Using Punctuations and Lengths for Bilingual Sub-sentential Alignment

Wen-Chi Hsien, Kevin Yeh, Jason S. Chang
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road, Hsinchu, 300, Taiwan, ROC
{g904307, jschang)@cs.nthu.edu.tw

Thomas C. Chuang
Department of Computer Science
Van Nung Institute of Technology
1 Van-Nung Road, Chung-Li, Taiwan, ROC
tomchuang@cc.vit.edu.tw

Abstract

We present a new approach to aligning bilingual English and Chinese text at sub-sentential level by interleaving alphabetic texts and punctuations matches. With sub-sentential alignment, we expect to improve the effectiveness of alignment at word, chunk and phrase levels and provide finer grained and more reusable translation memory.

1. Introduction

Recently, there are renewed interests in using bilingual corpus for building systems for statistical machine translation (Brown et al. 1988, 1991), including data-driven machine translation (2002), computer-assisted revision of translation (Jutras 2000) and cross-language information retrieval (Kwok 2001). It is therefore useful for the bilingual corpus to be aligned at the sentence level and even sub-sentence level with very high precision (Moore 2002; Chuang, You and Chang 2002, Kueng and Su 2002). Especially, for further analyses such as phrase alignment, word alignment (Ker and Chang 1997; Melamed 2000) and translation memory, high precision and quality alignment at sentence or sub-sentential levels would be very useful. Furthermore, alignment at sub-sentential level has the potential of improving the effectiveness of alignment at word, chunk and phrase levels and providing finer grained and more reusable translation memory.

Much work has been reported in the literature of computational linguistics studying how to align sentences. One of the most effective approaches is length-based approach proposed by Brown et al. and by Gale and Church. Length-based approach for aligning parallel corpora has commonly been used and produces surprisingly good results for the language pair of French and English at success rates well over 96%. However, it does not perform as well for alignment of two distant languages such as Chinese-English. Furthermore, for sub-sentential alignment, length-based approach gets less effectiveness than running it in sentence level since sub-sentence has less information in length.

Punctuations based approach (Yeh, Chuang and Chang 2003) for sentence alignment produces high accuracy rates as same as length based approach and was independent of languages. Although the ways different languages around the world use punctuations vary, symbols such as commas and full stops are used in most languages to demarcate writing, while question and exclamation marks are used to show emphasis. However, for sub-sentential alignment, punctuation-based approach has the same problem as length-based approach — no enough information in sub-sentence since sub-sentence might be very short and just include one or two punctuations within it.

Yeh, Chuang and Chang (2003) examined the results of punctuation-based sentence alignment and observed:

> "Although word alignment links do cross one and other a lot, they general seem not to cross the links between punctuations. It appears that we can obtain sub-sentential alignment at clause and phrase levels from the alignment of punctuation."

Building on their work, we develop a new approach to sub-sentential alignment by interleaving the alignment of text and punctuations. In the following, we first give an example for bilingual sub-sentential alignment in Section 2. Then we introduce our probability model in Section 3. Next, we describe experimental setup and results in Section 4. We conclude in Section 5 with discussion and future work.

2. Example

Consider a pair of aligned sentences in a parallel corpus as below:

> "My goal is simply this - to safeguard Hong Kong's way of life. This way of life not only produces impressive material and cultural benefits; it also incorporates values that we all cherish. Our prosperity and stability underpin our way of life. But, equally, Hong Kong's way of life is the foundation on which we must build our future stability and prosperity."

> 我的目標很簡單，就是要保障香港的生活方式。這個生活方式，不單在物質和文化方面爲我們帶來了重大的利益，而且更融合了大家都珍惜的價值觀。香港的安定繁榮是我們生活方式的支柱。同樣地，我們未來的安定繁榮，亦必須以香港的生活方式爲基礎。

We can observe that although word alignment links might cross one and other a lot, there exist some text-blocks as follow that general seem not to cross the links between punctuations:

> "My goal is simply this –"

> "我的目標很簡單，"

> "to safeguard Hong Kong's way of life."

> "就是要保障香港的生活方式。"

> "This way of life not only produces impressive material and cultural benefits;"

> "這個生活方式，不單在物質和文化方面爲我們帶來了重大的利益，"

“it also incorporates values that we all cherish.”

“而且更融合了大家都珍惜的價值觀。”

…

That’s what we call sub-sentences here. From the examples above, we can define that a sub-sentence is a text-block that include at least one or more punctuations. That’s an unclear definition since a sentence and a paragraph also fit the definition too. However, what we want is to find out the shortest parallel text-block pairs that fit the definition. That’s why in the third pair of above examples, “這個生活方式，” is a Chinese text-block but we have to combine it with ”不單在物質和文化方面爲我們帶來了重大的利益，”, because we can’t find any English text-block correspond to “這個生活方式，”, we have to combine the two Chinese above first, than we can find the corresponding one : “This way of life not only produces impressive material and cultural benefits;”.

3. Probability Model

In this section we describe our probability model. To do so, we will first introduce some necessary notation. Let E be an English paragraph $e_1, e_2,\ldots,e_m$ and C be a Chinese paragraph $c_1, c_2,\ldots,c_n$, which e_i and c_j is a text-blocks as described in Section 2. We define a **link** $l(e_i, c_j)$ to exist if e_i and c_j are translation (or part of a translation) of one another. We define **null link** $l(e_i, c_0)$ to exist if e_i does not correspond to a translation of any c_j. The null link $l(e_0, c_j)$ is defined similarly. An **alignment** A for two paragraphs E and C is a set of links such that every text-block in E and C participates in at least one link, and a text-block linked to e_0 or c_0 participates in no other links.

We define the alignment problem as finding the alignment A that maximizes P(A|E, C). An alignment A consists of t links $\{l_1, l_2,\ldots, l_t\}$, where each $l_k = l(e_{ik}, c_{jk})$ for some i_k and j_k.We will refer to consecutive subsets of A as $l_i^j = \{l_i, l_{i+1},...,l_j\}$, Given this notation, P(A|E, C) can be decomposed as follows:

$$P(A \mid E,F) = P(l_1^t \mid E,F) = \prod_{k=1}^{t} P(l_k \mid E,C,l_1^{k-1})$$

For each condition probability, given any pair e_i and c_j, the link probabilities can be determined directly from combining the probability of length-based model with punctuation-based model. From the paper of Gale and Church in 1993 for length-based model, we know the match probability is $Prob(\delta \mid match)$ $Prob(match)$ and $Prob(\delta \mid match)$ can be estimated by

$$Prob(\delta \mid match) = 2\,(1 - Prob(|\delta|))$$

Where *Prob*(|δ|) is the probability that random variable, z, with a standardized (mean zero, variance one) normal distribution, has magnitude at least as large as |δ|. That is,

Where

$$\text{Prob}(\delta) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\delta} e^{-z^2/2} dz$$

We compute δ directly from the length of two portions of text, l_1 and l_2, and the two parameters, c and s^2. (Where c is the expected number of characters in L_2 per character in L_1, and s^2 is the variance of the number of characters in L_2 per character in L_1.) That is, $\delta = (l_2 - l_1 \times c) / \sqrt{l_1 s^2}$. Then, *Prob*(|δ|) is computed by integrating a standard normal distribution (with mean zero and variance 1).

Then, from Yeh, Chuang and Chang (2003), for punctuation-based model, we know:

$$P(e_i, c_j) = P(pe_i, pc_j)\, p(|pe_i|, |pc_j|) \times \prod_{i=1}^{m-1} P(pe_i, null)\, p(|pe_i|, 0) \times \prod_{j=1}^{n-1} P(null, pc_j)(0, |pc_j|)$$

where e_i and c_i is λ, one, or two punctuations,
e_i, c_j = English and Chinese text-block
$pe_1\, pe_2 \ldots pe_m$ = p*E*, the English punctuations,
$pc_1\, pc_2 \ldots pc_n$ = p*C*, the Chinese punctuations,
$|pe_i|$ and $|pc_i|$ are the number 0f punctuations in
pe_i and pc_i respectively,
$P(pc_i, pe_i)$ = probability of pc_i translates into pe_i,

Thus, for each link l_k given E, C and l, we can computing the probability as following:

$P(l_k | E, C, l^{k-1}) = P(\delta \mid match) P(match) * P(e_i, c_i)$, So

$$P(A \mid E, F) = \prod_{k=1}^{t} P(\delta \mid match) P(match) P(e_{ik}, c_{jk})$$

4. Experimental result

In order to assess the performance of our sub-sentential alignment model, we selected top ten bilingual articles from official record of proceedings of Hong Kong Legislative Council at Oct. 7, 1992 as our experimental data. For probability of punctuation, We use all the data such as punctuation translation probability(Table 1)and category frequency Prob(match) (Table 2) from Yeh, Chuang and Chang (2003) directly. For probability of length, we set c = 3.23 , standard variance = 0.93 and match probability as Table 3:

Table 1. Punctuation Translation probability

English Pun.	Chinese Pun.	Match Type	Counts	Probability

,	，	1-1	541	0.809874
,	、	1-1	56	0.083832
,	。	1-1	41	0.061377
,	「	1-1	10	0.01497
,	：	1-1	5	0.007485
,	；	1-1	4	0.005988

Table 2: P(match) Category Frequency Prob(match)

Match type	1-1	1-0, 0-1	1-2	2-1	1-3	1-4	1-5
Probability	0.65	0.000197	0.0526	0.178	0.066	0.0013	0.00013

Table 3. Match probability of sentences

Match Type	Probability
1-0	0.000197
0-1	0.000197
1-1	0.6513
2-2	0.0066
1-2	0.0526
2-1	0.1776
1-3	0.0066
3-1	0.0658
1-4	0.00132
4-1	0.0132

After aligning by our model, we got 94 parallel records from the ten articles, and had precision rate at 92.55%. To calculate precision rate, we count English and Chinese sub-sentences isolated, so were the error records. For detail, refer to Appendix. Following table show the result:

Article	# of sub-sentence	errors	Prec(%)
Official record of proceedings of Hong Kong Legislative Council	188	14	92.55

5. Discussion and future work

We propose a model combining length-based approach with punctuation-based approach to do sub-sentential alignment and we got about 93% precision rates here. It was not bad but still had a lot of space to improve. We should change the sub-sentence match type probability first of all. We use the probability of sentence match type instead of sub-sentence match type in this experiment since we don't do sub-sentence training first. It causes a problem, because a sub-sentence has higher probability to include two

or three text-blocks within it than a sentence do. An inverted sentence causes the second problem here, no matter length-based or punctuation-based approach you used; they cannot solve this kind of problem. We might add lexical information in it to solve this kind of problem in the future.

References

Brown, P. F., J. C. Lai and R. L. Mercer (1991), 'Aligning sentences in parallel corpora', in 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, USA. pp. 169-176.

Chen, Stanley F. (1993), Aligning Sentences in Bilingual Corpora Using Lexical Information. In Proceedings

Chuang, T., G.N. You, J.S. Chang (2002) Adaptive Bilingual Sentence Alignment, Lecture Notes in Artificial Intelligence 2499, 21-30.

Gale, William A. & Kenneth W. Church (1993), A program for aligning sentences in bilingual corpus. In Computational Linguistics, vol. 19, pp. 75-102.

Jutras, J-M 2000. An Automatic Reviser: The TransCheck System, In Proc. of Applied Natural Language Processing, 127-134.

Ker, Sue J. & Jason S. Chang (1997), A class-based approach to word alignment. In Computational Linguistics, 23:2, pp. 313-344.

Kueng, T.L. and Keh-Yih Su, 2002. A Robust Cross-Domain Bilingual Sentence Alignment Model, In Proceedings of the 19th International Conference on Computational Linguistics.

Kwok, KL. 2001. NTCIR-2 Chinese, Cross-Language Retrieval Experiments Using PIRCS. In Proceedings of the Second NTCIR Workshop Meeting, pp. (5) 14-20, National Institute of Informatics, Japan.

Melamed, I. Dan (1997), A portable algorithm for mapping bitext correspondence. In The 35th Conference of the Association for Computational Linguistics (ACL 1997), Madrid, Spain.

Piao, Scott Songlin 2000 Sentence and word alignment between Chinese and English. Ph.D. thesis, Lancaster University.

Kevin C. Yeh, Thomas C. Chuang, Jason S. Chang (2003), Using Punctuations for Bilingual Sentence Alignment - Preparing Parallel Corpus for Distribution by the ACLCLP

Appendix

Table A. all incorrect alignments of this experiment. Shaded parts indicate imprecision in alignment results. We calculated the precision rates by dividing the number of unshaded sentences (counting both English and Chinese sentences) by total number of sentences proposed. Since we did not exclude aligned pair using a threshold, the recall rate should be the same as the precision rate.

Sub-sentence alignment based on length and punctuation	
English text	Chinese Text
[Now] [is the time] [to show] [how we mean to prepare] [for Hong Kong's future] [under that far-sighted concept],	現在也是時候表明我們打算怎樣按照「一國兩制」這個極具遠見的構思，
["one country, two systems"].	爲香港的未來作好準備。
[- we shall maintain] [an economy] [which] [continues to thrive] [and prosper,]	— 我們便可令經濟持續繁榮蓬勃，創造所需財富，
[generating the wealth] [required to provide] [the standards of public service] [that people rightly demand;]	使提供的公共服務，能達到市民要求的合理水平；
[Our prescription for prosperity] [is straightforward.]	我們締造繁榮的配方清楚簡單。我們相信，
[We believe that] [businessmen] [not politicians or officials] [make the best commercial decisions.]	最佳的商業決定是由商人，而不是由政治家或政府官員作出的。
[We believe that] [government spending] [must follow] [not outpace] [economic growth.]	我們相信，政府開支必須跟隨經濟增長，
[We believe in competition] [within] [a sound,	而不應超逾經濟增長。我們更相信，
fair framework of regulation and law.]	應在健全而公平的法規下進行競爭。
[I am inviting distinguished members] [of the business community] [to join it.]	並會邀請商界傑出人士加入。他的職責是，
[Their mandate] [will be] [to advise me] [on:]	就下開事項向我提供意見：