.

Therefore, the terms in Gale's method are defined as follow:

Estimation

$freq\not{c}(w_s, w\not{c}_t) = freq(w_s, w\not{c}_t) - \sum freq(w_s, w_t, w\not{c}_t)$
$freq\not{c}(w_s) = freq(w_s)$
$freq\not{c}(w_t) = freq(w_t)$
$N\not{c} = N$
$(freq\not{c}(w_s, w_t) = freq(w_s, w_t))$

In case that w_s, w_t and $w¢_t$ occur in the same sentence bead, the co-occurrence of w_s and w_t is preferred while that of w_s and $w¢_t$ is ignored. Thus, $freq¢(w_s, w¢_t)$ is obtained by subtracting the frequency of all those cases from the real co-occurrence frequency of w_s and $w¢_t$. In addition, $freq¢(w_s)$ and $freq¢(w_t)$ are the same as the real frequencies ($freq(w_s)$ and $freq(w_t)$) and the estimated word correspondences reflect the real co-occurrence frequencies in the input text.

Threshold Function

Let x = co-occurrence frequency
y = estimated value for word correspondence
a = constant for eliminating low frequency words (1.0 for h_g)
b = constant for eliminating words with low estimated value (0.1 for h_g)
c = lower bound of word frequency (2.5 for h_g)

Then, the threshold function $g(x,y)$ is defined as follow:

$$g(x,y) = \frac{x\,(y - b)}{a}, \quad (x>c)$$

The condition for extracting the corresponding word pairs is given as follow:

$$g(x,y) > 1, \quad x>c$$

4. RESULTS

We used 2,000 pairs of sample sentences randomly selected from English-Thai dictionary, *So Sethaputra*, as a parallel texts. We assume that alignment at the sentence level is already done because each English-Thai sentence is the translation of each other. After filtering the registered word pairs, about 12,000 possible word pairs are generated as the candidates for finding the word correspondence. Followings are sample of word correspondences together with their statistic values.

The threshold value of $g(x,y)> 1$ is determined because of the appropriate result of 3 errors from 36 word correspondences extracted. Comparing with the threshold value of $g(x,y) > 0$, the results becomes as worse as 13 errors for 78 word correspondences extracted.

Table 1. Samples of the Result of Word Correspondence Computation.

W_{st}	f_{st}	w_s	f_s	W_t	f_t	h_g	$g(x,y)$
has>ÁÕ	49	Has	195	ÁÕ	235	2.05	95.49
has>¤Ô***	5	Has	195	¤Ô	6	5.54	27.22
war>Ê§¤ÀÒÁ	16	War	19	Ê§¤ÀÒÁ	18	0.75	10.34
Disease>âÅ¤	8	Disease	8	âÅ¤	9	0.89	6.31
Japanese> - ÕØ	6	Japanese	6	- ÕØ	6	1.00	5.40
Please >â»Å´	3	Please	33	â»Å´	33	1.87	5.31
Idea >¤Ô	6	Idea	8	¤Ô	6	0.75	3.90

The results are improved after adding a new condition for the extracting corresponding word pairs. Following is the new added condition.

$$\frac{|\,ft\text{-}fs|}{fst} \leq 5$$

From the observation, the incorrect word pairs have a high difference of word frequency in English and Thai (shown in column $|f_t - f_s|$, Table 2). Our new condition comes from the ratio of the difference between word frequency in English and Thai to their co-occurrence frequency ($|f_t - f_s| / f_{st}$). In other words, the word pair is likely not correspondent to each other if the value of the ratio is higher than the threshold.

Table 3 shows the comparison of word correspondence estimated by Gale' method and the one estimated by Gale's method with new condition. The first row shows the result and their accuracy at the threshold $g(x,y)$ higher than 1 and the second row shows the result and their accuracy at the threshold $g(x,y)$ higher than 0.

Figure 2 shows graph of varying thresholds with the precision and recall. A high threshold results in relatively higher precision and relatively lower recall. The black lines are the results of experiment with Gale's method. The dash lines are the results of experiment with Gale's method with new condition. The intersection point between the precision and the recall graphs of this figure shows that the Gale's method with new condition provides higher precision than the traditional Gale's method.

Table 2. The Incorrect Word Pairs after Adding New Condition
(The incorrect word pairs shown in the table with ∗)

W_{st}	f_{st}	W_s	f_s	W_t	f_t	H_g	$g(x,y)$	$\lvert f_t - f_s \rvert$	$\lvert f_t - f_s \rvert / f_{st}$
has>¤Õ	5	has	195	¤Õ	6	5.54	27.22	189	37.80
many>ÁÕ	9	many	11	ÁÕ	235	0.35	2.24	224	24.89
cholera>âÅ¤ ***	4	cholera	4	âÅ¤	9	0.44	1.37	5	1.25
case>à»Š	7	case	24	à»Š	318	0.20	0.69	294	42.00
action>à»Š	3	action	4	à»Š	318	0.29	0.57	314	104.67
week>ÇÑ ***	5	week	8	ÇÑ	17	0.18	0.41	9	1.80
is>»ÃÐËÅÒ´	6	is	410	»ÃÐËÅÒ´	10	0.16	0.34	400	66.67
are>ÁÕ	16	are	102	ÁÕ	235	0.11	0.16	133	8.31
was>¤¹	6	was	192	¤¹	124	0.13	0.15	68	11.33
was>à¢ŒE	3	was	192	à¢ŒE	90	0.13	0.09	102	34.00
is>·Ò	18	is	410	·Ò	117	0.10	0.08	293	16.28
war>ÂÊÕ***	4	war	19	ÂÊÕ	7	0.12	0.07	12	3.00
carry>·Ò	5	carry	30	·Ò	117	0.10	0.001	87	17.40

Table 3: The Accuracy of Word Correspondence Estimates

Threshold $g(x,y)$	Gale's Method					Gale's Method + New Condition				
	Number of Words			Precision (%)	Recall (%)	Number of Words			Precision (%)	Recall (%)
	Total	Correct	Wrong			Total	Correct	Wrong		
$g(x,y)>1$	36	33	3	91.66	17.64	34	33	1	97.05	17.46
$g(x,y)>0$	78	65	13	83.33	34.4	67	64	3	95.52	33.86

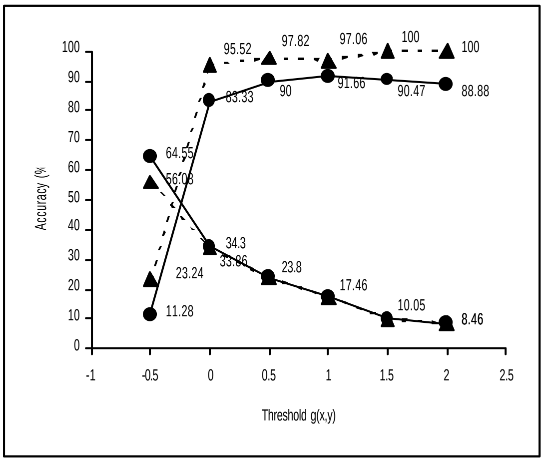

Figure 2. Graph of Varying Thresholds with the Precision and the Recall.

5. EVALUATION

To evaluate this method, we have estimated English translation of Thai sentences for finding word correspondence. After applying our method with 2,000 English-Thai sentences, we found our problem is the co-occurrence of word in the sentence. If two words in many sentences are likely to occur together and both of words do not appear in our dictionary, the two words tend to be aligned incorrectly when applying Gale's method.

From Figure 3, the word "many" often occur with " has", "there are" (meaning in Thai "ÁÕ") in the same sentence. And "many", " has", "there are" are not found in our bilingual dictionary. Therefore, "many" is always matched with "ÁÕin many sentences.

Other problems are caused by complex sentences that include and embedded sentence. We intend not to use our method to find word correspondence in complex or compound sentences. For better result, our bilingual texts should be simple sentence, so as to make the corresponding accuracy and frequency in word higher. Also bilingual text corpus should be larger.

6. CONCLUSION

This research proposes a method for developing bilingual dictionary by estimating word correspondences not included in bilingual dictionary, we adopt also quite simple method based-on the number of word correspondence by co-occurrence-frequency-based techniques of Gale's. The accuracy when applying Gale's method is 91.66% (for threshold more than 1) and 83.33% (for threshold more than 0). After adding new condition, the accuracy is improved to be 97.05% (for threshold more than 1) and 95.52% (for threshold more than 0). In the future the extracted word correspondences will improve the accuracy of alignment by combining in the bilingual dictionary. And also can be applied in bilingual concordances, for automatically constructing bilingual lexicons.

1. In writing there are many pitfalls.
   ¡ÒÃà¢ÕÂ¹Ë¹Ñ§Ê×ÍÁÕ¢éÍ¼Ô´¾ÅÒ´ÁÒ¡ÁÒÂ
2. This problem has many phases.
   »˜ÞËÒ¹ÕéÁÕËÅÒÂ´éÒ¹
3. There were many illegitimate children during the war years.
   ã¹ªèÇ§Ê§¤ÃÒÁÁÕà´ç¡ÁÒ¡ÁÒÂ·Õèà¡Ô´ÁÒ¹Í¡ÊÁÃÊ
4. The book was followed by many successive editions.
   Ë¹Ñ§Ê×ÍàÅèÁ¹ÕéÁÕ¡ÒÃ¾ÔÁ¾ìµèÍ¡Ñ¹ËÅÒÂ¤ÃÑé§
5. The military successes of the Japanese have been many.
   ªÑÂª¹Ð·Ò§·ËÒÃ¢Í§ÕèÍØè¹ÁÕÁÒ¡
6. His speech is interposed with many quotations.
   ¤Ó»ÃÒÈÃÑÂ¢Í§à¢ÒÁÕ¤Ó¾Ù´á·Ã¡ËÅÒÂ¢éÍ
7. How many cases are there under the doctor care.
   á¾·ÂìÁÕ¼Ùé»èÇÂã¹¤ÇÒÁ´ÙáÅ¡Õè¤¹
8. Since its birth the earth has gone through many cataclysmic changes.
   µÑé§áµè¡ÓàÔ´¢Öé¹â Å¡ä´é¼èÒ¹¡ÒÃà»ÅÕèÂ¹á»Å§ÁÒËÅÒÂ¤ÃÑé§
9. This musical comedy has many catchy tunes.
   ÅÐ¤ÃÃéÍ§µÅ¡àÃ×èÍ§¹ÕéÁÕà¾Å§à¾ÃÒÐËÅÒÂà¾Å§

Figure 3. Example of Co-Occurrence Words.

References

Brown, P.F., Lai, J.C. and Mercer, R.L., 1991. Aligning Sentences in Parallel Corpora, *Proceedings of 29th Annual Meetings of the Association for Computational Linguistics*, pp. 169-176.

Gale, W.A. and Church, K.W., 1993. A Program for Aligning Sentences in Bilingual Corpora, *Proceedings of Computational Linguistics*, Vol. 19, No. 1, pp. 75-90.

Chen, S.F., 1993. Aligning Sentences in Bilingual Corpora using Lexical Information, *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pp. 9-16.

Kepiec, J.1993. An algorithm for finding noun phrase correspondences in bilingual corpora. *In Proc. Of the 31st Annual Meeting of the ACL*, 17-22.

Yamamoto, Y. and Sakamoto, M., 1993. Extraction of Technical Terms Bilingual Dictionary from Bilingual Corpus, *IPSJ SIGNotes Natural Language 094-012 (in Japanese)*.

Utsuro T., Ikeda, H., Yamane, M., Matsumoto, Y. and Nagano, M.., 1994. Bilingual Text Matching using Bilingual Dictionary and Statistics, *Proceedings of 15th Computational Linguistics*, pp. 1076-1082.

Gale, W. and Church, K., 1991. Identifying Word Correspondence in Parallel Text, *Proceedings of the DARPA NLP Workshop*.

Kay, M. and Röscheisen, M. 1993. Text translation alignment, *Computational Linguistics*, 19(1): 121-142.

Brill, E., 1992. A Simple Rule-Based Part of Speech Tagger, *Proceedings of Applied Natural Language Processing '92*, Vol. 24, No. 3, pp. 173-202.

Charoenpornsawat, P., 1998. Featured Based Thai Word Segmentation, *Master of Engineering Thesis*, Computer Engineering Program, Chulalongkorn, pp. 34-36.