Search Between Chinese and Japanese Text Collections

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

Abstract

For NTCIR Workshop 6 UC Berkeley participated in Phase 1 of the bilingual task of the CLIR track. Our focus was upon Japanese topic search against the Chinese News Document Collection and upon Chinese topic searches retrieving from Japanese News document collection. We performed search experiments to segment and use Chinese search topics directly as if they were Japanese topics and vice versa. We also utilized Machine Translation (MT) software between Japanese and Chinese, with English as a pivot language. While Chinese search without translation against Japanese documents performed credibly well for title only runs, the reverse (Japanese topic search of Chinese documents without translation) was poor. We are investigating the reasons.

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

1 Introduction

UC Berkeley has participated in all six 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] Workshop 5 [7] and Workshop 6 tasks, we limited our participation to a portion of the Bilingual task, specifically this time to search between the Japanese and Chinese 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 and NTCIR-5 [6, 7], our methodology for processing Japanese documents in NTCIR-6 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 and NTCIR-6, we preserved Hiragana for further indexing. We choose this approach because in NTCIR-3 we found that word indexing performed equally to n-gram indexing with less 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]. Dictionary segmentation of Chinese is limited by dictionary coverage and presents the usual out-of-vocabulary problems.

3 No translation for Chinese or Japanese topics

We know that a portion of the Japanese language (Kanji alphabet) is derived originally from the Chinese language Thus one approach to Chinese → Japanese CLIR is to utilize the Chinese topics without translation. This approach is similar 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. In NTCIR-5 we reasoned that some portion of many Chinese topic titles, descriptions and narratives can be carried over into their Japanese equivalent without change. For example, NTCIR 6 CLIR Topic 077 ("Director Takeshi Kitano's films"), we may compare the Japanese version of this topic,

<TITLE>北野武監督作品</TITLE> with its Chinese version,
<TITLE>北野武導演的電影 </TITLE>
We see that the two versions seem to be visually similar, and that the Japanese version consists almost entirely of Kanji characters. Of course while the topics above may be 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 018 ("Teenager, Social Problem") have the following Chinese and Japanese versions, respectively:
<TITLE>青少年，社會問題</TITLE>
<TITLE>ティーンエージャー，社会問題</TITLE> wherein the Japanese Kanji overlap to the Chinese seems only consists of the general term "Social Problem," while the critical word "Teenager" is represented in Katakana (phonetically rendered as "Dean ager" by the GOOGLE translator).

Thus the simple approach of assuming an identity between Chinese and Japanese might work very well for topic 077 and poorly for the topic 018. Indeed Berkeley's NTCIR-6 CLIR results confirm this supposition. For topic 77, Berkeley's official no-translation title run *BRKLY-C-J-T-04* achieved the highest Mean Average Precision (MAP over precision at 11 recall points) of 0.3902 of all Chinese to Japanese runs for this topic. The same method for topic 018 retrieved only retrieved only 2 of 43 relevant documents for an overall MAP 0.0004 precision, the minimum over all Chinese to Japanese runs for this topic. Similarly for Japanese to Chinese cross-language search, Berkeley's unofficial no-translation run for topic 077 achieved maximum MAP (0.5835) over all Japanese to Chinese runs and again for topic 018 achieved 0.000 precision (retrieving 0 of 77 relevant documents), 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.

4 Official bilingual results

Berkeley submitted eight official CLIR runs to the NTCIR cross-language information retrieval task, focusing particularly on the bilingual subtask with the document collections in Japanese or Chinese. Our Japanese to Chinese runs used the Google's translation capability for Japanese to English and the Systran CJK personal MT package for English to Chinese. We were unable to create the translation runs from Chinese to Japanese due to technical difficulties, so we only submitted "no translation" runs for **C→J**.

Rigid relevance performance of the runs is summarized below and is compared to the NTCIR workshop 6 maximum performance for either **C→J** or **J→C** by type.

Run BRKLY	Translate Process	Berkeley MAP	MaxMAP (by type)
C-J-T-04	No transl. Chinese	0.2738	0.3233
C-J-TDNC-01	No transl. Chinese	0.0606	0.2840
C-J-D-03	No transl. Chinese	02519	0.3118
C-J-DN-02	No transl. Chinese	0.2840	0.2840
J-C-TDNC-01	Google+ SYST CJK	0.1748	†
J-C-DN-02	Google+ SYST CJK	0.1659	†
J-C-D-03	Google+ SYST CJK	0.0770	†
J-C-T-04	Google+ SYST CJK	0.0471	†

† Not meaningful because only Berkeley submitted runs for this task.

We should note the wide disparity between different types of runs. Unexpectedly, the **C→J** Title and Description-Narrative run are the best performing, both by Berkeley and overall, while the J-C-TDNC run performs poorly. It seems than using more descriptive text from the D, N and C fields increases the noise of the translation between Chinese and Japanese. Conversely the **J→C** runs behave as expected, with the Title only run performing considerably worse than others.

5 Blind Feedback Query Expansion

For NTCIR-6 (similar to NTCIR-5), Berkeley 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 [2]. Our official results for NTCIR-6 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 6, 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**	**0.2738***	**0.2519**	**0.0606***
No BF	**C→J**	0.1098	0.1058	0.0320
30terms 20 docs	**J→C**	**0.0471***	**0.0770***	**0.1748***
No BF	**J→C**	0.0283	0.0518	0.1157

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

The results show that blind feedback more than doubles the performance for **C→J**, except for TDNC where the performance improvement is still a remarkable 89%. All Chinese→Japanese **(C→J)** CLIR runs shown above are for no translation, i.e. Chinese bi-grams converted from BIG5 to EUC-J character sets. The Japanese→Chinese **(J→C)** runs are for Google translation of Japanese topics to English and then SYSTRAN CJK personal translation of English topics to Chinese topics (a pivot language approach).

6 No Translation of J→C

Berkeley did not submit any official "no translation" runs between Japanese and Chinese. However in preparing this paper we decided to perform such experiments to see if the "no translation" option would work as well as it did for Chinese to Japanese. The results are summarized in the table below:

Run Type	CLIR Type	Title only	Desc, only	TDNC
MAP Pivot Translat.	**J→C**	**0.0471***	**0.0770***	**0.1748***
MAP No Translat.	**J→C**	0.0429	0.0407	0.0636
Topic18 Pivot Tr	**J→C**	0.0004*	0.0004*	0.0021*
T18 No Translat.	**J→C**	0.0000	0.0000	0.0464
Topic77 Pivot Tr.	**J→C**	0.0000*	0.0000*	0.0093*
T77 No Translat.	**J→C**	0.5835	0.6264	0.5517

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

We have added results for Topics 18 and 77 which were described in section 3. Topic 18 contained the Katakana word for "teenager" which was not translated by the GOOGLE online translation system.

7 Conclusions and future research

Berkeley participated in NTCIR workshop 6 Phase 1 by experimenting with approaches to Cross Language Information Retrieval from Japanese to Chinese and vice versa. W continued to explore our hypothesis that the Chinese and Japanese languages are have partially shared alphabets and to test whether this supposition can lead to decent retrieval results in searching between the two languages. We have again found that when a Japanese version of an NTCIR topic consists of primarily Kanji text, then use of the Chinese topic directly (after character code conversion) against Japanese documents can produce very impressive results in terms of mean average precision for that topic. However the reverse direction **J→C** no-translation did not provide comparable performance as was observed for **C→J**. We are proceeding with a failure analysis of why this asymmetry of performance exists. We note that MT systems do not adequately translate Katakana words, and that fuzzy matching of transliteration of Japanese Katakana text may also 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 work of Aitao Chen, now with Yahoo Research, who wrote the basic logistic regression retrieval software, which was used for all the Berkeley NTCIR-6 retrieval runs.

9 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. 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.

[3] Chen, A, J He, L Xu, F Gey and J Meggs, Chinese text retrieval without using a dictionary, *in SIGIR '97: Proceedings of the 20th annual international conference on Research and Development in information retrieval,* Philadelphia, USA, 1997, pp 42-49.

[4] Cooper W. S., Chen A and Gey F.C. 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,* April 1995 pp 57–64.

[5] Fujii, A 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] Gey, F, Chinese and Korean Topic Search of Japanese News Collections, in *Proceedings of Fourth NTCIR, Tokyo,* June 2004.

[7] Gey, F, How Similar are Chinese and Japanese for Cross-Language Information Retrieval, in *Proceedings of Fifth NTCIR Workshop, Tokyo,* December 2005, pp. 171-174.

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:

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

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

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

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

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].