Interleaving Text and Punctuations for Bilingual Sub-sentential Alignment

Wen-Chi Hsie, 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 (Richardson et al. 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 alignment at sub-sentential levels would be very useful. Alignment at sub-sentential level has the potential of improving the effectiveness of alignment at word and phrase levels and providing finer grained and more reusable translation memory.

Much work has been reported in the literature of computational linguistics on how to align sentences, while very little is touched on alignment just below the sentence level. The most effective approach for sentence alignment is the length-based approach proposed by Brown et al. (1991) and by Gale and Church (1991). Both methods use normal distribution to model the ratio of lengths between the counter-part sentences measured in number of words or characters. 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 text in two distant languages such as Chinese and English.

Yeh (2003) proposed a punctuation-based approach for sentence alignment which produces even high accuracy rates than the length based approach. It was pointed out that the ways different languages use punctuations are more or less similar and the correspondence of punctuations across different languages can be obtained using a small set of training data. By soft matching punctuations of the two languages in ordered comparison, the probabilities of mutual translation for a pair of bilingual sentences can be estimated more effectively than lengths. This is not surprising since the average sentence contains many punctuations which carry more information than lengths. Yeh also 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."

This observation indicates that in bilingual corpus pieces of text delimited by punctuations behave much the same way as sentences with non-crossing alignment links. Therefore, it is reasonable to align pieces of text ending with a couple of punctuations, much the same way as sentence alignment. Building on their work, we develop a new approach to sub-sentential alignment by interleaving the matches of alphabetic texts 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 counterpart paragraphs in the official records of Hong Kong Legislative Council:

> "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。(Source: Oct. 7, 1992, Governor Christopher Francis Patten's address to the HK LEGCO)

By sub-sentential alignment, we mean identifying the shortest possible pair of counterpart texts ending with punctuations. From the example above, the following is the intended results of sub-sentential alignment:

- 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。

Notice that longer pairs such as the following translation equivalent pair of sentences

My goal is simply this – to safeguard Hong Kong's way of life.
我的目標很簡單，就是要保障香港的生活方式。

does not fit the bill, since a finer grained subdivision into two 1-1 matches, (My goal is simply this –, 我的目標很簡單，) and (to safeguard Hong Kong's way of life., 就是要保障香港的生活方式。) also preserve translation equivalence. Not unlike the situation in sentence alignment, there are many to one, one to many, and many to many matches. For instance, it is not possible to find a 1-1 match for "This way of life not only produces impressive material and cultural benefits;" since it only corresponds to "這個生活方式，" in part. Therefore, we have to combine the subsequent clause "不單在物質和文化方面爲我們帶來了重大的利益，" for a 1-2 match.

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 fragment $e_1, e_2,\ldots,e_m$ and C be a Chinese paragraph $c_1, c_2,\ldots,c_n$, which e_i and c_j is a text-fragment as described in Section 2. We define a **link** $l(e_i, c_j)$ for e_i and c_j that are translation (or part of a translation) of one another. We define **null link** $l(e_i, c_0)$ for e_i which does not correspond to a translation. 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-fragment 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 $l_k = (e_{ik}, c_{jk})$ 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*(δ | *match*) *Prob*(*match*) and *Prob*(δ | *match*) can be estimated by

Prob(δ | *match*) = 2 (1 – *Prob*(|δ|))

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 (2003), for punctuation-based model, we know:

$P_{pun}(e_i, c_j) = P(pe_i, pc_j)P(|pe_i|, |pc_j|)$ for some $l_k = (e_i, c_j)$

where e_i and c_i is λ, one, or two punctuations,

e_i, c_j = English and Chinese text-block
pe_i = the ending English punctuations of e_i, $i = 1, m$
pc_j = the ending Chinese punctuations c_j, $j = 1, n$,
$P(pc_i, pe_i)$ = probability of pc_i translates into pe_i,

Thus, for each link l_k given *E*, *C* and *l*, we can compute the probability as follows:

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

$$P(A \mid E,F) = \prod_{k=1}^{t} P(\delta_k)P(m_k)P_{pun}(l_k)$$

4. Experimental results

In order to assess the performance of our sub-sentential alignment model, we run the system on two test cases:

1. Official record of proceedings of Hong Kong Legislative Council at Oct. 7, 1992,

2. Harry Potter Book I Chapter one.

For probability of punctuation, we use a small set of hand aligned data which led to the following model parameters:

1. Punctuation translation probability (Table 1),

2. Sentence match type probability (Table 2).

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. Match probability of clauses

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
Other	0.0066

Table 3. Performance evaluation for the two test cases

Test cases	# of para-graphs	# of matches	# of correct matches	Precision (%)
Official record of proceedings of Hong Kong Legislative Council	10	188	174	93
Harry Potter Book I Chapter 1	110	634	540	85

Preliminary results shown in Table 3 indicate precision rates of 85% and 93% for the two test cases.

5. Discussion and future work

We propose a model interleaving length-based text alignment and punctuation alignment to carry out sub-sentential alignment. The method seems to work reasonably well with an average precision rate around

90% in the evaluation of a preliminary implementation. There is still a lot of room for improvement. We are currently working on identification of more punctuations useful for sub-sentential alignment, proper segmentation of text ending with punctuations, and better model for lengths of sub-sentential fragment. We are also looking into the issues of best weighting scheme of length and punctuation information. Finally, the cases where there is inverion of translated fragments are difficult to handle with length information alone. We are also preparing to work with additional lexical information to solve this kind of problem in the future.

Acknowledgements

We acknowledge the support for this study through grants from Ministry of Education, Taiwan (MOE EX-91-E-FA06-4-4). Thanks are also due to Jim Chang for preparing the training data and evaluating the experimental results.

References

Church, K and P. Hank, "Word Association Norms, Mutual Information, and Lexicography," Computational Linguistics, 16:1, 1990, pp. 22-29.

Sproat, Chinese Word Segmentation, *First International Conference on Language Resources & Evaluation: Proceedings*, 1998, pp. 417— 420.

Richard Sproat, Chilin Shih, 1990, A statistical method for finding word boundaries in Chinese text. Computer Processing of Chinese & Oriental Languages, 4(4): 336-351.

R. Sproat, Chilin Shih, William Gale, and Nancy Chang. 1996. A stochastic finite-state word-segmentation algorithm for chinese. Computational Linguistics, 22(3): 377–404.

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.

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：