RICeSTQUTranslate: A suggestion oriented machine translation system for English-Persian cross-lingual information retrieval in medical domain

Amin Rahmani
Computational linguistics department
MSc candidate
RICeST
shiraz, Iran

Mohammad Reza Falahati
Computational linguistics department
Associate professor
RICeST
shiraz, Iran

Mohammd Bagher Dastgheib
Computational linguistics department
Associate professor
RICeST
shiraz, Iran

Abstract

Recently, due to the fast information growth in cyberspace, information retrieval (IR) is not only limited to English language, but it is extended to other languages. The need to gain information in different languages, motivated researchers to develop cross-lingual information retrieval (CLIR) systems. The most important reason to develop such systems is language barrier problem. Since establishment, officials of the RICeST, have created several English and Persian databases. This international center meets the information needs of Iranian as well as foreign users. Due to the language barrier problem, foreign users are not able to use RICeST's Persian databases; consequently, the aim of doing this research was to develop a CLIR system, called "RICeSTQUTranslste", to help foreign users to use RICeST's Persian databases. To do this, 250 English paper titles were used to develop a query suggestion system. as well, 15000 Persian paper titles and their abstracts were utilized to develop a machine translation system. RICeSTQUTranslate has three main components. The first part is the query suggestion system which enables users to select the proposed alternatives. In this system, a new query re-ranker algorithm is suggested to prioritize the alternatives. The second part called, query translation, translates and searches users' queries. This part is designed based on a novel translation algorithm. The last part, called text translation, translates the titles of the retrieved documents from Persian to English in order to help the users to gain related information. RICeSTQUTranslate was evaluated based on the TREC medical test-set and two conventional methods, glass-box and black-box evaluation. As glass-box method, machine translation was evaluated according to the BLEU (0.455). As black-box evaluation, RICeSTQUTranslate was tested based on the MAP score. According to the results, CLIR system and monolingual systems gained 0.41 and 0.54 respectively. As conclusion, performance of the CLIR system was 75% identical to that of monolingual system. Consequently, the suggested approach was suited the English-Persian CLIR. due to the simplicity of the proposed translation algorithm, regarding the hardware resource usage and algorithm order, it could be replaced with other algorithms such as statistical ones in CLIR process.

Keywords: Colmeauer, IR, CLIR, machine translation, MAP score, BLEU, Black-Box, Glass-Box