Translation Events in Cross-Language Information Retrieval

Anne R. Diekema
School of Information Studies
Syracuse University
diekemar@ syr.edu

Cross-Language Information Retrieval (CLIR) systems enable users to formulate queries in their native language to retrieve documents in foreign languages. Because queries and documents in CLIR do not necessarily share the same language, translation is needed before matching can take place. This translation step tends to cause a reduction in the retrieval performance of CLIR as compared to monolingual information retrieval. To understand the events occurring in cross-language retrieval query translation and the relation of these events to retrieval performance, the study explored the following research questions:

1) *What kinds of translation events affect cross-language retrieval?*
2) *In what way does the presence of certain translation events in query translation affect retrieval performance?*

In answer to the first research question, a detailed taxonomy was developed which incorporated possible translation events from a query analysis and a literature review. The translation event taxonomy included three main translation event categories in addition to a translation correctness assessment.

To answer the second research question, a large number of English target queries (translated from Dutch into English, thus representing cross-lingual retrieval) was coded using the translation event taxonomy. In a retrieval experiment the performance of the English source queries (representing monolingual retrieval) was compared to that of the English target queries to assess the impact of the translation events on retrieval performance. A multiple regression analysis was carried out to see which translation events had an impact on retrieval performance. It was found that four of the six independent variables contributed significantly to prediction of the difference in retrieval performance when comparing monolingual to cross-lingual information retrieval: *missing specialized vocabulary*, *missing general terms*, *wrong translation due to ambiguity*, and *correct identical translation*. Although the contribution of each of these variables is significant, their contribution to the total variance of the independent variable is small (adjusted $R^2 = .223$).

A traditional query analysis showed that more than half of the queries showed a performance difference between monolingual and cross-lingual retrieval of over 70%. The queries with less than 10% performance difference were shorter and less ambiguous than other queries in this study. The results also suggest that not all query terms are created equal and that some terms appear to be more important from a retrieval perspective than others.