# HIGH BASELINE JAPANESE INFORMATION RETRIEVAL FOR QUESTION-ANSWERING

Ray R. Larson and Fredric C. Gey
School of Information and UC Data Archive
University of California, Berkeley 94720-4600, USA
ray@ischool.berkeley.edu, gey@berkeley.edu

## Abstract

*For NTCIR Workshop 7 UC Berkeley participated in IR4QA (Information Retrieval for Question Answering) as well as the Patent Mining tracks. For IR4QA we only did Japanese monolingual search. Our focus was thus upon Japanese topic search against the Japanese News document collection as in past NTCIR participations. We preprocessed the text using the ChaSen morphological analyzer for term segmentation. We utilized a time-tested logistic regression algorithm for document ranking coupled with blind feedback. The results were satisfactory, ranking second among IR4QA overall submissions..*

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

## 1 Introduction

UC Berkeley has participated in all seven NTCIR workshops, concentrating primarily on the Cross-Language Information Retrieval Tasks. In NTCIR-3 we also participated in the Patent Retrieval task [1]. For the NTCIR Workshops NTCIR-4 [3], NTCIR-5 [4] and NTCIR-6 [5] tasks, we limited our participation to a portion of the Bilingual task, specifically this search between Japanese and Chinese languages. For NTCIR-7 we also participated in the Patent Mining task. Our document ranking algorithm is a probability model based using the technique of logistic regression [2] (see Appendix in our Patent Mining paper for details [6]). Between NTCIR-6 and NTCIR-7, the Berkeley Information School changed computer systems from Sun servers to Linux servers based upon the Intel architecture. This change meant that our production processing software would not work, so we chose this opportunity to move to the first author's Cheshire system. Cheshire has been used for the INEX XML retrieval evaluations [7, 8].

## 2 Japanese processing

As in NTCIR-4, NTCIR-5 and NTCIR-6 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-4 through NTCIR-7, 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.

## 3 Official Japanese monolingual results

Berkeley submitted four official runs for Japanese monolingual retrieval to the NTCIR-7 information retrieval for question answering task. Two of these runs were using the question title (T) only and two utilized title-narrative (DN). The only difference between the runs was the use of blind feedback for the better performing runs.

For NTCIR-7 (similar to previous NTCIRs), Berkeley augmented its document ranking formula with the application of blind relevance feedback (BF) to add terms to a query which might not be found in the initial natural language formulation of the topic. Feedback chose the top 10 terms from the top 10 ranked documents.

Relevance performance of the runs is summarized below and is compared to the NTCIR workshop 7 maximum performance for **J-J** by type.

| Run BRKLY | Type | BRKLY MAP | Max MAP T / DN |
|---|---|---|---|
| **JA-JA-02-T** | BF | 0.5838 | 0.6979 |
| **JA-JA-03-T** | No-BF | 0.5407 | 0.6979 |
| **JA-JA-02-DN** | BF | 0.6278 | 0.6278 |
| **JA-JA-02-DN** | No-BF | 0.5767 | 0.6278 |

As expected, utilization of blind feedback improved performance, but less than ten percent, unlike in previous NCTIR workshops where our blind feedback runs sometimes showed improvements of 40 to 50 percent. Also, as expected, our narrative runs performed about 7.5 percent better than our best title run. However, we should note that the best Japanese monolingual run overall for IR4QA was a title run only, outperforming our best title-narrative run [9].

## 5 Suggestions for IR4QA future

For NTCIR-7, the IR4QA task has only required a standard IR retrieval pattern and result structure. From the beginning of work on factoid question answering, it has been clear that QA is dependent more upon accurate *passage retrieval* than document retrieval. We recommend that if the IR4QA task continues in the future, the task should become somewhat more difficult – the IR4QA participants should be required to return ranked lists of best passages. Passages could either be defined as 250 byte overlapping chunks, or as "best paragraph" within the document, assuming the organizers would wish to consider tagging individual paragraphs within the IR4QA document collections.

## 6 Conclusions and future research

Berkeley participated in NTCIR workshop 7. IR4QA Japanese monolingual task only, because of a change in our hardware and software infrastructure. Our goal was to create a high baseline monolingual retrieval as a basis for IR4QA participation in future NTCIR evaluations.

## 7 References

[1] 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,* October 2002, pp 173-182.

[2] 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.

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

[4] 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.

[5] Gey, F, Search Between Chinese and Japanese Text Collections, in *Proceedings of Sixth NTCIR Workshop, Tokyo,* May 2007, pp. 73-76.

[6] Gey, F and R Larson, Patent Mining: A Baseline Approach, in *Proceedings of Seventh NTCIR Workshop, Tokyo,* December 2008 (this volume)

[7] Larson, R., A fusion approach to XML structured document retrieval. *Information Retrieval*, 8:601–629, 2005.

[8] Larson, R., Probabilistic retrieval approaches for thorough and heterogeneous xml retrieval. In *Advances in XML Information Retrieval: INEX2006,* Springer (LNCS #4518), 2007, pp 318–330..

[9] Sakai, T et al, Overview of the NTCIR-7 ACLIA IR4QA Subtask, in *Proceedings of Seventh NTCIR Workshop, Tokyo,* Dec. 2008 (this volume).