Chinese and Korean Topic Search of Japanese News Collections

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

Abstract

UC Berkeley participated in the pivot bilingual task of the CLIR track at NTCIR Workshop 4. Our focus was on Chinese and Korean searches against the Japanese News document collection, using English as a pivot language. For comparison of our pivot techniques, we submitted Japanese monolingual and English → Japanese bilingual search rankings as well. Two different commercial translation software packages were used in quite different ways – one did standard query translation from Chinese or Korean topics to English and then to Japanese, while the other was used to translate the Japanese corpus to English word-by-word using 'fast document translation'. Another interesting search approach was to segment and use Chinese search topics directly as if they were Japanese topics

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

1 Introduction

This UC Berkeley team has participated in all four 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 tasks, we limited our participation to a portion of the Pivot Bilingual task. Our approach to CLIR has always been to apply translation resources to translate from the source language topics 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. For NTCIR-4 we tested a technique called 'fast document translation' which we have used with some success for European languages. In this way we translated the entire Japanese collection into English and then used the English translated queries for monolingual English retrieval and ranking.

2 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].

3 Japanese preprocessing

Our methodology for processing Japanese documents in NTCIR-4 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. In past NTCIR participation, Berkeley has used both n-grams and segmentation along alphabet boundaries to obtain word groupings of Katakana and Kanji character strings (in prior participations we discarded all Hiragana words – by using Chasen in NTCIR-4, 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.

4 Translation

4.1 Translation software

In NTCIR-4, our approach to translation from Korean and Chinese topics to English was to utilize a widely available software package, the SYSTRAN CJK personal system available for less than $US100. from www.systransoft.com. The package provides bi-directional translation between English natural language text and Chinese, Japanese and Korean. The other package used was L&H J-Surf which translates Japanese web pages to English; J-Surf was available as part of an EasyTranslate software package – this software was purchased at an E-Bay auction for $US22 – it only runs on Windows 98 and not later operating systems.

4.2 Fast Document translation

As a general practice, systems which do cross-language search do not usually attempt to translate the document collection into the query language because of the overwhelming computational and time resources required. However there is an approach which, somewhat imperfectly, can do this translation without an astronomical computational burden. The idea is to do word-by-word translation using a simple 1 ←→ 1 bilingual lexicon produced by a translation resource from the document collection corpus. The process is to collect the unique words from the corpus and submit each individually to the translation engine to obtain a unique word in the topic language. This involves a "guess without context" by the translation engine, but has sometimes proven useful for European languages [2].

To apply this process for Berkeley's submissions to NTCIR-4, we printed out the list of 214,906 unique Japanese words in the NTCIR-4 CLIR corpus as identified by Chasen morphologic processor. This excludes all occurrences of the 241 stop-words mentioned in the last-section This word list was split into 42 files of 5,000 words each and one file of 4906 words. Each file was edited to make it an HTML file and the J-Surf program was run individually on each file to attempt to translate each word in the file to its English equivalent. While we don't have complete statistics, a casual examination showed that less than half of the words in the corpus were actually translated into English. Even so, the official E-J run **(E-J-TDNC-04)** performed quite well. In other venues [2] we have merged rankings from independent translation sources to improve results. No attempt to do this was made for NTCIR-4.

4.3 No translation for Chinese

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 is somewhat akin to Buckley's approach to English → French CLIR in the first TREC CLIR experiments [1], in 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 most Chinese topic words can be carried over into their Japanese equivalent without change. If these words 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 three runs which applied this technique, either directly or as augmentation to query translation. The methodology is simply to convert character sets from BIG5 (Chinese) to UTF-8 (Unicode) to EUC-J (Japanese) using the Unix ICONV utility. In general the technique did not work very well as a stand-alone technique, although

for a few topics, such as Topic 030 (Animal Cloning Technique) the performance was excellent. The following xml fragment of this topic in 3 languages:

<TOPIC>
<NUM>030</NUM>
<TITLE-CH>動物複製技術</TITLE-CH>
<TITLE-JA>動物クローン技術</TITLE-JA>
<TITLE-EN>Animal Cloning Technique</TITLE-EN>
</TOPIC>

Shows the resemblance of the original Chinese to the Japanese if the non-Kanji characters are removed. Performance for this topic was especially high for "Chinese only" retrieval. After the NTCIR-4 Workshop, further examination was made of the topics which found that the Chasen analyzer often segmented the Chinese text into unigram characters which may have led to the poor results. A further experiment with bigram segmentation is described in Section 9 below.

5 Results

5.1 Berkeley bilingual and mono-lingual official runs

Berkeley submitted seventeen official runs to the NTCIR cross-language information retrieval task, focusing particularly on the pivot-bilingual subtask with the document collection in Japanese. Rigid relevance performance of the runs is summarized below:

Run Name BRKLY	Translate Process	Berkeley MAP	% of BRKLY Mono (by type)
J-J-T-03	none	0.3307	100.00%
J-J-D-02	none	0.3222†	100.00%
J-J-TDNC-01	none	0.3487	100.00%
C-J-T-01	SYST CJK + Chinese	0.1913	57.85%
C-J-TDNC-02	SYST CJK + Chinese	0.2511	72.01%
C-J-D-03	SYST CJK	0.1904	59.08%
C-J-T-04	SYST CJK	0.1748	52.86%
C-J-TDNC-05	No transl. Chinese	0.0893	25.61%
K-J-T-01	SYST CJK +manual web transl.	0.1626	49.17%
K-J-T-04	SYST CJK	0.1573	47.57%
K-J-D-03	SYST CJK	0.1402	49.17%
K-J-TDNC-02	SYST CJK	0.1852	53.11%
E-J-T-03	SYST CJK	0.1917	53.11%
E-J-D-02	SYST CJK	0.1874	57.96%
E-J-TDNC-01	SYST CJK	0.2522†	72.33%
E-J-TDNC-04	fast docum. translation	0.2267	65.01%

5.2 Comparing to NTCIR-4 medians

In this section we present the results of the 'best' Berkeley run for each language combination compared to the median MAP for that bilingual (or monolingual language combination (**O** means 'Other runs' in NTCIR terminology, for Berkeley the run was TDNC):†

Tlang-Dlang-Type	NTCIR Median	Berkeley MAP	% diff
J-J-T	0.3135	0.3307	5.49%
J-J-D	0.3352	0.3222†	-3.88%
J-J-O	0.3487	0.3487	0.00%
C-J-T	0.1748	0.1913	9.44%
C-J-D	0.1680	0.1904	13.37%
C-J-O	0.0893	0.2511	181.19%
K-J-T	0.1626	0.1626	0.00%
K-J-D	0.1648	0.1402	-14.93%
K-J-O	0.3303	0.1852	-43.92%
E-J-T	0.2284	0.1917	-16.07%
E-J-D	0.2615	0.1874	-28.34%
E-J-O	0.3099	0.2522†	-18.61%

In general, our title and description runs are below the median, while our TDNC runs perform better than NTCIR median. We attribute our poor performance primarily to the non-translation or mis-translation of central concept terms in title and narrative, as discussed in the sections below.

6 Translation problems

As is often the case in CLIR, the major reason for poor performance is the lack of translation or incorrect translation of critical words or phrases from the source language topic to the document collection language. This is best exemplified by Topic 008 with Japanese title: バイアグラ(Viagra). The xml fragment below summarizes the various translation values obtained by the Systran commercial translation package:

† For the two runs (identified by the symbol †), Berkeley submitted incorrectly identified results from other runs, the tables show the performance of corrected runs.

<TOPIC>
<NUM>008</NUM>
<SLANG>CH</SLANG>
<TITLE-CH>威而鋼</TITLE-CH>
<TITLE-CH-EN>prestige but steel</TITLE-CH-EN>
<TITLE-CH-EN-JP>威信しかし鋼鉄</TITLE-CH-EN-JP>

<TITLE-KR>비아그라</TITLE-KR>
<TITLE-KR-EN>Viagra</TITLE-KR-EN>
<TITLE-KR-EN-JP>Viagra</TITLE-KR-EN-JP>
<TITLE-KR-EN-JP-web>Viagra バイアグラ</TITLE-KR-EN-JP-web>
<TITLE-EN>Viagra</TITLE-EN>
<TITLE-JP>バイアグラ</TITLE-JP>
</TOPIC>

The SYSTRAN CJK package incorrectly translates the original Chinese title to the English phrase "prestige but steel", which is, in turn mis-translated into Japanese, yielding an mean average precision for Berkeley's C-J-T run for this topic of 0.0000. The CJK package correctly translates the Korean title to the English word "Viagra", but the package fails to translate the English word to its Japanese equivalent. The submission of this untranslated English word as a Japanese query retrieved no documents from the NTCIR-4 Japanese collection. When we used a manual web search (described in the next section), we found the correct Japanese translation "バイアグラ" and performance went from 0.0000 to 0.4468, close to the best NTCIR-4 performance of 0.5411 for K-J-T runs. This single improvement took our title overall MAP performance from 0.1573 to 0.1626.

7 Web translation techniques for out-of vocabulary words

As the previous section indicates, the primary reason for poor CLIR performance for our runs in NTCIR-4 were poorly translated or untranslated words or concepts. During NTCIR-3 Berkeley introduced a technique of searching the web for possible Chinese-English translation of out-of-vocabulary words. The central idea was to search for web pages containing the known word (usually in English) and attempt to find its Chinese translation equivalent within a few words of the location of the English word in the web page text. The technique is described in [3]. The general approach is to do a GOOGLE Boolean search of the type:

Find 'viagra' and language = 'Japanese'

The following is a screen capture of a web page found with this search:

Simply by choosing the Japanese term prominently displayed next to the English word on the web page and cut-and-paste from the page into the text for Topic 008 yields the translation needed with the results described above. We did not have time to implement an automatic technique for Japanese, so we simulated the process by doing a single manual search for the example of "Viagra" cited above. A more complete methodology and analysis of web vocabulary extraction techniques has been done by Zhang and Vines [5].

8 Query expansion with blind feedback

During the past two years, Berkeley has 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-4 were all submitted using blind relevance feedback by selecting thirty additional terms from the top 20 ranked documents of the initial retrieval. This parameterization was chosen based upon prior experience. After receipt of official results we ran some additional experiments to test the validity of query expansion and of our choice of parameters.

The experiments, for English-Japanese cross-language retrieval are summarized in the

table below, with the results of the official runs in boldface.

ndocs	nterms	title	TDNC
none	none	0.1194	0.1813
5	10	0.1608	0.2134
10	10	0.1831	0.2209
20	**30**	**0.1917**	**0.2522**

The results confirm our choice of parameters and decision to use blind feedback.

9 A further experiment using bi-gram segmentation for Chinese

After we discovered that the Chasen analyzer performed rather poorly in segmenting Chinese sentences as if they were sentences composed of Japanese Kanji characters, we decided to perform one last experiment. The Chinese topics were segmented into overlapping bi-grams and then matched to Japanese documents as if they were Japanese queries. This performance (in terms of average precision overall all points of recall) of 0.0908 was virtually indistinguishable from the 0.0893 precision of Chasen-segmented Chinese, and considerably below the 0.2511 precision of the official run **C-J-TDNC-02** which used MT for translation with English as a pivot language. However, there were five topics for which "Chinese as Japanese" performed comparable or better than translation:

Topic	C-J-TDNC-02	C-J-TDNC-05	C-J-bigram
3	0.3239	0.2428	0.3727†
9	0.1906	0.2606†	0.1023
11	0.3214	0.3368†	0.2316
14	0.4141†	0.4007	0.3597
16	0.2460	0.0113	0.2735†

† - best performance for topic

10 Summary

UC Berkeley participated in the Cross-Language Information Retrieval task by doing searches of the NTCIR Japanese News document collections from Chinese and Korean search topics. In particular we participated in the Pivot Language subtask using English as a pivot language. For comparison purposes we submitted Japanese monolingual and English → Japanese bilingual runs. Translation was done using two commercial translation packages, primarily the SYSTRAN CJK Personal package which translates from Chinese, Japanese and Korean to English. Our results, while acceptable, suffered mainly from poor translation of key concepts either from Chinese to English or from English to Japanese. The three innovative ideas we explored in these experiments were 1) segmentation and use of Chinese queries as if they were Japanese, 2) fast document translation of the NTCIR-4 Japanese CLIR corpus to English and 3) the search for Web pages which would translate key concepts from English to Japanese. Within the limited scope of our experiments, these techniques show promise for future improvements in CLIR.

10 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-4 runs.

11 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] 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*, 57–64, April 1995.

[5] Zhang, Y. and P. Vines, Using the Web for Automated Translation Extraction in Cross-Language Information Retrieval, *Proceedings of ACM-SIGIR-2004*, 162-169, July 2004.