ICE-TEA: an Interactive Cross-language Search Engine with Translation Enhancement

Dan Wu
School of Information Management
Wuhan University, Hubei 430072, China
woodan.cn@gmail.com

Daqing He
School of Information Sciences
University of Pittsburgh, PA 15260, USA
dah44@pitt.edu

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Relevance Feedback*

General Terms: Design, Experimentation

Keywords: CLIR, Translation Enhancement (TE), Query Expansion (QE), Relevance Feedback (RF), ICE-TEA

EXTENDED ABSTRACT

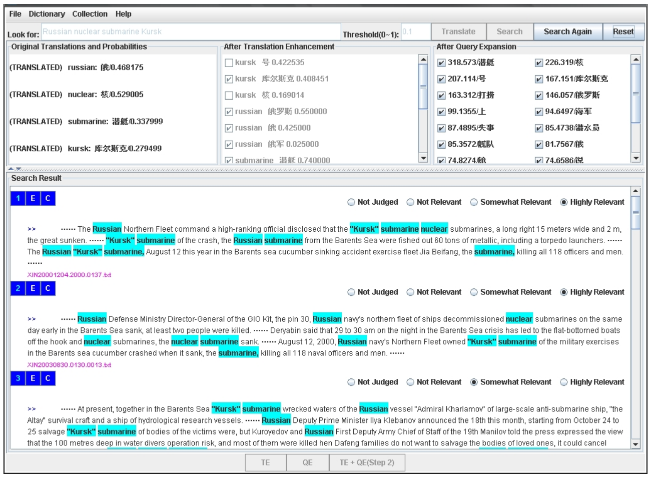
In cross-language information retrieval (CLIR), relevance feedback (RF) has been demonstrated to be effective in improving retrieval results, especially when reliable RF information can be obtained from users. Though query expansion (QE) is the leading RF approach in CLIR and it can take place before or/and after translating the query, it should not be the only possible RF method. In our demonstration, besides an implementation of post-translation QE, we also implement a novel RF approach called translation enhancement (TE) and the integration of TE and QE.

ICE-TEA is an interactive multilingual information access system with complete relevance feedback techniques. It is developed to utilize the RF information from users to conduct TE, QE, and the integration of them to improve CLIR results. ICE-TEA performs query translation based CLIR. It is built on top of Indri, a leading open-source monolingual search engine. As illustrated in Figure 1, the followings are the main functions of ICE-TEA:

Translation Enhancement. As discussed in iCLEF, and as demonstrated by Google's cross-language search engine, returned documents in CLIR need to be translated back to the query language side so that they can be selected and examined by the users. The main idea of TE, therefore, is to extract intended translation relationships of query terms from the relevant document pairs (original returned documents and their translations) and then to enhance the query translation. Through a set of experiments, we design a TE method that acquires translation relationships in relevant document pairs by word alignment obtained using GIZA++. Such translation relationships are used to re-calculate the corresponding translation probabilities of the query terms, so that the translated query obtained from the new translation probabilities will be closer to the true meaning of the user's request.

Query Expansion. In current version of ICE-TEA, we implement post-translation QE, which has been proved to perform better than pre-translation QE. Our system utilizes Indri's QE mechanism which is an adaptation of Lavrenko's relevance models. More interestingly, since QE and TE occur at different stages of the CLIR process, ICE-TEA also combines both QE and TE.

User-assisted Query Translation with Relevance Judgment. Following the user-assisted query translation idea [1], ICE-TEA invites users into CLIR process. It lets users type queries and set cumulative probability threshold (CPT) so that initial CLIR search can be initiated. It also displays the original returned documents and their translations (both surrogate and full-text) to help users to make multi-level relevance judgments. After RF, it displays both TE and QE results, and helps users to modify easily the query translations by selecting or deselecting translation alternatives and their probabilities.



Look for: Russian nuclear submarine Kursk

Threshold(0~1): 0.1

Original Translations and Probabilities

(TRANSLATED) russian: 俄/0.468175

(TRANSLATED) nuclear: 核/0.529005

(TRANSLATED) submarine: 潜艇/0.337999

(TRANSLATED) kursk: 库尔斯克/0.279499

After Translation Enhancement

□ kursk 号 0.422535
☑ kursk 库尔斯克 0.408451
□ kursk 核 0.169014
☑ russian 俄罗斯 0.550000
☑ russian 俄 0.425000
☑ russian 俄军 0.025000
☑ submarine 潜艇 0.740000

After Query Expansion

☑ 318.573/潜艇	☑ 226.319/核
☑ 207.114/号	☑ 167.151/库尔斯克
☑ 163.312/打捞	☑ 146.057/俄罗斯
☑ 99.1355/上	☑ 94.6497/海军
☑ 87.4895/失事	☑ 85.4738/潜水员
☑ 85.3572/舰队	☑ 81.7567/俄
☑ 74.8274/舱	☑ 74.6586/说

Search Result

1 E C   ○ Not Judged ○ Not Relevant ○ Somewhat Relevant ● Highly Relevant

>> ······ The Russian Northern Fleet command a high-ranking official disclosed that the "Kursk" submarine nuclear submarines, a long right 15 meters wide and 2 m, the great sunken. ······ "Kursk" submarine of the crash, the Russian submarine from the Barents Sea were fished out 60 tons of metallic, including a torpedo launchers. ······ The Russian "Kursk" submarine, August 12 this year in the Barents sea cucumber sinking accident exercise fleet Jia Beifang, the submarine, killing all 118 officers and men. ······

XIN20001204.2000.0137.txt

2 E C   ○ Not Judged ○ Not Relevant ○ Somewhat Relevant ● Highly Relevant

>> ······ Russian Defense Ministry Director-General of the GIO Kit, the pin 30, Russian navy's northern fleet of ships decommissioned nuclear submarines on the same day early in the Barents Sea sank, at least two people were killed. ······ Deryabin said that 29 to 30 am on the night in the Barents Sea crisis has led to the flat-bottomed boats off the hook and nuclear submarines, the nuclear submarine sank. ······ August 12, 2000, Russian navy's Northern Fleet owned "Kursk" submarine of the military exercises in the Barents sea cucumber crashed when it sank, the submarine, killing all 118 naval officers and men. ······

XIN20030830.0130.0013.txt

3 E C   ○ Not Judged ○ Not Relevant ● Somewhat Relevant ○ Highly Relevant

>> ······ At present, together in the Barents Sea "Kursk" submarine wrecked waters of the Russian vessel "Admiral Kharlamov" of large-scale anti-submarine ship, "the Altay" survival craft and a ship of hydrological research vessels. ······ Russian Deputy Prime Minister Ilya Klebanov announced the 18th this month, starting from October 24 to 25 salvage "Kursk" submarine of bodies of the victims were, but Kuroyedov and Russian First Deputy Army Chief of Staff of the 19th Manilov told the press expressed the view that the 100 metres deep in water divers operation risk, and most of them were killed hen Dafeng families do not want to salvage the bodies of loved ones, it could cancel

Figure 1: ICE-TEA's English-to-Chinese Search Interface

ICE-TEA right now is implemented to conduct CLIR search between English queries and Chinese collections. However, the system's architecture is generic enough to be converted for any major language pairs. The basic required language resource is a bilingual dictionary. However, with a machine translation system or the ability to access a machine translation service between the document and the query language, the translated returned document will be in better quality, so does the TE results.

REFERENCE

[1] He, D., Oard, D. W., Wang, J., et al. Making MIRACLEs: Interactive Translingual Search for Cebuano and Hindi. *ACM Transactions on Asian Language Information Processing*. 2003, 2(3): 219-244.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.