Cross-Lingual Information Retrieval System for Indian Languages

Jagadeesh Jagarlamudi and A Kumaran

Abstract

This paper describes our first participation in the Indian language sub-task of the main Adhoc monolingual and bilingual track in CLEF competition. In this track, the task is to retrieve relevant documents from an English corpus in response to a query expressed in different Indian languages including Hindi, Tamil, Telugu, Bengali and Marathi. Groups participating in this track are required to submit a English to English monolingual run and a Hindi to English bilingual run with optional runs in rest of the languages. We had submitted a monolingual English run and a Hindi to English cross-lingual run.

We used a word alignment table that was learnt by a Statistical Machine Translation (SMT) system trained on aligned parallel sentences, to map a query in source language into an equivalent query in the language of the target document collection. The relevant documents are then retrieved using a Language Modeling based retrieval algorithm. On CLEF 2007 data set, our official cross-lingual performance was 54.4% of the monolingual performance and in the post submission experiments we found that it can be significantly improved up to 73.4%.