

Adaptive Sentence Alignment based on Length and Lexical Information

Thomas C. Chuang
Department of Computer Science
Van Nung Institute of Technology
Chungli, Taoyuan, Taiwan, ROC
tomchuang@cc.vit.edu.tw

Jason S. Chang
Department of Computer Science
National Tsing Hua University
Hsinchu, Taiwan, ROC
jschang@cs.nthu.edu.tw

Abstract

This prototype system demonstrates a novel sentence alignment method for bilingual texts based on adaptive learning and lexical information. The system aligns bilingual text at the paragraph level first and acquires length related statistics for the subsequent sentence alignment process. In addition to lengths, a probabilistic translation lexicon is utilized to further enhance the precision. The system is especially effective in the case of noisy translations produced in either translation direction that may involve different domains.

1 Introduction

The length-based approach (Church and Gale 1991; Brown et al. 1992) to sentence alignment produces surprisingly good results for French and English pair at success rates well over 95%. However, it does not fair as well for the English-Chinese task. Gale and Church indicated that the length-based alignment method was insensitive to length distribution parameters of mean and variance when dealing with translation languages in the same family. However, when working with the languages from different language families, we found that the performance of sentence alignment is very sensitive to the quantities of these parameters. The value of length ratio ranges from 2.8 to 4.8 for Chinese and English translations. The associated variance also varies in a wide range.

The performance sensitivity can be further accentuated by the translation direction and domain of the text. For instance, the accuracy rate of a length-based aligner dropped from 98.2% down to

85.6% when trained on English-to-Chinese technical menus and tested on Chinese-to-English bilingual articles in a general-interest magazine (Sinorama Magazine). Directionality is a more critical factor that accounts for the wide variation in length ratio and performance than the domain.

We demonstrate a sentence alignment program developed to cope with such problems arising from handling bilingual texts from diverse language families. The solution hinges on adaptive, tighter estimation of length-related parameters and incorporation of lexical information.

2 Adaptive Sentence Alignment

The system proceeds in two passes: paragraph alignment and sentence alignment within aligned paragraph. The two-pass scheme allows the values of paragraph length ratio to be used for sentence alignment. This simple feed-forward scheme works out with remarkable robustness and accuracy in aligning texts across different languages, domains, genres and translation directions. *A priori* information is not required on the domain or translation directionality of the bilingual text.

In this prototype, the sentence alignment process involves:

- Supplying a set of initial values of the mean and the variance of length ratio measured in characters between the source and target languages from an independent source (Longman 1992).
- Generating an optimum set of length distribution parameters by aligning paragraphs based on initial length distribution estimation and a probabilistic word translation model trained on a 90,000-entry BDC Chinese-to-English Online Dictionary (Chang, Yu and Lee 2002).
- Calculating the mean and variance based on the length ratios of the aligned paragraphs to be used

in a Gaussian estimation of probability of mutual translatability of sentences based on lengths.

- Finding sentence alignment using a dynamic programming procedure to optimize both length-based and lexical probabilities.

The prototype shows some interesting observations and results:

- For the Chinese-English task, the character length distribution is indeed close to Gaussian and length based sentence alignment scheme can be applied effectively to the Chinese-English task.
- Longer, modern Chinese word forms are used to translate technical menus written in English, while more condensed, classic word forms are used in reportage genre where frequent uses of allusion, idioms, and proverbs are quite evident.
- A self-adaptive calibration method can be developed to cope with wide variation in length ratios.
- The combination of length and lexical information generates remarkable results

The aligner's innovative way of handling diversity in length statistics and incorporation of lexical information enables it to effectively align bilingual texts across diverse languages, domains, genres and translation directions with accuracy rates approaching 99%, which has been achieved only for French-English task. The program aligns bilingual texts from *Harry Potter* (English-to-Chinese, novel, children reading), *Scientific American* (English-to-Chinese, reportage, science and technology), and the *bilingual Sinorama Magazine* (Chinese-to-English, reportage, general interest) with consistently high accuracy rates.

The special capability of the prototype to handle omission and insertion differentiates itself from other methods described in the literature. Experiment showed that the prototype spotted three untranslated sentences in *Harry Potter*, Vol. I, Chapter 2, published in Taiwan:

"Oh." Dudley sat down heavily and grabbed the nearest parcel. "All right then."

3 Conclusion

Combining adaptive learning and lexical information represents an innovative way to automatically optimize the alignment of text and translation in very different languages. It is especially effective

in the case of cross-domain noisy translations with omission and/or insertion, produced in either direction.

Acknowledgements

This work is partially supported by a NSC grant, NSC 90-2411-H-007-033-MC. Thanks are due to Wong Ying of Sinorama Magazine and Ivan Tsai of Yuan Liu Publishing and Scientific American, Taiwan for providing bilingual corpora for the experiments.

References

- Chang, JS, D. Yu and CJ Lee 2002. Statistical Phrase Translation Model, *Journal of Computational Linguistics and Chinese Language Processing*, forthcoming.
- BDC 1992 The BDC Chinese-English electronic dictionary (version 2.0), Behavior Design Corporation, Taiwan.
- Brown, P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Mercer R. L., and Roosin P. S. 1988 A Statistical Approach to Language Translation, In *Proceedings of the 12th Coling Conference*, Budapest, Hungary, pp. 71-76.
- Kay, M. and Röscheisen M. 1988 Text-Translation Alignment, Technical Report P90-00143, Xerox Palo Alto Research Center.
- Ker, S. J. and Chang J. S. 1997 A Class-base Approach to Word Alignment, *Computational Linguistics*, 23/2, pp. 313-343.
- Longman Group 1992 Longman English-Chinese Dictionary of Contemporary English, Published by Longman Group (Far East) Ltd., Hong Kong.
- Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer, 1991. Aligning Sentences in Parallel Corpora, In *Proc. of the Annual Meeting of the ACL* 29, pp. 169-176.
- Stanley F. Chen, 1993. Aligning Sentences in Bilingual Corpora Using Lexical Information, In *Proc. of ACL* 30, pp. 9-16.
- William A. Gale and Kenneth W. Church, 1991. A Program for Aligning Sentences in Bilingual Corpora, *Computational Linguistics* 19:75-102.
- Dekai Wu, 1994. Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria, In *Proc. of ACL* 31, pp. 80-87.