Cross-Lingual Information Retrieve in Sogou Search

JingFang Xu
Sohu.com internet plaza, No.1 Park, Zhongguancun East Road, Haidian District, Beijing 100084, P.R.China
xujingfang@sogou-inc.com

Feifei Zhai
Sohu.com internet plaza, No.1 Park, Zhongguancun East Road, Haidian District, Beijing 100084, P.R.China
zhaifeifei@sogou-inc.com

Zhengshan Xue
Sohu.com internet plaza, No.1 Park, Zhongguancun East Road, Haidian District, Beijing 100084, P.R.China
xuezhengshan@sogou-inc.com

ABSTRACT

In recent years, more and more Chinese people desires to be able to access the large amount of foreign language information and understand what is happening all over the world. However, language barrier is always a problem to them. In order to break the language barrier and connect Chinese people to the foreign language information in the world, Sogou has built a cross-lingual information retrieval (CLIR) system named *Sogou English* (http://english.sogou.com), which enables Chinese people to search and browse foreign language information with Chinese.

In *Sogou English*, when the user inputs a Chinese query, it will first translate the Chinese query into English, and then search over the Internet, and finally translate the search results into Chinese so that users can understand them better. Hence with *Sogou English*, people can read and browse the information from English world without actually knowing English.

Sogou English is built based on the second largest search engine in China, *Sogou Search*. Besides, the neural machine translation (NMT) technology is adopted to do the translation part. As far as we know, *Sogou English* is the first CLIR system in the world that integrates NMT into search engine. NMT has made a significant progress in recent years, and by leveraging the advantages of NMT, *Sogou English* is able to produce high-quality translation outputs, especially for the search results to help people understand them. However, we are still facing many technical challenges arising from integrating the translation system into search engine.

One challenge is the ambiguity originates from the polysemy phenomenon in target language. In query translation process, when an input query is translated into a target language word that has several different meanings, search engine does not know which meaning the user wants to utilize for searching.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.
SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan
© 2017 Copyright is held by the owner/author(s).
ACM ISBN 978-1-4503-5022-8/17/08.
http://dx.doi.org/10.1145/3077136.3096463

For example, if a user wants to find some information about the clothing brand "Ports", and he inputs Chinese term "宝姿 (baozi)" for searching. The translation system will translate "宝姿 (baozi)" into "Ports", which is correct. However, when "Ports" serves as the English query, search engine doesn't know what exactly the user needs, the clothing brand, the term used for communication in computer networking, or a harbour. As a result, it will consider all the possible search intentions, and return many results that are actually not what the user wants. Therefore, how to keep the original meaning of source language query and let it supervise the search process is one problem we need to solve.

Another challenge is the translation ambiguity when we translate the search results into Chinese. Currently, same as the traditional machine translation setting, different sentences are translated separately. Therefore, the same source language phrase in different search results might be translated into different target phrases (due to the different context), or even wrong translations (lack of context). For example, suppose now we have got some search results about movie "Doctor Strange (Its Chinese translation is 奇异博士)". One search result contains a phrase "IMDB score", indicating the translation system that it is about a movie. By this context, the translation system is able to follow the word distribution in movie domain implicitly, and correctly translate "Doctor Strange" into "奇异博士". However, for the results lacking that kind of information, it is possible that the translation system prefers to follow the general word distribution from its training data, and wrongly translate "Doctor Strange" into "奇怪医生 (means weird medical doctor)", or "陌生医生 (means strange medical doctor)". Thus, how to get more accurate and robust translation by using the context from all search results is also a big challenge to us.

In this talk, we will give a brief introduction to our efforts on building our CLSI system, *Sogou English*: the architecture, the challenges we face and our current solutions.