Using Part-of-speech Information in Word Alignment

Jyun-Sheng Chang and Huey-Chyun Chen
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan 30043
jschang@cs.nthu.edu.tw

Abstract

We have developed a new program called *PosAlign* for identifying word correspondence in parallel text by using a translation model based on part-of-speech. The program takes the result from the part-of-speech taggers, and applies the POS-level constraints to produce word alignments. The parallel text of 55 suits of representative English sentence patterns (454 English sentences and their Chinese translations), were used to test the feasibility of this approach. By aligning cognates first before applying Brown et al.'s Model 2, the program converges faster and more correctly. Advantages of using a POS-based translation model include: (1) only parallel text of modest size is required for training. (2) Smaller number of model parameters are used and less storage is required. (3) A high degree generality is obtained, and the model is likely to be more applicable to text of other domains.

1. Introduction

The statistical machine translation model proposed by Brown et al. (1990) provides researchers with a new way of attacking the MT problem and has established itself as a new paradigm in MT research. Basically, there are two components in the statistical MT model, a prior and post prior probabilities. The a prior probabilities measure the possibility of a translation without considering the source given. These probabilities are often called the language model and the n-gram model seems to be the model of choice. The post prior probabilities measure the possibility of a translation under the condition that the source is given. This view of post prior probability of machine translation was first proposed and called translation model in Brown (1990). Later Brown et al. (1993) proposed 5 translation models with different levels of approximation.

In order to estimate the parameters in the translation model, research has been conducted to find alignment between sentences (Brown et al. 1992, Gale and Church 1992, Chen 993), syntactic structures (Mutsumoto 19930), noun phrases (Kupiec 1993), collocations (Smadja 1992), and words (Gale and Church 1990, 1993) in parallel text. Sentences from the two languages can be aligned with error rate as low as 0.6% using simple dynamic programming technique. The reason for the low error rates is three-folded. The alignments are predominantly 1-to-l. The lengths of counterpart sentences correlate closely. And finally, in most cases there are no cross dependencies between pairs of aligned sentences.

Words, on the other hand, are much more difficult to align reliably, because they are less predominantly 1-to-l and the non-crossing constrain does not hold. Brown et al. proposed an approximation implementation of the EM algorithm to estimate the word-to-word translation model and to find the most likely word alignments for the aligned sentence pairs (Brown et al. 1989). Even with the limitation of small dictionaries (1,700 most frequently used French words with translation sentences covered completely by a 1,000-word English vocabulary) and availability of 117,000 sentences, their method does not seem to produce robust estimation of word-based translation model (Gale and Church

1990, p. 154). Another EM-based algorithm *Word_align* (Ido, Church and Gale 1993) with character alignment as the starting point, was shown to align 60.5% percent of the words correctly, and in 84% of the cases the offset from the correct alignment is at most 3. Gale and Church (1990) proposed using an χ^2-like associate measure for the plausibility of one-to-one word alignment, instead of the conditional probability derived from EM algorithm. The association ratios are consequently used in a dynamic programming algorithm to align words partially (only words with association higher than a threshold are considered for alignment). A certain threshold allows the algorithm to align 60% of the words and 95% of the alignments are correct.

These lower success rates for word alignment are due to the predicament of very large parameter space of word-based translation models. Apparently, difficulties in training word-based translation model have constituted one of the major obstacles of using the statistical machine translation. We are interested in statistical MT model that can be trained sufficiently with modest amount of data. In this paper, we propose to work with part-of-speech-based translation models.

In the following sections, we will describe an efficient algorithm for word alignment that use part-of-speech information. The algorithm first lookups a small dictionary of cognates to obtain initial alignment for the sentences. The partially aligned sentences are used to in an EM algorithm to train a statistical machine translation model on the level of part-of-speech. A small parallel text is used to test the feasibility of the approach. The model converges rapidly for the test set with reasonably good performance.

2. The Word Alignment Algorithm

We use an example to introduce our framework for alignment Consider the parallel text *(C, E)* in Fig. 1. Both the Chinese and English sentences are tagged with part-of-speech information.

C:					E:			
wuo	zhen	xiang	ku	.	I	feel	like	crying
I	real	think	cry	.	PRON	VB	PREP	VBG
PRON	ADV	VERB	VERB	.	PRON	VERB	PREP	VERB
(I really want to cry)								

Figure 1. Parallel Text

The correct word alignments for the this pair of sentences are shown in Fig. 2. As with sentence alignments, word alignments are not necessarily 1-to-l. In general, in a so-called *i-to-j alignment, i* words in one sentence might correspond to *j* words in its translation sentence. For example, the correct alignments of the sentences in Fig. 1, consists of the following word alignments:

1-to-l alignment: (I, wuo)
0-to-l alignment: (-, zhen)
2-to-l alignment: (feel like, xiang)
1-to-l alignment: (crying, ku)

The POS information, position of words, and word alignments are shown in Fig. 2. For more example sentence of the parallel text, see Fig. 3.

English	POS	position	Chinese		POS	position
I	PRON	El	wuo	(I)	PRON	C1
			zhen	(really)	ADV	C2
feel like	VERB PREP	E2 E3	xiang	(think)	VERB	C3
crying	VERB	E4	ku	(cry)	VERB	C4
.	.	E5	.		.	C5

Figure 2. Word alignments

1.e Birds/NNP fly/VB./.

1.c Niao(bird)/Nfei(fly)/V./.

2.e We/PRP all/DT breathe,/,VB,/, eat,/,CD and/CC drink/NN./.

2.c Wuomen(we)/PRON dou(all)/ADV yiao(need)/V fuxi(breathe)/V ,/, chi(eat)/V ,/, he(drink)/V./.

3.e There/EX was/VBD no/DT money/NN./.

3.c Meiyou(not-have)/V qian(money)/N./.

4.e It/PRP seems/VBZ that/IN John/NNP is/VBZ an/DT intelligent/JJ boy/NN./.

4.c Yuehan(John)/PRON shihu(seem)/ADV shi(be)/V ge(piece)/M congmi(clever)/ADJ de(CLITCS)/S xiaohai(kid)/N./.

5.e The/DT sun/NN rises/VBZ in/IN the/DT east/NN./.

5.c Taiyang(sun)/N you(from)/PREP dongfang(east)/N shen(rise)/V qi(up)/PART./.

Figure 3. Tagged parallel text

If a word-based translation model is used, the optimal word alignments are obtained using the probability $t(c \mid e)$ of the Chinese word c being the translation of the English word e where $c \in$ *(-, wuo, zhen, xiang, ku }* and $e \in \{ -, I,$ *feel, like, crying }*. The problem is that there are so many of these t parameters and they are extremely difficult to estimate. So, we propose the alternative of using a POS-to-POS translation model. Under the POS-based translation model, we calculate optimal alignments for *(E, C)* using the probability $t(c \mid e)$ of the Chinese word with POS c being the translation of the English word with POS e where $c \in$ *(-, PRON, ADV, VERB }* and $e \in$ *(-, PRON, VERB, PREP }*. Obviously, there are far less part-of-speeches than there are words.

We can estimate $t(c \mid e)$ for POS-based model much the same way as for word-based model However, there are two common problems associated with statistical estimation of parameters. The first problem has to do with fertility involving deletion. It is considerably more difficult to estimate parameters involving deletion in sentence alignment (Simard 1990, Brown 1990). The second problem is related to errors introduced due to distribution regularity. Because of the very nature of reliance on distribution regularity, statistical methods such as EM algorithms and inside-outside reestimation algorithms tend to introduce errors due to distribution regularity. For example, under statistically learned stochastic context free grammar, the period tends to be grouped with the previous word as a constituent (Periera and Schabes 1992). Aligning words regardless of their being function or content words creates similar problems. Firstly,

the translation probability of any POS with "-" (empty) often unrealistically high. Similarly, the translation probability of a content POS with a pronoun is also unrealistically high due to the abundance of pronouns in the parallel text

We have found that through initial alignment of cognates, we can deal with the two problems at the same time very effectively. Since the bulk of the cognate dictionary is function words which tends to appear more often than normal (distribution regularity) and involve non 1-to-l translation (deletion on one language side or fertile l-to-2, 2-to-l and 2-to-2 alignments). By aligning these pairs first, we can improve the accuracy of parameter estimation for the rest of the data. Consequently, the content words will be less likely to align incorrectly with function words. The idea of initial alignment is similar to using anchors (Brown 1990) and cognates (Simard, Foster and Isabelle 1992) in sentence alignment to deal with the issue of recovering from deletion of sentences. It is also related to using a partially bracketed corpus in learning a stochastic context-free grammar (Periera and Schabes 1992).

The cognate dictionary contains pairs of words in the two languages which are very likely to be used as mutual translation. Cognates include such words as pronouns, proper nouns, numerical expressions, and punctuation marks. For the example in Fig. 1, we obtain the following initial alignments shown, in Fig. 4.

English	POS	position	Chinese		POS	position
I	PRON	E1	wuo	(I)	PRON	C1
		E5	.		.	C5

Initial alignment of cognates

E:					C:				
El	E2	E3	E4	E5	C1	C2	C3	C4	C5
I	feel	like	crying	.	wuo	zhen	xiang	ku	
PRON	VB	PREP	VBG	.	(I)	(real)	(think)	(cry)	
PRON	VERB	PREP	VERB	.	PRON	ADV	VERB	VERB	
C1				C5	E1				E5

Figure 4. Cognates and partially aligned sentences

The sequence of POS's {(-, VERB, PREP, VERB); (-, ADV, VERB, VERB)} is then used in training Brown et al.'s Model 2. The *PosAlign* algorithm is summarized as follows:

Algorithm *PosAlign*

1. Tag the parallel text with part-of-speeches
2. Initial alignment with the help of cognate lookup
3. Train the translation model iteratively using the unaligned part of the POS sequences

3. Implementation and Evaluation

PosAlign was first evaluated on a representative set of 55 English sentence patterns with 454 sentences and their translation in Chinese. Two taggers for English and Chinese are used in the experiment.

The entries of cognate dictionary came *from* the function words of an electronic Chinese-English dictionary (BDC 1990) and a translator handbook on function words (Wong 1989), totaling some 3000 entries with various kinds of fertility (1-to-l, l-to-0,0-to-l, 1-to-many, many-to-1, and many-to-many).

The sentences in both languages run through taggers for the two languages first. The tagged text is used unedited. See Fig. 3 for some examples of tagged sentences. To make the training of translation model more effective, we have combined tags. For example, various inflection forms of English verbs, *VB, VBD, VBG,* etc., are combined into a single tag *VERB*. Similar combination was also done for the Chinese tagset. Pairs of counterpart sentences were then partially aligned with the help of cognate lookup. Probabilities for both POS mapping and distortion in positions of mapped word are derived.

Table 1 presents the trained translation model. Only POS-to-POS mapping with highest probabilities are shown.

English POS		Chinese PQS		Prob.
2	infinite to	VERB	verb	0.162
2	infinite to	-	empty	0.119
D	cardinal number	Q	numeral	0.371
D	cardinal number	M	measure noun	0.120
E	existential there	VERB	verb	0.162
E	existential there	-	empty	0.129
PREP	preposition	PREP	preposition	0.358
PREP	preposition	-	empty	0.135
PREP	preposition	ADV	adverb	0.102
PREP	preposition	VERB	verb	0.088
PREP	preposition	N	noun	0.066
ADJ	adjective	ADJ	adjective	0.206
ADJ	adjective	VERB	verb	0.145
ADJ	adjective	-	empty	0.128
ADJ	adjective	ADV	adverb	0.100
M	modal auxiliary	X	modal auxiliary	0.177
M	modal auxiliary	VERB	verb	0.165
M	modal auxiliary	-	empty	0.109
N	noun	N	noun	0.218
N	noun	-	empty	0.128
ADV	adverb	ADV	adverb	0.203
ADV	adverb	-	empty	0.116
T	determiner/pronoun	-	empty	0.121
T	determiner/pronoun	D	determiner	0.057
VERB	verb	VERB	verb	0.244
VERB	verb	-	empty	0.134
VERB	verb	ADV	adverb	0.097

Table 1. POS Translation Model

Using the example in Fig. 1, we show how alignment is done using the trained model. As is being done in the training phrase, the cognates are aligned first, resulting in the following alignment being identified:

English	POS	position	Chinese		POS	position
I	PRON	E1	wuo	(I)	PRON	C1
		E5	.		.	C5

The sentences with initial alignments taken away become the following:

E: 1.- 2. feel/VERB 3.1ike/PREP 4. crying/VERB 5.-

C: 1.- 2. zhen (real)/ADV 3. xiang (think)/VERB 4. ku (cry)/VERB 5.-

From the probabilities listed in Table 1 for *t* (ADV | VERB), t (VERB | VERB), *t* (-| VERB), *t* (ADV | PREP), *t* (VERB | PREP) , *t* (- | PREP),

t (ADV | VERB)	=0.097	*t* (ADV | PREP)	=0.102
t (VERB I VERB)	=0.244	*t* (VERB | PREP)	=0.088
t (- | VERB)	=0.134	*t* (- | PREP)	=0.135

it is not difficult to see that the optimal solution for (E, C) under the trained model is the following:

English	POS	position	Chinese		POS	position
I	PRON	E1	wuo	(I)	PRON	C1
			zhen	(really)	ADV	C2
feel	VERB	E2	xiang	(think)	VERB	C3
like	PREP	E3				
crying	VERB	E4	ku	(cry)	VERB	C4
		E5	.		.	C5

Table 3. The alignment result

Notice that under the assumption of Model 2, we are not able to produce a 2-to-l alignment (feel like, xiang) as shown in Fig. 2.

4. Concluding remarks

Compared with other word alignment algorithms, *PosAlign* does not require large amount of data for training, and was shown to produce alignment with high precision in complete alignment (taking 0-1 and 1-0 mapping into consideration). The program is much more general in the sense that it has the potential to performs as well on unrestricted text in a different domain.

PosAlign provide an example of the way how statistical machine translation can be made more general and modular by incorporating other language processing modules such as part-of-speech taggers. We are currently expanded our cognate dictionary for better results. A second evaluation of PosAlign is

being carried out for some 70 short bilingual news reportage. Initial results from the experiments are quite encouraging.

References:

1. Behavior Design Co. (1990) "BDC Chinese-English Electronic Dictionary," Hsinchu, Taiwan.
2. Brown, P., et al., (1990) "A Statistical Approach to Machine Translation," Computational Linguistics, vol. 16, pp. 79-85.
3. Brown, P., et. al. (1991), "Word Sense Disambiguation using Statistical Methods," Proceedings of the 29th Annual ACL Meeting, pp. 264-270.
4. Brown, P., et al. (1992), "Analysis, Statistical Transfer, and Synthesis in Machine Translation" Proceedings of the 4th International Conference on Theorectical and Methodological Issues in Machine Translation, pp. 83-100.
5. Brown, P., S.A. Pietra, V.J.D. Pietra, and R. Mercer (1992), "Dividing and Conquering Long Sentences in a Translation System," Proceedings of the Speech and Natural Language Workshop, pp. 267-271.
6. Brown et al. (1991) "Aligning Sentences in Parallel Text," Proceedings of the Annual ACL meeting pp. 169-176.
7. Brown P., et al. (1993) "The Mathematics of Machine Translation: Parameter Estimation," Computational Linguistics, 19(2), pp. -.
8. Chen, S. (1993) "Aligning Sentences in Bilingual Corpora Using Lexical Information," Proceedings of the Annual ACL meeting pp. 9-17.
9. Church, K.W. (1988) "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text," Proceedings of The Second Applied Natural Language Processing, ACL, pp. 136-143.
10. Dagan, I., K.W. Church, and W.A. Gale. (1993) "Robust Bilingual Word Alignment for Machine Aided Translation," Proceedings of the Annual ACL meeting pp. 1-8.
11. Gale, W. and K. Church. (1991) "Identifying Word Correspondences in Parallel Texts," Proceedings of Speech and Natural Language Workshop, pp. 152-158.
12. Gale, W., K. Church, and D. Yarowsky (1992), "Using Bilingual Materials to Develop Word Sense Disambiguation Methods, "Proceedings of the 4th International Conference on Theoretic and Methodological Issues in Machine Translation, pp. 101-112.
13. Ker, SJ. and J.S. Chang. (1994) "Word Correspondence and Sense Tagging of Parallel Text," Technical Report NSC82-0408-E-007-195, National Science Council, Taiwan.
14. Krovetz, R. and W. Croft. (1992) "Lexical Ambiguity and Information Retrieval," ACM Transaction on Information Systems, pp. 115-141.
15. Kupiec. (1993) "An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora" Proceedings of the Annual ACL meeting, pp. 17-22.
16. Matsumoto, Y. et al. (1993) "Structural Matching of Parallel Texts" Proceedings of the Annual ACL meeting, pp. 23-30.

17. Pereira, F. and Y.Schabes. (1992) "Inside-Outside Reestimation from Partially Bracked Corpora," Proceedings of the Annual ACL meeting, pp. 128-135.

18. Smadja. (1993) "How to Compile a Bilingual Collocational Lexicon Automatically. Proceedings of the AAAI Workshop on Statistically-based Natural Language Programming Techniques, pp. 57-63.

19. Simard, M., G. Foster and P. Isabelle. (1993) "Using Cognates to Align Sentences in Parallel Corpora, Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation, Montreal, Canada

20. Wong, I. (1990) "A Translation Handbook of Chinese-English Function Words", Joint Publishing Co., Hong Kong.