TotalRecall: A Bilingual Concordance in National Digital Learning Project - CANDLE

Jian-Cheng Wu, Wen-Chi Shei , Jason S. Chang

Department of Computer Science

National Tsing Hua University

101, Kuangfu Road, Hsinchu, 300, Taiwan, ROC

{g904307, jschang }@cs.nthu.edu.tw

Abstract

This paper describes a Web-based English-Chinese concordance system, TotalRecall, being developed in National Digital Learning Project – CANDLE, to promote translation reuse and encourage authentic and idiomatic use in second language learning. We exploited and structured existing high-quality translations from the bilingual Sinorama Magazine to build the concordance of authentic text and translation. Novel approaches were taken to provide high-precision bilingual alignment on the sentence, phrase and word levels. A browser-based user interface also developed for ease of access over the Internet. Users can search for word, phrase or expression in English or Chinese. The Web-based user interface facilitates the recording of the user actions to provide data for further research.

1 Introduction

A concordance tool is particularly useful for studying a piece of literature when thinking in terms of a particular word, phrase or theme. It will show how often and where a word occurs, so can be helpful in building up some idea of how different themes recur within an article or a collection of articles. Concordances have been indispensable for lexicographers and increasingly considered instrumental for promoting learning effectiveness and motivation for language instructors and learners. A bilingual concordance tool is like a monolingual concordance, except that each sentence is associated with translation counterpart in a second language. It could be extremely useful for bilingual lexicographers, human translators and second language learners. Pierre Isabelle, in 1993, pointed out: “existing translations contain more solutions to more translation problems than any other existing resource.” It is particularly useful and convenient when the resource of existing translations is made available on the Internet. Web based bilingual concordances have proved to be very useful and popular. For example, the English-French concordance system, *TransSearch* (Macklovitch et al. 2000), provides a familiar interface for the users who only need to type in the expression in question, a list of citations will come up and it is easy to scroll down until one finds one that is useful.

In addition to the similar functionalities provided by *TransSearch*, **TotalRecall** comes with an additional feature making the *solution* more easily recognized: the user not only gets all the citations related to the expression in question, but also gets to see the translation counterpart highlighted.

TotalRecall extends the translation memory technology and provide an interactive tool intended for translators and non-native speakers trying to find ideas to properly express themselves. **TotalRecall** empowers the user by allowing her to take the initiative in submitting queries for searching authentic, contemporary use of English. These queries may be single words, phrases, or longer expressions, the system will search a substantial and relevant corpus and return bilingual citations that are helpful to human translators and second language learners.

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 collection of Chinese-English articles from the Sinorama magazine. Two additional bilingual collections: Studio Classroom English lessons and Hansard of Hong Kong Legislative Council are now in the work. That would allow us to offer bilingual texts in both translation directions and with different levels of difficulty. Currently, the articles from Sinaroma 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 decade.

The concordance database is composed of bilingual sentence pairs. 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 words or phrases.
3. Ranking that is pedagogically useful for translation and language learning.

We are currently running an experimental prototype with Sinorama articles, dated mainly from 1990 to 2000. There are approximately 50,000 bilingual sentences and over 2 million words in total. We also plan to continuously updating the database with newer information from Sinorama magazine so that the concordance is kept up to date and relevant.

The bilingual texts that go into **TotalRecall** must be rearranged and structured. We describe the main steps below:

2.1 Sentence Alignment

After parsing each article from files and put them into the database, we need to segment articles into sentences and align them into pairs of mutual translation. While the length-based approach (Church and Gale 1991) to sentence alignment produces surprisingly good results for the close language pair of French and English at success rates well over 96%, it does not fair as well for distant language pairs such as English and Chinese. Work on sentence alignment of English and Chinese texts (Wu 1994), indicates that the lengths of English and Chinese texts are not as highly correlated as in French-English task, leading to lower success rate (85-94%) for length-based aligners.

Simard, Foster, and Isabelle (1992) pointed out cognates in two close languages such as English and French can be used to measure the likelihood of mutual translation. However, for the English-Chinese pair, there are no orthographic, phonetic or semantic cognates readily recognizable by the computer. Therefore, the cognate-based approach is not applicable to the Chinese-English tasks.

At first, we used the length-based method for sentence alignment. The average precision of aligned sentence pairs is about 95%. We are now using a newer alignment method based on punctuation statistics. Although the average ratio of the punctuation counts in a text is low (less than 15%), punctuations provide valid additional evidence, helping to achieve high degree of alignment precision. It turns out that punctuations alone are telling evidences for sentence alignment, if we do more than hard matching of punctuations and take into consideration of intrinsic sequencing of punctuation in ordered comparison. Experiment results show that the punctuation-based approach outperforms the length-based approach with precision rates approaching 98%.

2.2 Phrase and Word Alignment

After sentences and their translation counterparts are identified, we proceeded to carry out finer-grained alignment on the phrase and word levels. We employ part of speech patterns and statistical analyses to extract bilingual phrases/collocations from a parallel corpus. The preferred syntactic patterns are obtained from idioms and collocations in the machine readable English-Chinese version of Longman Dictionary of Contemporary of English.

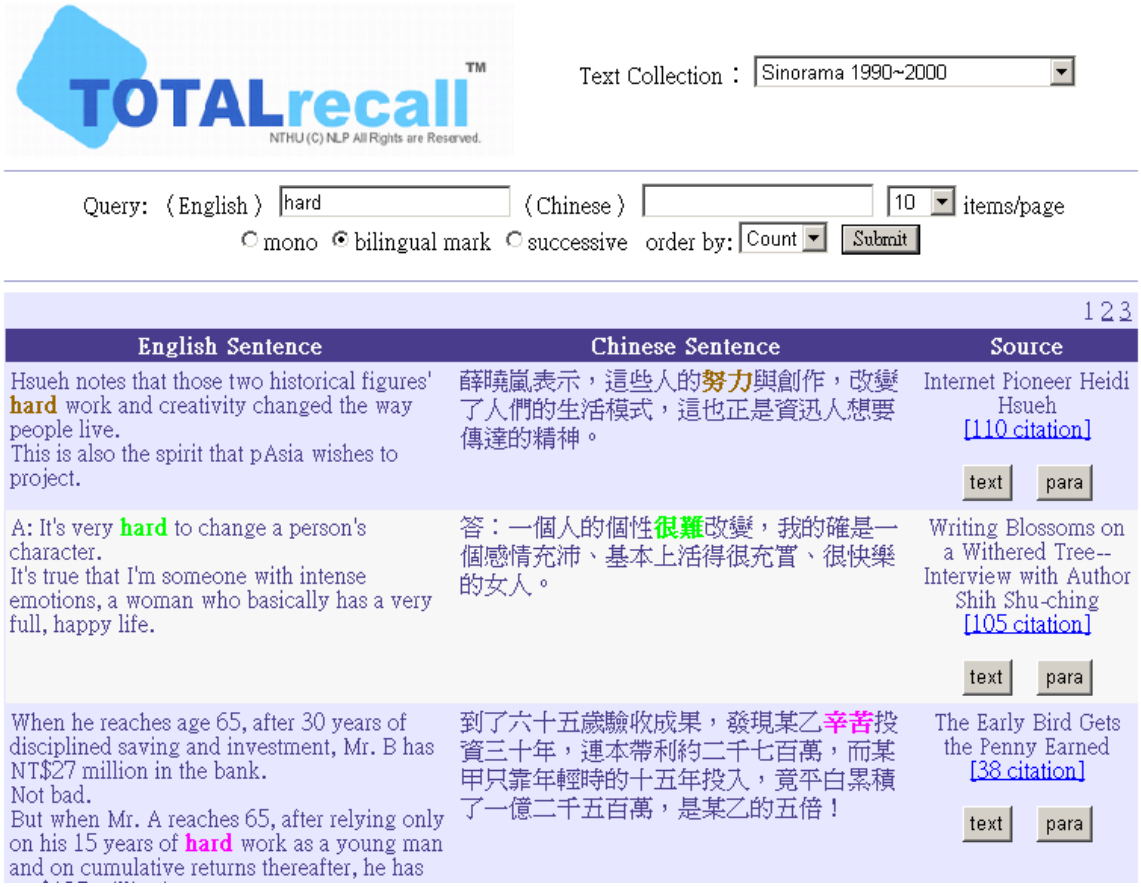

Figure 1. The results of searching for “hard” with citation ranking by counts of word and translation pairs.

Phrases matching the patterns are extract from aligned sentences in a parallel corpus. Those phrases are subsequently matched up via cross linguistic statistical association. Statistical association between the whole phrase as well as words in phrases are used jointly to link a collocation and its counterpart collocation in the other language. See Table 1 for an example of extracting bilingual collocations. The word and phrase level information is kept in relational database for use in processing queries, highlighting translation counterparts, and ranking citations. Sections 3 and 4 will give more details about that.

3 The Queries

The goal of the **TotalRecall** System is to allow a user to look for instances of specific words or expressions. For this purpose, the system opens up two text boxes for the user to enter queries in any one of the languages involved or both. We offer some special expressions for users to specify the following queries:

Table 1 The result of Chinese collocation candidates extracted. The shaded collocation pairs are selected based on competition of whole phrase log likelihood ratio and word-based translation probability. Un-shaded items 7 and 8 are not selected because of conflict with previously chosen bilingual collocations, items 2 and 3.

No.	English collocations	Chinese collocations	LLR	Word Prob
1.	iron rice bowl	鐵飯碗	103.3	0.0202
2.	performance review bonus	考績 獎金	63.03	0.1374
3.	year-end bonus	年終獎金	59.21	0.0700
4.	civil service rice	公家 飯	29.08	0.0378
5.	economic downturn	經濟 景氣 低迷	28.4	0.6961
6.	pay cut	減薪	28.4	0.0585
7.	year-end bonus	考績 獎金	27.35	0.2037
8.	performance review bonus	年終獎金	26.31	0.0370
9.	starve to death	餓不死	26.31	0.5670

- Exact single word query - W. For instance, enter “work” to find citations that contain “work,” but not “worked”, “working”, “works.”
- Exact single lemma query – W+. For instance, enter “work+” to find citations that contain “work”, “worked”, “working”, “works.”
- Exact string query. For instance, enter “in the work” to find citations that contain the three words, “in,” “the,” “work” in a row, but not citations that contain the three words in any other way.
- Conjunctive and disjunctive query. For instance, enter “give+ advice+” to find citations that contain “give” and “advice.” It is also possible to specify the distance between “give” and “advice,” so they are from a VO construction. Similarly, enter “hard | difficult | tough” to find citations that involve difficulty to do, understand or bear something, using any of the three words.

Once a query is submitted, **TotalRecall** displays the results on Web pages. Each result appears as a pair of segments, usually one sentence each in English and Chinese, in side-by-side format. The words matching the query are highlighted, and a “context” hypertext link is included in each row. If this link is selected, a new page appears displaying the bilingual paragraph or article, containing query.

4 Ranking

It is well known that the typical user has no patient to go beyond the first or second pages returned by a search engine. Therefore, ranking and putting the most useful information in the first one or two pages is of paramount importance for search engines. This is also true for a concordance.

Experiments with a focus group indicate that the following ranking strategies are important:

- Citations with a translation counterpart should be ranked first.
- Citations with a frequent translation counterpart appear before ones with less frequent translation
- Citations with same translation counterpart should be shown in clusters. The cluster can be called out entirely on demand.
- Ranking by nonlinguistic features should also be provided, including date, sentence length, query position in citations, relevancy as indicated via within document term frequency, etc.

With various ranking options available, the users can choose one that is most convenient and productive for the work at hand.

5 Conclusion

In this paper, we describe a bilingual concordance designed as a computer assisted translation and language learning tool. Currently, **TotalRecall** uses Sinorama Magazine corpus as the translation memory and will be continuously updated as new issues of the magazine becomes available. We have already put a beta version on line and experimented with a focus group of second language learners. The learners as well as their instructors seems to enjoy the novel features of **TotalRecall** including highlighting of query and corresponding translations, clustering and ranking of search results according translation and frequency.

TotalRecall enables the non-native speaker who is looking for a way to express an idea in English or Chinese. 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 90-2411-H-007-033-MC and MOE EX-91-E-FA06-4-4) and a special grant for preparing the Sinorama Corpus for distribution by the Association for Computational Linguistics and Chinese Language Processing.

References

Chuang, T.C. and J.S. Chang (2002), Adaptive Sentence Alignment Based on Length and Lexical Information, ACL 2002, Companion Vol. P. 91-2.

Gale, W. & K. W. Church, "A Program for Aligning Sentences in Bilingual Corpora" Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA, 1991.

Macklovitch, E., Simard, M., Langlais, P.: TransSearch: A Free Translation Memory on the World Wide Web. Proc. LREC 2000 III, 1201--1208 (2000).

Nie, J.-Y., Simard, M., Isabelle, P. and Durand, R.(1999) Cross-Language Information Retrieval based on Parallel Texts and Automatic Mining of Parallel Texts in the Web. Proceedings of SIGIR ’99, Berkeley, CA.

Simard, M., G. Foster & P. Isabelle (1992), Using cognates to align sentences in bilingual corpora. In Proceedings of TMI92, Montreal, Canada, pp. 67-81.

Wu, Dekai (1994), Aligning a parallel English-Chinese corpus statistically with lexical criteria. In The Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, New Mexico, USA, pp. 80-87.

Wu, J.C. and J.S. Chang (2003), Bilingual Collocation Extraction Based on Syntactic and Statistical Analyses, ms.

Yeh, K.C., T.C. Chuang, J.S. Chang (2003), Using Punctuations for Bilingual Sentence Alignment- Preparing Parallel Corpus for Distribution by the ACLCLP, ms.