

Research Statement

Wei Gao (wgao@se.cuhk.edu.hk)

Department of Systems Engineering and Engineering Management

The Chinese University of Hong Kong

Revised on September 21 2009

Background

World Wide Web has become a huge repository of texts in many languages as everybody can post content in any language. Due to the limit of knowledge across languages, however, the infrastructure and functionality underlying the modern Web information systems is essentially monoglot, meaning that information can be processed effectively merely inside each single language. As a result, it cannot satisfy the growing interests of users for cross-lingual and/or multilingual information access.

In spite of the long development of information processing across languages, such as machine translation, cross-language and multilingual information retrieval (CLIR/MLIR), and the recent advancement in cross-lingual text mining, the effectiveness of cross-lingual information access is still far to meet the requirement of daily use and application. Simply put, the main bottleneck lies not only in the complexity of human languages themselves but also in that of their interactions. It is therefore challenging for cross-lingual information processing.

The goal of my research is to develop effective techniques for helping bridge the gap in information access across languages. The problems falling into this category include cross-lingual (or multilingual) text mining and retrieval, machine translation, named entity translation and transliteration, and many other human language technologies based on translation knowledge and tools. In aligning with this goal, I worked at the intersection of information retrieval and natural language processing. I developed various learning-based methods to improve the effectiveness of cross-lingual information access at different levels. At the level of query processing, for example, I collaborated with colleagues developing algorithms that learn to suggest related queries in a different language by mining search engine query logs to improve CLIR effectiveness in Web search; meanwhile, at the level of relevance ranking, we investigated learning to rank for leveraging document correlations across languages to enhance (monolingual, cross-lingual and multilingual) search result ranking.

Among others, I am also interested in information access under Web 2.0 environment, especially knowledge discovery by mining opinions and social networks.

Cross-language mining and retrieval

The heavy reliance on translation vocabulary leads cross-lingual applications fragile to out-of-vocabulary (OOV) problem during the course of translation. This is especially prominent in cross-language Web search where new terms frequently appear in the query but are missing in the translation vocabulary. Another predicament to the translation effectiveness is caused by the ambiguity — polysemy of word sense. Cross-language mining acts as an effective means to improve the effectiveness of query translation by complementing the coverage of bilingual dictionary [1]. The common approach resorts to text mining to extract the translation of OOV query terms from Web corpora such as search results, anchor texts, etc. In contrast to these traditional approaches, I am interested in developing novel methods by mining search engine query logs to suggest the relevant queries of the target language given the source query [2, 6]. The method can not only better tackle

OOV problem and translation ambiguity, but also provide effective means to “expand” the query and achieves better retrieval effectiveness than pseudo-relevance feedback in CLIR.

An interesting finding is that the users of search engines in the same period of time have similar interests, and it is quite often that they submit queries on similar topics in different languages. Statistically, the overlap of interests become even greater for the queries that are searched most frequently [3]. Based on this fact, useful techniques can be developed for helping users better specify their information needs across languages. For example, one can discover multilingual query clusters based on the content and clickthrough information from the logs in different languages to form the most typical formulation of queries, representing the common search interests across languages. Random walks on click graph and the similar documents can be leveraged to infer the relevance of queries. It is expected that the query substituted with the new set of formulations can cover more relevant documents, thus to improve the retrieval effectiveness in CLIR or MLIR.

Cross-lingual and multilingual rank learning

Relevance ranking is the core component of information retrieval. Much recent effort has been made on using machine learning approaches for ranking known as learning to rank. Based on this formalism, the innovative thrust of my work is using the diversity of search quality across languages to enhance document ranking for Web search. For certain portion of queries that are inherently bilingual or multilingual, i.e., the queries that occur in the query logs of different languages but represent the same information needs, the relevant information in different languages may be distributed asymmetrically across the corresponding sides. Intuitively, the discrepancy of search quality can be leveraged to improve the ranking performance for the “poor” side with the information from the “rich” side. For example, it can be expected that ranking search results of *The Duke of Zhou*, an ancient Chinese politician, is harder in English than it is in Chinese. Thus the Chinese information can be used to help ranking English documents for this query, underling which is the various monolingual and cross-lingual similarities among these retrieved documents. This idea were investigated in different Web search settings in our studies [3, 4, 5].

One of the future directions of this work is to identify more cross-lingual relationships among documents that are likely to affect the ranking effectiveness. The correlations were derived previously by following the links of clickthrough data in query logs and measured according to the similarity of text content. I attempt to incorporate other useful streams of information such as anchor texts or query texts of the clicked documents. With the click graph, it may be possible to reinforce the estimation of document relation and query relation by leveraging iterative random walk model or spectral theorem based on the content-based similarities.

One may also try out *transfer learning* or *domain adaptation* to cross-lingual rank learning. How to transfer or adapt the knowledge of a ranker learned monolingually to the search setting of a different language is obviously an interesting problem. For instance, one may identify more sophisticated monolingual features that do not transfer cross-lingually but are inherently asymmetric for either side, such as document classification features built from domain taxonomies. The *Open Directory Project* (DMOZ) provides the most comprehensive human-reviewed directory hierarchy for classification, but mostly the hierarchy is much more comprehensive in a few dominant languages, which can be used for this purpose. The asymmetric knowledge base *Wikipedia* also can be used as one of the resources.

Mining opinions and social networks

The quick emergence of Web 2.0 is envisaged by the unprecedented freedom of interaction and participation in creating, expressing and changing Web content by individual users. Compared to the traditionally static non-interactive Web sites, this new form of media is characterized as diverse, informal and unplanned user connectivity and expression of personal opinion. I am interested in research topics on mining social relationship and opinions from Web documents, blogs and social networks based on Web search and information extraction technology to facilitate decision making in financial or public sectors. Due to these unique characteristics, the typical challenge is to develop adaptive retrieval and natural language processing techniques for this inimitable arena. One needs to design algorithms for detecting and learning relationship between people or organizations through websites and social networks, to develop high precision search tools to find and extract this specific information in the Internet before carrying out the learning, and to learn the behavior profiles of individuals that are parts of events and activities in the network. In addition, to deal with the anomalous expressions like the way of spoken language, radical transfer of human language technologies is required to make them adaptable to the linguistic styles in this dynamic and diverse environment. The research may need a wide spectrum of approaches and formalisms, such as kernel methods, supervised/semi-supervised learning, transfer learning, use of ontological knowledge and reasoning using world knowledge, and their application on retrieval, extraction and mining natural language texts.

References

- [1] Wei Gao and Cheng Niu. “Cross-Language Mining and Retrieval”. In Ling Liu and T. Omur (Eds.) *Encyclopedia of Database Systems*, Springer-Verlag, 2009.
- [2] Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Kam-Fai Wong, and Hsiao-Wuen Hon. “Exploiting Query Logs for Cross-Lingual Query Suggestion”. *ACM Transactions on Information Systems (TOIS)* (To appear).
- [3] Wei Gao, John Blitzer, Ming Zhou, and Kam-Fai Wong. “Exploiting Bilingual Information to Improve Web Search”. In *Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics (ACL09)*, Singapore, August 2009.
- [4] Wei Gao, Cheng Niu, Ming Zhou, and Kam-Fai Wong. “Joint Ranking for Multilingual Web Search”. In *Proceedings of the 31st European Conference on Information Retrieval Research (ECIR09)*, Toulouse, France, April 2009. LNCS-5478, pp. 114-125, Springer-Verlag.
- [5] Wei Gao, John Blitzer, and Ming Zhou. “Using English Information in Non-English Web Search”. In *Proceedings of the 2nd ACM International Workshop on Improving Non-English Web Searching (iNEWS’08)*, held in conjunction with *ACM 17th Conference on Information and Knowledge Management (CIKM08)*, pp.17-24, Napa Valley, California, USA, October 2008.
- [6] Wei Gao, Cheng Niu, Jian-Yun Nie, Ming Zhou, Jian Hu, Kam-Fai Wong, and Hsiao-Wuen Hon. “Cross-lingual Query Suggestion Using Query Logs of Different Languages”, In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR07)*, pp.463-470, Amsterdam, The Netherlands, July 2007.