中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

The ICT System Description for NTCIR-7

Weihua Luo[1,2], Tian Xia[1,2], Ji Guo[2,3] and Qun Liu[2]
[1] Graduate University of Chinese Academy of Sciences
[2] Key Laboratory of Intelligent Information Processing,
Institute of Computing Technology (ICT), Chinese Academy of Sciences
[3] School of Software and Microelectronics, Peking University
{luoweihua, xiatian, guoji, liuqun}@ict.ac.cn

1. System

The ICT system for NTCIR-7, ICT-Crossn, consists of two components.

1.1 Architecture

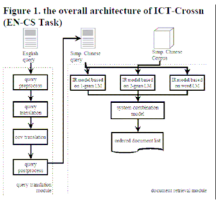
ICT-Crossn consists of a query translation component and a document retrieval component. When performing a CLIR task, e.g. reading English queries and searching in Chinese corpus, ICT-Crossn firstly translates the English queries into Chinese and searches for the relevant documents in the Chinese corpus.

In the translation process, we use the output of phrase-based statistical machine translation and accomplish the translation of OOV (out-of-vocabulary) words with search engines.

In the retrieval process, we combine the results generated by multiple IR models.

Figure 1 shows the overall architecture of ICT-Crossn.

Figure 1. the overall architecture of ICT-Crossn (EN-CS Task)



1.2 Query Translation

Our work focuses on phrase translation because we think words in sentences are ambiguous but phrases are more clearly defined. Translating multiple words as phrase introduces much less candidates . For IR4QA, the critical information in the question is often phrases. So it's crucial to recognize a phrase and to translate them correctly.

In the test set of the dry run and the formal run, it is easily observed that most questions follow a few fixed structures. Roughly, four types of questions are defined in four kinds of structures. Therefore, we predefine some structure templates to capture the constituents we need in the questions. For compound phrases, which consist of shorter phrases and words but have no translations in the phrase table, we divided them into "basic" ones with the help of phrase tables. These "basic" ones make up the whole constituent and have the minimal joining times.

For example, a template we defined is "*what is X*" and X is a constituent we want. For the question "*What is human genetic sequence determination*", we can extract "*human genetic sequence determination*" but it's compound phrase. And the we find 6 sub-phrases in the phrase table, i.e. "*human*", "*genetic*", "*sequence*", "*determination*", "*human genetic sequence*" and "*genetic sequence*". However, "*human genetic sequence*" and "*determination*" make up the constituent and have the minimal joining times. So the combination is what we want.

In order to provide more relevant key words, the system is designed to output multiple candidates. We rank the candidates by $p(c|e)$, the probability to translate an English phrase e to a Chinese one c.

We translate the phrases and words not appearing in the phrase table based on search engines. We construct the English query based on the OOV words and submit it to the search engine for Chinese web pages. We extract the possible translations within a predefined windows from the snippets returned from the search engine. A linear feature function is used to determine the translation of the English key words.

$$score(c,e) = \sum_i \omega_i f_i(c,e)$$

Three feature is adopted. The first one is whether the candidates appears directly behind the key words within a pair of parenthesis. The next feature is the similarity of the transliteration of a candidate and the keyword. The last is the co-occurrence probability of the key words and the Chinese candidates.

1.3 Document Retrieval

Retrieval model and index model are two categories of models in information retrieval. We think results of various models are complementary for each other. Proper combination methods would bring better results.

In retrieval, motivated by the idea of system combination in statistical machine translation, we design a combination model for a robust and better IR performance. The final score of a document for a query is a summed interpolation value of different models as follows:

$$score(d,q) = \sum_j \alpha_j score_j(d,q)$$

In ICT-Crossn, we prefer index model to construct our combination model.

To tune the parameters, we define the object function as follows:

$$\delta(\bigcup_{i=1}^{n} Model_i) = MAP(\bigcup_{i=1}^{n} Model_i)$$

The goal is to find the parameter setting maximizing the MAP score of the combination model. We adopt a greedy search algorithm to tune the parameter setting to achieve the locally optimal performance.

2. Data

Besides the data provided by the organizer, we used the following additional data:

(1) LDC corpus for phrase table:

LDC2002E18 LDC2003E07 LDC2003E14 LDC2004E12

LDC2004T07 LDC2004T08 LDC2005T10 LDC2005T06

The corpus contains about 153.5M Chinese words and 168.1M English words.

Table 1. Development data set in NTCIR-7

Task	Question Type	Question number
EN-CS	Event	14
	Relationship	10
	Biography	21
	Definition	33
CT-CT	Event	8
	Relationship	10
	Biography	24
	Definition	29

(2) Development data set

We construct our development set both for EN-CS and CT-CT tasks based on the dry run set. We download the dry run set and the answers through the EPAN interface. For each question, documents that contains an answer are marked as relevant. Table 1 gives a detail illustration about our development data set.

3. Experiments on dry run set

We choose language model with Dirrichlet smoothing and feedback as our primary retrieval model. Unigram, bigram and word indices are created separately. For each index model, tune the feedback document number and feedback term count to achieve its best performance. Then index model combination is conducted to obtain our final model. Figure 2 shows the combination results of the index models for EN-CS tasks on the dry-run data set. The results of index model combination for CT-CT task are illustrated in Figure 3.

Figure 2. Results of index models combination in EN-CS task for dry run

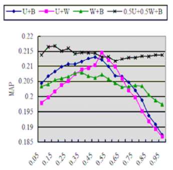
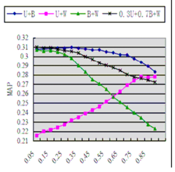


Figure 3. Results of index model combination in CT-CT task for dry run



4. Results on NTCIR-7 evaluation

We submit 5 runs for EN-CS cross-lingual IR task and 4 runs for CT-CT monolingual IR task. They are listed as following:

(1) MITEL-EN-CS-01-T: our primary system for EN-CS task. It uses the question part. The combination scheme is 0.5U+0.5W+0.15B, which outperform others in the dry run topics.

(2) MITEL-EN-CS-02-T: the combination scheme is 0.5U+0.5W, with the question part.

(3) MITEL-EN-CS-03-T: the combination scheme is 0.45U+0.55B, using the question part.

(4) MITEL-EN-CS-04-D: the combination scheme is the same as primary system, but only uses the narrative part of each topic.

(5) MITEL-EN-CS-05-TD: the same as primary system, except using both question and narrative part of each topic.

(6) MITEL-CT-CT-01-T: our primary system for CT-CT task, which is combined in the scheme of 0.3U+0.7B, with the question part of each topic.

(7) MITEL-CT-CT-02-T: the combination scheme is 0.3U+0.7B+0.05W, with the question part. We use it as our secondary system considering the word index's poor performance.

(8) MITEL-CT-CT-03-D: the same as primary system except using the narrative part.

(9) MITEL-CT-CT-04-T: only use the bigram index unit and the question part.

Table 2 shows the performance of each run of ICT-Crossn on the IR4QA subtasks in the NTCIR-7 formal runs. Our runs perform well in both subtasks. In the EN-CS CLIR subtasks, MITEL-EN-CS-03-T works well even compared to the monolingual runs. In CT-CT tasks, our runs rank top 4.

Table 2. Performance based on real qrels in NTCIR-7 formal fun

Task	Run id	Mean AP	Mean Q	Mean nDCG
EN-CS	MITEL-EN-CS-01-T	0.5849	0.6005	0.7949
	MITEL-EN-CS-02-T	0.5693	0.5858	0.7847
	MITEL-EN-CS-03-T	**0.5959**	**0.6124**	0.7947
	MITEL-EN-CS-04-D	0.5789	0.5950	0.7907
	MITEL-EN-CS-05-TD	0.5898	0.6058	**0.8003**
CT-CT	MITEL-CT-CT-01-T	0.5791	0.5963	0.7835
	MITEL-CT-CT-02-T	**0.5839**	**0.6018**	**0.7873**
	MITEL-CT-CT-03-D	**0.5839**	0.6013	0.7869
	MITEL-CT-CT-04-T	0.5645	0.5783	0.7648

5. Conclusion

In this paper, we give a brief introduction to our system in NTCIR-7 evaluation. We report the resources used, development data set construction, and results achieved on both the development set and the final test sets. ICT-Crossn is based on the traditional CLIR framework. First, the system translates the questions into the target language of the corpus with a phrase table generated by a SMT system and an OOV word translation module based on search engines. Second, the system performs information retrieval with different index units, and combines the document lists into the final result. To obtain the optimal models and the parameters, we construct the development set based on the dry run set. The results on the official set show our system achieves a good performance.

MITEL
中国科学院计算技术研究所
多语言交互技术实验室
MITEL
Multilingual Interaction Technology Lab