Using Google Translation in Cross-Lingual Information Retrieval*

He Xiaoning[1] , Wang Peidong[1], Qi Haoliang[2], Yang Muyun[3], Lei Guohua[2], Xue Yong[2]

1 Harbin University of Science and Technology
2 Heilongjiang Institute of Technology
3 Harbin Institute of Technology

Abstract

HIT2 Lab participated in NTCIR 7 IR4QA task. In this task many topics consist of name entities, so Google translation was used to translate query terms because of its high performance on name entity translation. We use KL-divergence model to perform retrieval and Chinese character bigram as our indexing unit. Pseudo feedback was used trying to improve average precision. We achieved competitive results in the task.

Keywords: *CLIR, term translation, KL-divergence*

1. Introduction

This is the first time that HIT^{2}Lab takes part in NTCIR. HIT^{2}Lab is the joint lab between Harbin Institute of Technology and Heilongjiang Institute of Technology. This paper reports our method used in the subtask IR4QA of NTCIR 7. In IR4QA, the goal is to evaluate how good an IR system is at returning documents that are relevant to the information needs on average, given a set of natural language questions or question analysis results. The main evaluation metric is Mean Average Precision (MAP).

Term translation is vital to cross-lingual information retrieval. We used Google translation to translate NTCIR 7 CLIR task queries, and made mono-lingual IR on it.

Relevance feedback is considered as pseudo (or blind) relevance feedback when there is an assumption that the top documents retrieved have a higher precision and that their terms represent the subject expected to be retrieved [4]. In other words, it is assumed that the documents on the top of the retrieval list are relevant to the query, and information from these documents is extracted to generate a new retrieval set. We compared results with pseudo-feedback results.

2. CROSS-LINGUAL IR BASED ON GOOGLE TRANSLATION

2.1. Term Translation Base on Google Translation

Google translation is from Google API family. It offers translation service based on Google translation technology. Terms in NTCIR topics are mostly name entities, which needs much human efforts to make translations accurate, industrial products have the availability to use very much human work, and that’s one reason that Google translation may work well on NTCIR topics. We used Google translation API to translate CLIR queries in terms of sentence. For example, the query “Users want to know what is Moore's Law.” is translated as “用户想知道什么是摩尔定律”, and it’s the same as translated query in topics file. “List events related to the Centenary Celebration of Peking University” is translated as “名单事件有关的百年庆祝北京大学”, the words: events, related, centenary, celebration, pecking university are also correctly translated. Only the word “list” is translated as a noun, as the anonym of “table”. Accurate translation can greatly guarantee high average precision.

2.2. Retrieval Model.

For retrieval model, we chose KL-divergence, KL-divergence is a widely used language model for information retrieval[6], and has good effect on IR task. Given two probability mass functions p(x) and q(x), The KL-divergence between p and q is defined as:

$$D(p \parallel q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

And in information retrieval, we can model the risk of returning a document d as relevant to a query q by KL-divergence between their respective language models[2]:

$$R(d;q) = KL(M_d \parallel M_q) = \sum_t P(t \mid M_q) \log \frac{P(t \mid M_q)}{P(t \mid M_d)}$$

[*] This work is done at HIT^{2}Lab, Heilongjiang Institute of Technology.

The research is supported by the Key Project of the National Science Foundation of China (60736044), National Science Foundation of China (60873105), the Science and Technology Break-through Project of Heilongjiang (GZ07A108), the Project of the Natural Science Foundation of Heilongjiang Province (F2007-14).

where d is a document, q is a query. M_d and M_d are the language models for documents and queries language model respectively. $P(t|M_q)$ and $P(t|M_d)$ are the probabilities that term t appears in M_q and M_d.

Jelinek-Mercer is widely used in information retrieval as smoothing method. It uses linear combination of doc language model with background language model. We choose Jelinek-Mercer as our smoothing method, which is as follows:

$$\hat{P}(t \mid d) = \lambda \hat{P}_{mle}(t \mid M_d) + (1-\lambda)\hat{P}_{mle}(t \mid M_q)$$

where d represents document, t represents term, and M_d and M_d are the language models for documents and queries language model respectively.

2.3. Indexing Unit

The common problem in Chinese, Japanese and Korean processing is the lack of natural word boundaries. Even though some spaces are added in Korean sentences, they often separate a sentence into phrases instead of words. For all these languages, we have to determine the indexing units by an additional process – either using word segmentation or by cutting the sentence into n-grams[5]. The latter is a simple method that does not require any linguistic resource. Unit indexing using Chinese Character bigram has been proved effective for Chinese IR research. We indexed documents by Chinese character bigram, every adjacent two Chinese characters form an index unit, for example, English words “nice to see you” can generate these bigram units: nice to, to see, see you. And for Chinese string “什么是摩尔定律” (meaning: What is Moore's Law), it will generate the following units: 什么(what), 么是, 是摩, 摩尔(more), 尔定, 定律(law) . Characters separated by space or punctual don’t form a unit.

With bigrams most of the correct Chinese words in a piece of text will be generated and they are much more specific in meaning than single characters. The drawback is that many meaningless character-pairs would also be produced and they could lead to noisy matchings between queries and documents, adversely impacting precision. With the number of commonly used single characters being over 6700, there could in theory be 40 million or more bigrams. Even 2.5 percent of this would result in a million pairs, potentially much larger than normal common English content terms for a similar size collection.

2.4. Pseudo Feedback.

Pseudo feedback, also known as blind relevance feedback, provides a method for automatic local analysis. It automates the manual part of RF, so that the user gets improved retrieval performance without an extended interaction. The method is to do normal retrieval to find an initial set of most relevant documents, to then assume that the top k ranked documents are relevant, and finally to do RF as before under this assumption.. Indri also provides pseudo feedback, according to [4], the feedback model is as follows:

$$P(c \mid R) \approx \frac{P(c, q_1 \ldots q_k)}{P(q_1 \ldots q_k)}$$

c could be any possible representation concept. P(c, q_1...q_k) is calculated as follows:

$$P(c, q_1 \ldots q_k) = \sum_D P(c \mid D) P(q_1 \ldots q_k \mid D) P(D)$$

3. EXPERIMENTS AND RESULTS

The experiments are run on two corpora Xinhua and Lianhezaobao, which are from Xinhua News of People’s Republic of China and Lianhezaobao newspaper of Singapore. Xinhua is simplified Chinese style and Lianhezaobao is traditional Chinese style. We used KL-divergence model, and JM smoothing method, lambda set to 0.8, which has been proven effective in previous experiments.

Documents indexing and querying is processed by Indri 2.6[1], which is a toolkit aimed at making information retrieval research easier. It provides basic indexing functionalities and retrieval models, such as TF-IDF, Okapi and KL-divergence.

We have submitted 4 results, one has none pseudo-feedback, and the others are with pseudo feedback, for the 3 different run types. The results have been evaluated using 3 metrics: Mean Average Precision, Q-measure and Discounted Cumulative Gain[3]. The following table lists our results on CS topics. The 4 rows titled with HIT-EN-CS-* are our results, and the last one is average results.

Table1: Experimental Results

	Mean AP	Mean Q	Mean nDCG
HIT-EN-CS-01-DN	0.3948	0.4417	0.6435
HIT-EN-CS-02-T	0.3702	0.4182	0.6252
HIT-EN-CS-02-D	0.3438	0.3883	0.6037
HIT-EN-CS-02-DN	0.3210	0.3657	0.5889
AVERAGE	0.3889	0.4236	0.6224

For the name of the column titles, HIT represents our lab, EN-CS means the task is to return simplified Chinese documents for English queries, 01 and 02 are priority parameters used for pooling, HIT-EN-CS-01-DN, HIT-EN-CS-02-T, HIT-EN-CS-02-D are pseudo feedback submissions, the only difference between HIT-EN-CS-02-DN and the other three submissions is HIT-EN-CS-02-DN has no pseudo feedback. T, D and DN are run-types. In T-run, only QUESTION field is used, in D-run, only NARRATIVE field is used, while in DN-run both QUESTION and NARRATIVE field is used.

References

[1] The Indri Search Engine. http://www.lemurproject.org/indri/

[2] Cheng-Xiang Zhai, John Lafferty. Model-based feedback in the KL-divergence retrieval model. In Tenth International Conference on Information and Knowledge Management. 2001.

[3] Tetsuya Sakai, Noriko Kando, Chuan-Jie Lin, Teruko Mitamura, Donghong Ji, Kuang-Hua Chen, Eric Nyberg. Overview of the NTCIR-7 ACLIA IR4QA Subtask. NTCIR 7. 2008.

[4] Victor Lavrenko, W. Bruce Croft. Relevance based language models. SIGIR 2001. 2001.

[5] Kui-Lam Kwok. Comparing representations in Chinese information retrieval, SIGIR 1997. 1997.

[6] John Lafferty, Chengxiang Zhai. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. SIGIR 2001. 2001.