HIT2Lab at CLEF: an Exploration to TEL Task*

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

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

Abstract

This is the first time HIT^{2}Lab has participated in CLEF. The Adhoc TEL task used a collection of electronic card catalog records from British Library, it is different from all the text retrieval tasks we've ever involved. In our runs, we used language modeling approach for retrieval model and incorporated pseudo-feedback, which have been proved effective in previous experiments. For bilingual tasks, we used Google translation service for query translation. We also adopted methods that have showed good effects in TEL 2009 task, including a stopword list provided by UniNE and partial fields indexing.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval

Key Words

Cross-lingual Information Retrieval, Query Translation, KL-divergence

1 Introduction

HIT^{2}Lab is the joint lab between Harbin Institute of Technology and Heilongjiang Institute of Technology. We are currently interested in several natural language processing tasks, including information retrieval, text classification, web data mining, and etc.

Since this is the first time we've taken part in TEL@CLEF task, we've focused on getting familiar with the task and implementing the methods which produced the best results in TEL@CLEF 2008. We took lemur toolkit to build up our retrieval platform [1]. For monolingual retrieval, we experimented language model and pseudo relevance feedback, for bilingual retrieval, we experimented query translation based on Google translation service.

2 Methodology

2.1 Indexing

TEL@CLEF 2009 involves the British Library collection, which is consisted of electronic card catalog records[2]. The dataset has several features, such as:

- The data is sparse, that is, each document has very few texts.
- A document is organized in several fields, for example, a <dc:title> field indicates the title of the record.
- The data is multilingual, only about half of the collection is in English.

It is convenient to treat the whole text in a document as unstructured text, however, some have noticed that removal of several useless fields can help to improve the retrieval performance[3][4]. For example, <mods:location> field includes a text version that indicates the name and location of the repository responsible for the stewardship of the resource and its content, such field is regardless of the document's content. During indexing process, we only kept such five fields: relation, tableOfContents, abstract, subject, and title, other fields were abandoned.

* 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).

We also used porter stemmer to normalize the terms. To eliminate the influence of very high frequent words, we used the stopword list released by UniNE last year.

2.2 Retrieval Model

Language modeling approach has been used widely in information retrieval researches. KL-divergence is a language model for information retrieval [5] which makes language models from both the document and query, then measures the similarity between the estimated models. Given two probability mass functions p(x) and q(x), The KL-divergence between p and q is defined as:

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

Given the two models generated from document and query, we can model the risk of returning a document d as relevant to a query q by KL-divergence between their respective language models[6]:

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

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.

2.3 Pseudo Feedback

Pseudo feedback is a technique to improve information retrieval performance, it has proven to be an effective strategy for improving retrieval accuracy in all retrieval models. The method is to do normal retrieval to find an initial set of most relevant documents, then assume that the top k ranked documents are relevant, and finally to do relevance feedback as before under this assumption. The feedback model we used in this year's TEL task is as follows[7]:

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

where q_i is an arbitrary query term, and c could be any possible representation concept. $P(c, q_1 \ldots q_k)$ is calculated as follows:

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

2.4 Query Translation for Bilingual Retrieval

Cross-language information retrieval is a subfield of information retrieval dealing with retrieving information written in a language different from the language of the user's query. For example, a user may pose their query in Chinese but retrieve relevant documents written in English. One commonly used method is to translate the query into document language, called query translation. Online translation service based query translation has been proved to be an effective resource for cross-lingual information retrieval [8]. We submitted bilingual retrieval runs based on query translation. Chinese queries are translated into English by Google translation service and then a mono lingual retrieval on the target collection is performed.

Experimental results showed that most queries have been translated well. For example, the title of the first query is translated as “Arctic animals”, which is the same as the English version; and its description is translated as “Find the Arctic fauna species and of related documents.” while the original one is “Find documents about arctic fauna species”. However there were still incorrectly translated topics. the title and description of topic 41 are “Sailing for Beginners” and “Find publications suitable for beginners or non-experts that provide information on any kind of sailing or boating”, and after it is translated into Chinese manually and translated back into English by Google translation, the contents of the two fields became “Portal navigation” and “Suitable for beginners or find information about non-professionals of any type of sailing or rowing motion information publications”. Portal navigation and sailing for beginners are definitely two different concepts, and so it caused the average precision of this topic to be 0%.

3 Experiments and Results

We submitted both monolingual and bilingual runs. Since we are familiar with English collection retrieval tasks, all our runs used only the English version dataset which is called TEL corpus from Britain Library as mentioned above. 50 topics are released for monolingual retrieval subtask. Since we are the only one who asked for Chinese version topics in bilingual retrieval subtask, we were responsible for translating the topics into

Chinese. An undergraduate student who did not take part in the task took the charge of this job, and we required an associate professor from Harbin Institute of Technology to proof the translated topics.

For monolingual runs, we experimented with KL-divergence method and pseudo feedback. We compared partial fields indexing with full indexing, the effectiveness of stopword list provided by UniNE was also tested. For bilingual runs, we experimented with query translation based on Google translation. We also tried a corpus based method which takes a Chinese-English parallel corpus for converting Chinese topics to English version, however it did not show good performance on TEL@CLEF 2008 data, and we analyzed that the bad performance was due to the limit of the corpus we used. That is, the parallel corpus we used was too small and couldn't reveal the true performance of the method, so we did not use the method in this year's task.

Monolingual results without pseudo are listed below. All the experiments took KL-divergence as retrieval model. Run tag including "u" means running with UniNE stopword list, "p" means partial indexing, "t" means title field of the topic is used, "d" means description field is used, and "td" means title and description field are used.

Table 1. Monolingual results

Method	Average Precision	Binary Preference	R Precision
kl-t	0.2816	0.2984	0.3057
kl-d	0.1321	0.2154	0.1689
kl-td	0.2685	0.3062	0.3071
kl-u-t	0.2896	0.2975	0.3081
kl-u-td	0.3401	0.3457	0.3606
kl-p-t	0.2922	0.3012	0.3134
kl-p-td	0.3020	0.3264	0.3404
kl-u-p-t	0.3029	0.3001	0.3156
kl-u-p-td	0.3443	0.3476	0.3751

From the results we see that title combining with description can improve retrieval performance, especially after applying UniNE stopword list which effectively cut down meaningless terms in topics. Partial indexing also gets positive effect, and so suggests a potential research direction for the task.

Bilingual results are as follows.

Table 2. Bilingual results

Method	Average Precision	Binary Preference	R Precision
kl-u-p-t	0.2766	0.2767	0.2864
kl-u-p-td	0.3047	0.3115	0.3292

The retrieved documents set we submitted used pseudo feedback. The top 10 documents returned for each topic were considered to be relevant. We submitted 4 runs for monolingual subtask and 3 runs for bilingual subtask, the evaluation results of these runs are listed below. The meanings of "u", "p", "t", "d", "td" are as mentioned above, and "ft" means how many terms were contained in the expanded query generated from pseudo relevant documents.

Table 3. Submitted run results

Method	Subtask	Average Precision	Binary Preference	R Precision
kl-u-p-td	monolingual	0.3390	0.3420	0.3688
kl-u-p-td-ft20	monolingual	0.3924	0.3810	0.4037
kl-u-p-td-ft40	monolingual	0.3936	0.3808	0.4051
kl-u-p-t-ft40	monolingual	0.3503	0.3422	0.3618
kl-u-p-td-ft20	bilingual	0.3523	0.3415	0.3614
kl-u-p-td-ft40	bilingual	0.3527	0.3410	0.3613
kl-u-p-t-ft40	bilingual	0.3172	0.3192	0.3335

4 Conclusion

While classic language model coupled with pseudo feedback is suitable for the TEL task, particular features of the TEL corpus can also benefit to the retrieval performance. Different fields in TEL collection have different affect, and removal of some fields that have no relation with the document content can help to improve retrieval performance. Combining title and description field in a topic also works, especially after applying a stopword list.

Query translation based on online translation service works well for cross-lingual information retrieval. Online translation service is able to translate the most queries correctly, although there are a small portion of the queries be translated as absolutely different meaning from their real meaning.

There are still many issues to consider, including document fields weighting, the collection's multilinguality and the document contents' insufficiency. We will explore such issues in future work.

References

[1] The Lemur Toolkit for Language Modeling and Information Retrieval. http://www.lemurproject.org/

[2] Eneko Agirre1, Giorgio M. Di Nunzio, Nicola Ferro, Thomas Mandl, and Carol Peters. CLEF 2008: Ad Hoc Track Overview. 9th workshop of the cross-language evaluation forum. 2008.

[3] Ljiljana Dolamic, Claire Fautsch, Jacques Savoy: UniNE at CLEF2008: TEL, Perisan and Robust. 9th workshop of the cross-language evaluation forum. 2008

[4] Paul McNamee. JHU Ad Hoc Experiments at CLEF 2008. 9th workshop of the cross-language evaluation forum. 2008

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

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

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

[8] Jens K ursten, Thomas Wilhelm and Maximilian Eibl. CLEF 2008 Ad-Hoc Track: On-line Processing Experiments with Xtrieval. 9th workshop of the cross-language evaluation forum. 2008