Subsentential Translation Memory for Computer Assisted Writing and Translation

Jian-Cheng Wu
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road, Hsinchu, 300, Taiwan, ROC
D928322@oz.nthu.edu.tw

Thomas C. Chuang
Department of Computer Science
Van Nung Institute of Technology
No. 1 Van-Nung Road
Chung-Li Tao-Yuan, Taiwan, ROC
tomchuang@cc.vit.edu.tw

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

Abstract

This paper describes a database of translation memory, *TotalRecall*, developed to encourage authentic and idiomatic use in second language writing. *TotalRecall* is a bilingual concordancer that support search query in English or Chinese for relevant sentences and translations. Although initially intended for learners of English as Foreign Language (EFL) in Taiwan, it is a gold mine of texts in English or Mandarin Chinese. *TotalRecall* is particularly useful for those who write in or translate into a foreign language. We exploited and structured existing high-quality translations from bilingual corpora from a Taiwan-based Sinorama Magazine and Official Records of Hong Kong Legislative Council to build a bilingual concordance. Novel approaches were taken to provide high-precision bilingual alignment on the subsentential and lexical levels. A browser-based user interface was developed for ease of access over the Internet. Users can search for word, phrase or expression in English or Mandarin. The Web-based user interface facilitates the recording of the user actions to provide data for further research.

1 Introduction

Translation memory has been found to be more effective alternative to machine translation for translators, especially when working with batches of similar texts. That is particularly true with so-called delta translation of the next versions for publications that need continuous revision such as an encyclopaedia or user's manual. On another area of language study, researchers on English Language Teaching (ELT) have increasingly looked to concordancer of very large corpora as a new re-source for translation and language learning. Concordancers have been indispensable for lexicographers. But now language teachers and students also embrace the concordancer to foster data-driven, student-centered learning.

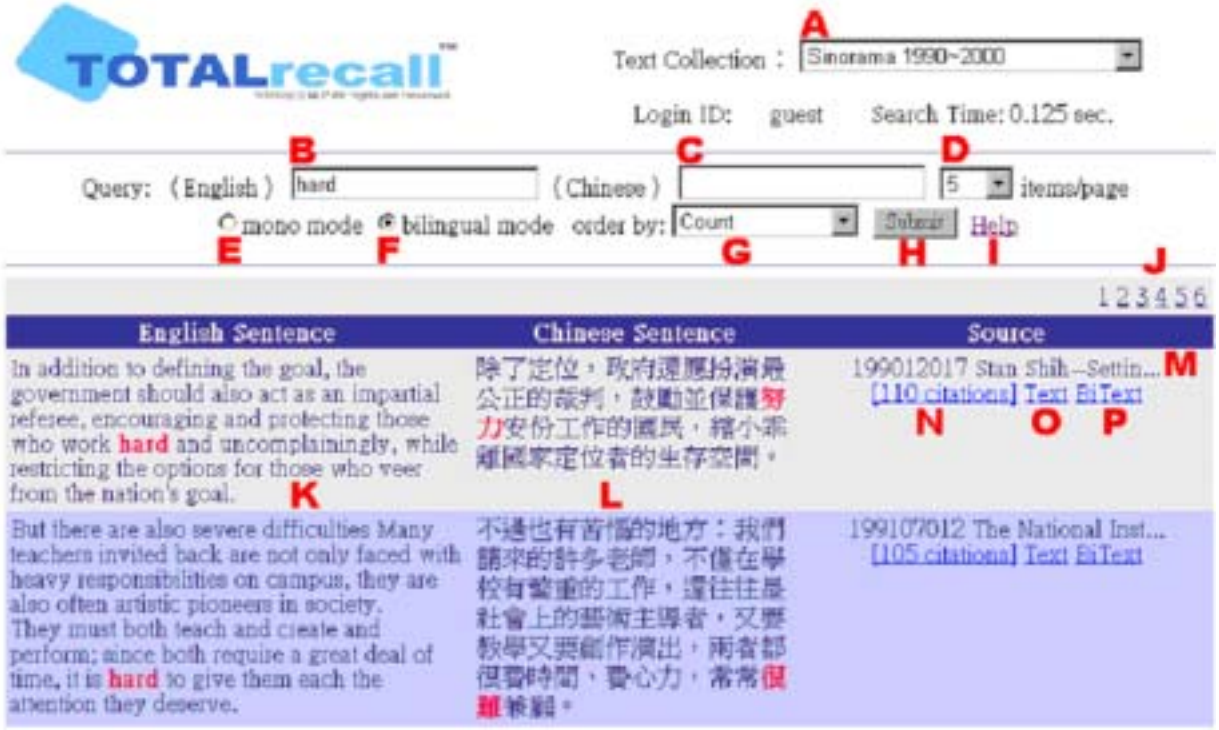
A bilingual concordance, in a way, meets the needs of both communities, the computer assisted translation (CAT) and computer assisted language learning (CALL). A bilingual concordancer is like a monolingual concordance, except that each sentence is followed by its translation counterpart in a second language. "Existing translations contain more solutions to more translation problems than any other existing resource." (Isabelle 1993). The same can be argued for language learning; existing texts offer more answers for the learner than any teacher or reference work do.

However, it is important to provide easy access for translators and learning writers alike to find the relevant and informative citations quickly. For in-stance, the English-French concordance system, TransSearch provides a familiar interface for the users (Macklovitch et al. 2000). The user type in the expression in question, a list of citations will come up and it is easy to scroll down until one finds translation that is useful much like using a search engine. TransSearch exploits sentence alignment techniques (Brown et al 1990; Gale and Church 1990) to facilitate bilingual search at the granularity level of sentences.

In this paper, we describe a bilingual concordancer which facilitate search and visualization with fine granularity. *TotalRecall* exploits subsentential and word alignment to provide a new kind of bilingual concordancer. Through the interactive interface and clustering of short subsentential bi-lingual citations, it helps translators and non-native speakers find ways to translate or express them-selves in a foreign language.

2 Aligning the corpus

Central to *TotalRecall* is a bilingual corpus and a set of programs that provide the bilingual analyses to yield a translation memory database out of the bilingual corpus. Currently, we are working with



A: Database selection B: English query C: Chinese query D: Number of items per page
E: Normal view F: Clustered summary according to translation G: Order by counts or lengths
H: Submit bottom I: Help file J: Page index K: English citation L: Chinese citation M: Date and title
N: All citations in the cluster O: Full text context P: Side-by-side sentence alignment

Figure 2. The results of searching for "hard"

bilingual corpora from a Taiwan-based Sinorama Magazine and Official Records of Hong Kong Legislative Council. A large bilingual collection of Studio Classroom English lessons will be provided in the near future. That would allow us to offer bilingual texts in both translation directions and with different levels of difficulty. Currently, the articles from Sinorama seems to be quite usefully by its own, covering a wide range of topics, reflecting the personalities, places, and events in Taiwan for the past three decades.

The concordance database is composed of bi-lingual sentence pairs, which are mutual translation. In addition, there are also tables to record additional information, including the source of each sentence pairs, metadata, and the information on phrase and word level alignment. With that additional information, *TotalRecall* provides various functions, including 1. viewing of the full text of the source with a simple click. 2. highlighted translation counterpart of the query word or phrase. 3. ranking that is pedagogically useful for translation and language learning.

We are currently running an operational system with Sinorama Magazine articles and HK LEGCO records. These bilingual texts that go into *TotalRecall* must be rearranged and structured. We describe the main steps below:

2.1 Subsentential alignment

While the length-based approach (Church and Gale 1991) to sentence alignment produces very good results for close language pairs such as French and English at success rates well over 96%, it does not fair as well for disparate language pairs such as English and Mandarin Chinese. Also sentence alignment tends to produce pairs of a long Chinese sentence and several English sentences. Such pairs of mutual translation make it difficult for the user to read and grasp the answers embedded in the retrieved citations.

We develop a new approach to aligning English and Mandarin texts at sub-sentential level in parallel corpora based on length and punctuation marks.

The subsentential alignment starts with parsing each article from corpora and putting them into the database. Subsequently articles are segmented into subsentential segments. Finally, segments in the two languages which are mutual translation are aligned.

Sentences and subsentenial phrases and clauses are broken up by various types of punctuation in the two languages. For fragments much shorter than sentences, the variances of length ratio are larger leading to unacceptably low precision rate

for alignment. We combine length-based and punctuation-based approach to cope with the difficulties in subsentential alignment. Punctuations in one language translate more or less consistently into punctuations in the other language. Therefore the information is useful in compensating for the weakness of length-based approach. In addition, we seek to further improve the accuracy rates by employing cognates and lexical information. We experimented with an implementation of the pro-posed method on a very large Mandarin-English parallel corpus of records of Hong Kong Legislative Council with satisfactory results. Experiment results show that the punctuation-based approach outperforms the length-based approach with precision rates approaching 98%.

Subsentential alignment results

From 1983 to 1991, the average rate of wage growth for all trades and industries was only 1.6%.
八三至九一年全部行業的平均工資增長率僅得 1.6%,
This was far lower than the growth in labour productivity, which averaged 5.3%.
遠較勞動生產力平均增長率的 5.3%為低,
But, it must also be noted that the average inflation rate was as high as 7.7% during the same period.
但同期的平均通脹率卻高達 7.7%,
As I have said before, even when the economy is booming, the workers are unable to share the fruit of economic success.
正如我之前所說，縱使經濟前景良好，勞工也無從分享經濟成果。

Figure 1 The result of subsentential alignment and collocation alignment.

2.2 Word and Collocation Alignment

After sentences and their translation counterparts are identified, we proceeded to carry out finer-grained alignment on the word level. We employed the Competitive Linking Algorithm (Melamed 2000) produce high precision word alignment. We also extract English collocations and their translation equivalent based on the result of word alignment. These alignment results were subsequently used to cluster citations and highlight translation equivalents of the query.

3 Aligning the corpus

TotalRecall allows a user to look for instances of specific words or expressions and its translation counterpart. For this purpose, the system opens up two text boxes for the user to enter queries in any or both of the two languages involved. We offer some special expressions for users to specify the following queries:

- Single or multi-word query – spaces between words in a query are considered as “and.” For disjunctive query, use “||” to de-note “or.”
- Every word in the query will be expanded to all surface forms for search. That includes singular and plural forms, and various tense of the verbs.
- *TotalRecall* automatically ignore high frequency words in a stoplist such as “the,” “to,” and “of.”
- It is also possible to ask for exact match by submitting query in quotes. Any word within the quotes will not be ignored. It is useful for searching named entities.

Once a query is submitted, *TotalRecall* displays the results on Web pages. Each result appears as a pair of segments in English and Chinese, in side-by-side format. A “context” hypertext link is included for each citation. If this link is selected, a new page appears displaying the original document of the pair. If the user so wishes, she can scroll through the following or preceding pages of context in the original document. *TotalRecall* present the results in a way that makes it easy for the user to grasp the information returned to her:

- When operating in the monolingual mode, *TotalRecall* presents the citation according to lengths.
- When operating in the bilingual mode, TotalRecall clusters the citations according to the translation counterparts and presents the user with a summary page of one example each for different translations. The query words and translation counterparts are high-lighted.

4 Conclusion

In this paper, we describe a bilingual concordance designed as a computer assisted translation and language learning tool. Currently, TotalRecll uses Sinorama Magazine and HKLEGCO corpora as the databases of translation memory. We have already put a beta version on line and experimented with a focus group of second language learners. Novel features of *TotalRecall* include highlighting of query and corresponding translations, clustering and ranking of search results according translation and frequency.

TotalRecall enable the non-native speaker who is looking for a way to express an idea in English or Mandarin. We are also adding on the basic functions to include a log of user activities, which will record the users’ query behavior and their background. We could then analyze the data and find useful information for future research.

Acknowledgement

We acknowledge the support for this study through grants from National Science Council and Ministry of Education, Taiwan (NSC 91-2213-E-007-061 and MOE EX-92-E-FA06-4-4) and a special grant for preparing the Sinorama Corpus for distri-bution by the Association for Computational Lin-guistics and Chinese Language Processing.

References

Brown P., Cocke J., Della Pietra S., Jelinek F., Lafferty J., Mercer R., & Roossin P. (1990). A statistical approach to machine translation. Computational Linguistics, vol. 16.

Gale, W. & K. W. Church, "A Program for Aligning Sen-tences in Bilingual Corpora" Proceedings of the 29th An-nual Meeting of the Association for Computational Linguistics, Berkeley, CA, 1991.

Isabelle, Pierre, M. Dymetman, G. Foster, J-M. Jutras, E. Macklovitch, F. Perrault, X. Ren and M. Simard. 1993. Translation Analysis and Translation Automation. In Pro-ceedings of the Fifth International Conference on Theoreti-cal and Methodological Issues in Machine Translation, Kyoto, Japan, pp. 12-20.

I. Dan Melamed. 2000. Models of translational equivalence among words. Computational Linguistics, 26(2):221–249, June.