How Similar are Chinese and Japanese for Cross-Language Information Retrieval?

Fredric C. Gey
UC Data Archive & Technical Assistance
University of California, Berkeley 94720-5100, USA
gey@berkeley.edu

Abstract

For NTCIR Workshop 5 UC Berkeley participated in the bilingual task of the CLIR track. Our focus was on Chinese topic searches against the Japanese News document collection, and on Japanese topic search against the Chinese News Document Collection. Extending our work of NTCIR 4 workshop, we performed search experiments to segment and use Chinese search topics directly as if they were Japanese topics and vice versa.

We also utilized a commercial Machine Translation (MT) between the two languages, with English as a pivot language. The best performance of Chinese topic search for Japanese documents was achieved using a hybrid approach which combined MT pivot translation with direct use of Chinese topic expressions.

Keywords: ***NTCIR, Cross-Language Information Retrieval***

1 Introduction

UC Berkeley has participated in all five NTCIR workshops, concentrating primarily on the Cross-Language Information Retrieval Tasks. In NTCIR-3 we also participated in the Patent Retrieval task. With reduced time and resources available to work on the NTCIR Workshop 4 [6] and 5 tasks, we limited our participation to a portion of the Bilingual task, specifically this time to search between the Chinese and Japanese languages. Our approach to CLIR has always been to apply translation resources to translate from the source language topics (query translation) to the target language of the document collection and then utilize tested monolingual retrieval document ranking algorithms. Our document ranking algorithm is probability model based using the technique of logistic regression (see Appendix).

2 Japanese and Chinese processing

As in NTCIR-4 [6], our methodology for processing Japanese documents in NTCIR-5 was to utilize the Chasen morphological analysis software (available from the site http://chasen.aist-nara.ac.jp/) to segment the Japanese document collection into words. Prior to NTCIR-4 participation, Berkeley used both n-grams and segmentation along alphabet boundaries to obtain word groupings of Katakana and Kanji character strings. In NTCIR-1 and NTCIR-2 we discarded all Hiragana words. By using Chasen in NTCIR-5, we preserved Hiragana for further indexing. In NTCIR-3 we found that word indexing performed equally well to n-gram indexing with less computational and storage overhead. All indexing was done excluding 241 Japanese stop-words prepared from Berkeley's participation in previous NTCIR workshops.

For Chinese retrieval we have found that overlapping bi-grams (sets of two Chinese characters extracted from a moving window which shifts forward one character at a time) have often produced the best results.

3 Using no translation for Chinese or Japanese topics

Because a significant fraction of the Japanese language (Kanji alphabet) is derived originally from the Chinese language, one approach to Chinese → Japanese CLIR is to utilize the Chinese topics without translation. This approach may be likened to Buckley's approach to English → French CLIR in the first TREC CLIR experiments [1], for which French words were assumed to be English cognates which could be identified through simple phonetic matching or spell-correction software. We reason that some portion of many Chinese topic titles, descriptions and narratives can be carried over into their Japanese equivalent without change. Consider, for example, NTCIR 5 CLIR Topic 003 ("Kim Dae Jun, Kim Jong Il, Inter-Korea Summit"). Compare the Chinese version of this topic,

<TITLE>金大中，金正日，南北韓高峰會
</TITLE>

with its Japanese version,

<TITLE>金大中，金正日，南北首脳会談
</TITLE>

We note that the two versions seem to be all visually nearly identical, and that the Japanese version consists entirely of Kanji characters. Of course while the topics above are visually similar, the underlying character representations are usually different because of the differing practices of data processing in Japanese and Chinese. To preserve the content while enabling term matching between the two languages, the methodology is simply to convert character sets from BIG5 (Chinese) to UTF-8 (Unicode) to EUC-J (Japanese) using the Unix ICONV utility.

By contrast, CLIR Topic 008 (“"ILOVEYOU", computer virus”) has the following Chinese and Japanese versions, respectively:

<TITLE>我愛你，電腦病毒</TITLE>

<TITLE>『ILOVEYOU』，コンピュータ・ウイルス</TITLE> wherein the Japanese version contains no content represented by Kanji characters.

Thus the simple approach of assuming an identity between Chinese and Japanese should work well for the topic 003 and very poorly for the topic 008. Indeed Berkeley's CLIR results confirm this supposition. For topic 003, Berkeley's official no-translation run *BRKLY-C-J-TDNC-02* achieved the highest MAP (0.5599) of all Chinese to Japanese runs for this topic. The same method for topic 008 achieved 0.0000 precision, the minimum over all Chinese to Japanese runs for this topic. Similarly for Japanese to Chinese cross-language search, Berkeley's official no-translation run *BRKLY-J-C-TDNC-03* achieved maximum MAP (0.4076) over all Japanese to Chinese runs for topic 003 and again for topic 008 achieved 0.0000 precision, the minimum over all Japanese to Chinese runs for this topic.

We can see that the approach shows considerable promise, but needs to be used judiciously in combination with other methods. If words from Chinese or Japanese topics cannot be translated into English or are mis-translated into English by the translation software, then the simple expedient of carrying over the Chinese words as if they were Japanese should help mitigate the damage of non-translation. We submitted a number of CLIR runs which applied this technique, either directly or as augmentation to query translation.

4 Official bilingual results

All Chinese→Japanese **(C→J)** CLIR runs shown below are for queries created from using SYSTRAN pivot translation between Chinese and Japanese in combination with Chinese bi-grams converted from BIG5 to EUC-J character sets. The Japanese→Chinese **(J→C)** runs are for SYSTRAN translation from Japanese topics to Chinese topics using English as a pivot language. Other runs used no translation (mere character representation conversion from the Japanese EUC-J to Chinese BIG5 via UTF-8). We were unable to create the combination runs for Chinese due to technical difficulties.

Berkeley submitted ten official CLIR runs to the NTCIR cross-language information retrieval task, focusing particularly on the pivot-bilingual subtask with the document collections in Japanese or Chinese. Rigid relevance performance of the runs is summarized below and is compared to the NTCIR workshop 5 maximum performance for either **C→J** or **J→C** by type.

Run BRKLY	**Translate Process**	**Berkeley MAP**	**MaxMAP (by type)**
C-J-T-05	SYST CJK + Chinese	0.2047	0.2684
C-J-TDNC-01	SYST CJK + Chinese	**0.2747†**	0.2747
C-J-D-04	SYST CJK	0.1639	0.2471
C-J-DN-03	SYST CJK	0.2692	0.2747
C-J-TDNC-02	No transl. Chinese	0.1231	0.2747
J-C-TDNC-01	SYST CJK	0.2695	0.2873
J-C-T-05	SYST CJK	**0.0925†**	0.0925
J-C-DN-02	SYST CJK	**0.2873†**	0.2873
J-C-TDNC-03	No transl. Japanese	0.1852	0.2873
J-C-D-04	SYST CJK	**0.1568†**	0.1568

† Best NTCIR MAP of type.

5. Effect of combination for Chinese to Japanese CLIR

In addition to our official runs, we performed additional experiments in order to determine the effect of the three components of translation (or just the use of Chinese as if it were Japanese). The following table displays mean average precision performance of Berkeley runs over the 47 topics with relevant Japanese documents.

Translate	**Title**	**Desc,**	**TDNC**
SYST CJK + Chinese	**0.2047***	0.2037	**0.2747***
SYST CJK only	0.1602	0.1639	0.2286
No translation	0.0925	0.0884	**0.1231***

*** Berkeley official run**. Other runs done for comparison

6 Query expansion with blind feedback

For NTCIR-5, Berkeley also augmented its document ranking formula with the application of blind relevance feedback to add terms to a query which might not be found in the initial natural language formulation of the topic. The process has three elements. First, an initial 'trial' retrieval is performed using the initial formulation of the query. Second, some number of top-ranked documents are assumed to be relevant and mined for additional query terms to be added to the initial query. Third, all query terms of the expanded are re-weighted and a second feedback retrieval run is performed to obtain the final document ranking. Details of this procedure may be found in our NTCIR-3 paper [3]. Our official results for NTCIR-5 were all submitted using blind relevance feedback by selecting 30 additional terms from the top 20 ranked documents of the initial retrieval. Choice of number of terms and documents for expansion was justified by experiments described in our NTCIR 4 paper [6]. After receipt of official results for NTCIR 5, we ran some additional experiments to test the validity of blind feedback query expansion.

The experiments, for **C**hinese←→**J**apanese cross-language retrieval are summarized in the table below, with the results of our official runs in boldface.

BF run	**CLIR Type**	**Title only**	**Desc, only**	**TDNC**
30terms 20 docs	**C→J**	**0.2047***	0.2037	**0.2747***
No BF	**C→J**	0.1098	0.1058	0.1855
30terms 20 docs	**J→C**	**0.0925***	**0.1568***	**0.2695***
No BF	**J→C**	0.0534	0.0752	0.1819

* **Berkeley official run**. Other runs done for comparison

The results show that blind feedback almost doubles the performance, except for TDNC where the performance improvement is still a remarkable 48%.

7 Conclusions and future research

Berkeley participated in NTCIR workshop 5 by experimenting with approaches to Cross Language Information Retrieval from Japanese to Chinese and vice versa. Our most novel idea for this bi-directional CLIR was to hypothesize that the Chinese and Japanese languages are identical and test whether this supposition can lead to decent retrieval results in searching between the two languages. We have found that when a Japanese version of an NTCIR topic consists of primarily Kanji text, then the use of the Chinese topic directly (after character code conversion) against Japanese documents (and vice versa) can produce very impressive results in terms of mean average precision for that topic. In addition, combining this approach with pivot language (English) machine translation produced substantially better results than translation alone.

The next area of research which should be investigated is whether transliteration (Romanization) of Chinese phonetic text using the Pinyin system can be matched to equivalent transliteration of Japanese Katakana text as a further technique to improve cross language search between the two languages. This would be similar to the work of Fujii and Ishikawa on transliteration between English, Korean and Japanese for NTCIR Workshop 4 [5].

8 Acknowledgment

This work could not have been accomplished without the advice and support of Aitao Chen, who wrote the basic logistic regression retrieval software, which was used for all the Berkeley NTCIR-5 runs. Dr. Chen, who now is associated with Yahoo Research, kindly lent some programming support to facilitate the NTCIR-5 runs from Berkeley.

References

[1] C Buckley, M Mitra, J Walz and C Cardie. Using Clustering and SuperConcepts within SMART: TREC-6, In: E M Voorhees and D K Harman, eds. *The Sixth Text Retrieval Conference (TREC-6), NIST Special publication 500-240.* pp. 107–124.

[2] A Chen and F Gey. Multilingual Information Retrieval Using Machine Translation, Relevance Feedback and Word Decompounding. *Information Retrieval,* 7 (1-2), 149-182, January - April 2004.

[3] A. Chen and F. Gey. Experiments in Cross-language and Patent Retrieval at NTCIR-3 Workshop, In *Proceedings of the Third NTCIR Workshop on research in Information Retrieval, Automatic Text*

Summarization and Question Answering, Tokyo, 173-182, October 2002.

[4] W Cooper, A Chen and F Gey. Full Text Retrieval based on Probabilistic Equations with Coefficients Fitted by Logistic Regression. In: Harman DK, ed. *The Second Text Retrieval Conference (TREC-2, NIST Special publication 500-215,* 57–64, April 1995.

[5] A Fujii and N Ishikawa. Cross-Language IR at University of Tsukuba: Automatic Transliteration for Japanese, English and Korean, in *Proceedings of Fourth NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, Tokyo,* June 2004.

[6] F Gey, Chinese and Korean Topic Search of Japanese News Collections, in *Proceedings of Fourth NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering, Tokyo,* June 2004.

Appendix: Document ranking

Berkeley has used a monolingual document ranking algorithm which uses statistical clues found in documents and queries to predict a dichotomous variable (relevance) based upon logistic regression fitting of prior relevance judgments. The exact formula is:

$$\log O(R \mid D,Q) = \log \frac{P(R \mid D,Q)}{1-P(R \mid D,Q)}$$

$$= \log \frac{P(R \mid D,Q)}{P(\bar{R} \mid D,Q)}$$

$$= -3.51 + 37.4 * x_1 + 0.330 * x_2$$

$$- 0.1937 * x_3 + 0.0929 * x_4$$

where $O(R \mid D,Q)$, $P(R \mid D,Q)$ mean, respectively, *odds* and *probability* of relevance of a document with respect to a query, and

$$x_1 = \frac{1}{\sqrt{n}+1} \sum_{i=1}^{n} \frac{qtf_i}{ql+35}$$

$$x_2 = \frac{1}{\sqrt{n}+1} \sum_{i=1}^{n} \log \frac{dtf_i}{dl+80}$$

$$x_3 = \frac{1}{\sqrt{n}+1} \sum_{i=1}^{n} \log \frac{ctf_i}{cl}$$

$$x_4 = n$$

where n is the number of matching terms between a document and a query, and
ql : query length
dl: document length
cl: collection length
qtf_i: the within-query frequency of the ith matching term
dtf_i: the within-document frequency of the ith matching term
ctf_i: the occurrence frequency of the ith matching term in the collection.

This formula has been used since the second TREC conference and for all NTCIR and CLEF cross-language evaluations [4].