MARS: Multilingual Access and Retrieval System with Enhanced Query Translation and Document Retrieval

Lianhau Lee, Aiti Aw, Thuy Vu, Sharifah Aljunied Mahani, Min Zhang, Haizhou Li
Institute for Infocomm Research
1 Fusionopolis Way, #21-01 Connexis, Singapore 138632
{lhlee, aaiti, tvu, smaljunied, mzhang, hli}
@i2r.a-star.edu.sg

Abstract

In this paper, we introduce a multilingual access and retrieval system with enhanced query translation and multilingual document retrieval, by mining bilingual terminologies and aligned document directly from the set of comparable corpora which are to be searched upon by users. By extracting bilingual terminologies and aligning bilingual documents with similar content prior to the search process provide more accurate translated terms for the in-domain data and support multilingual retrieval even without the use of translation tool during retrieval time. This system includes a user-friendly graphical user interface designed to provide navigation and retrieval of information in browse mode and search mode respectively.

1 Introduction

Query translation is an important step in the cross-language information retrieval (CLIR). Currently, most of the CLIR system relies on various kinds of dictionaries, for example WordNets (Luca and Nurnberger, 2006; Ranieri et al., 2004), in query translation. Although dictionaries can provide effective translation on common words or even phrases, they are always limited in the coverage. Hence, there is a need to expand the existing collections of bilingual terminologies through various means.

Recently, there has been more and more research work focus on bilingual terminology extraction from comparable corpora. Some promising results have been reported making use of statistics, linguistics (Sadat et al., 2003), transliteration (Udupa et al., 2008), date information (Tao and Zhai, 2005) and document alignment approach (Talvensaari et al., 2007).

In this paper, we introduce our Multilingual Access and Retrieval System – MARS which addresses the query translation issue by using in-domain bilingual terminologies extracted directly from the comparable corpora which are to be accessed by users. And at the same time, bilingual documents are paired up prior to the search process based on their content similarities to overcome the limitation of traditional keyword matching based on the translated terms. These would provide better retrieval experiences as not only more accurate in-domain translated term will be used to retrieve the documents but also provide a new perspective of multilingual information retrieval to process the time-consuming multilingual document matching at the backend.

The following sections of this paper will describe the system architecture and the proposed functionalities of the MARS system.

2 MARS System

The MARS system is designed to enhance query translation and document retrieval through mining the underlying multilingual structures of comparable corpora via a pivot language. There are three reasons for using a pivot language. Firstly, it is appropriate to use a universal language among potential users of different native languages. Secondly, it reduces the backend data processing cost by just considering the pair-wise relationship between the pivot language and any other languages. Lastly, the dictionary resources between the pivot language and all the other languages are more likely to be available than otherwise.

There are two main parts in this system, namely data processing and user interface. The data processing is an offline process to mine the underlying multilingual structure of the compa-

Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, pages 21–24,
Suntec, Singapore, 3 August 2009.

rable corpora to support retrieval. The structure of the comparable corpora is presented visually in the user interface under browse mode and search mode to facilitate navigation and retrieval of information respectively.

3 Data Processing

For demo purpose, three different language newspapers from the year 1995 to 2006 published by Singapore Press Holding (SPH), namely Strait Times[1] (English), ZaoBao[2] (Chinese) and Berita Harian[3] (Malay), are used as comparable corpora. In these particular corpora, English is chosen as the pivot language and noun terms are chosen as the basic semantic unit as they represent a huge amount of significant information. Our strategy is to organize and manipulate the corpora in three levels of abstraction – clusters, documents and terms. And our key task over here is to find the underlying associations of documents or terminologies in each level across different languages.

First, monolingual documents are grouped into clusters by k-means algorithm using simple word vectors. Then, monolingual noun terms are extracted from each cluster using linguistic patterns and filtered by occurrence statistics globally (within cluster) and locally (within document), so that they are good representatives for cluster as a whole as well as individual documents (Vu et al., 2008). The extracted terms are then used in document clustering in a new cycle and the whole process is repeated until the result converges.

Next, cluster alignment is carried out between the pivot language (English) and the other languages (Chinese, Malay). Clusters can be conceptualized as the collection of documents with the same themes (e.g. finance, politics or sports) and their alignments as the correspondents in the other languages. Since there may be overlaps among themes, e.g. finance and economy, each cluster is allowed to align to more than one cluster with varying degree of alignment score.

After that, document alignment is carried out between aligned cluster pairs (Vu et al., 2009). Note that the corpora are comparable, thus the aligned document pairs are inherently comparable, i.e. they are similar in contents but not identical as translation pairs. Also as important to note that, document alignment harvested over here is independent of user query. In other words, document alignment is not simply determined by mere occurrence of certain keyword and its absence does not hinder documents to be aligned. Hence mining of document alignment beforehand improves document retrieval afterward.

Finally, term alignment is likewise generated between aligned document pairs. The aligned terms are expected to be in-domain translation pairs since they are both derived from documents of similar contents, and thus they have similar contexts. By making use of the results provided by each other, document alignment and term alignment can be improved over iterations.

All the mentioned processes are done offline and the results are stored in a relational database which will handle online queries generated in the user interface later on.

4 User Interface

As mentioned, there are two modes provided in the user interface to facilitate navigation and retrieval of information, namely browse mode and search mode. Both modes can be switched simply by clicking on the respective tabs in the user interface. In the following, the functionalities of the browse mode and the search mode will be explained in details.

4.1 Browse Mode

Browse mode provides a means to navigate through the complex structures underneath an overwhelming data with an easily-understood, user-friendly graphical interface. In the figure 1, the graph in the browse mode gives an overall picture of the distribution of documents in various clusters and among the different language collections. The outer circles represent the language repositories and the inner circles represent the clusters. The sizes of the clusters are depending on the number of contained documents and the color represents the dominant theme. The labels of the highlighted clusters, characterized by a set of five distinguished words, are shown in the tooltips next to them. By clicking on a cluster, the links depicting the cluster alignments will show up. The links to the clusters in the other languages are all propagated through the pivot language.

[1] http://www.straitstimes.com/ an English news agency in Singapore. Source © Singapore Press Holdings Ltd.
[2] http://www.zaobao.com/ a Chinese news agency in Singapore. Source © Singapore Press Holdings Ltd.
[3] http://cyberita.asia1.com.sg/ a Malay news agency in Singapore. Source © Singapore Press Holdings Ltd.

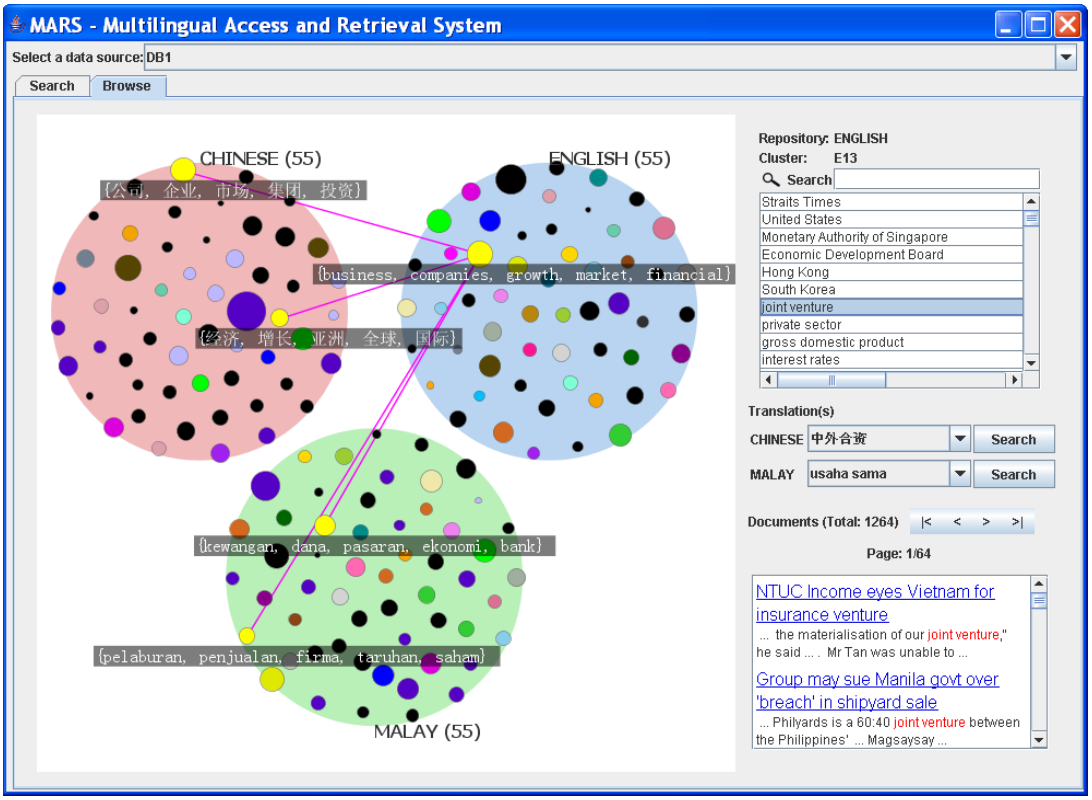

Fig. 1 Browse mode in the MARS System

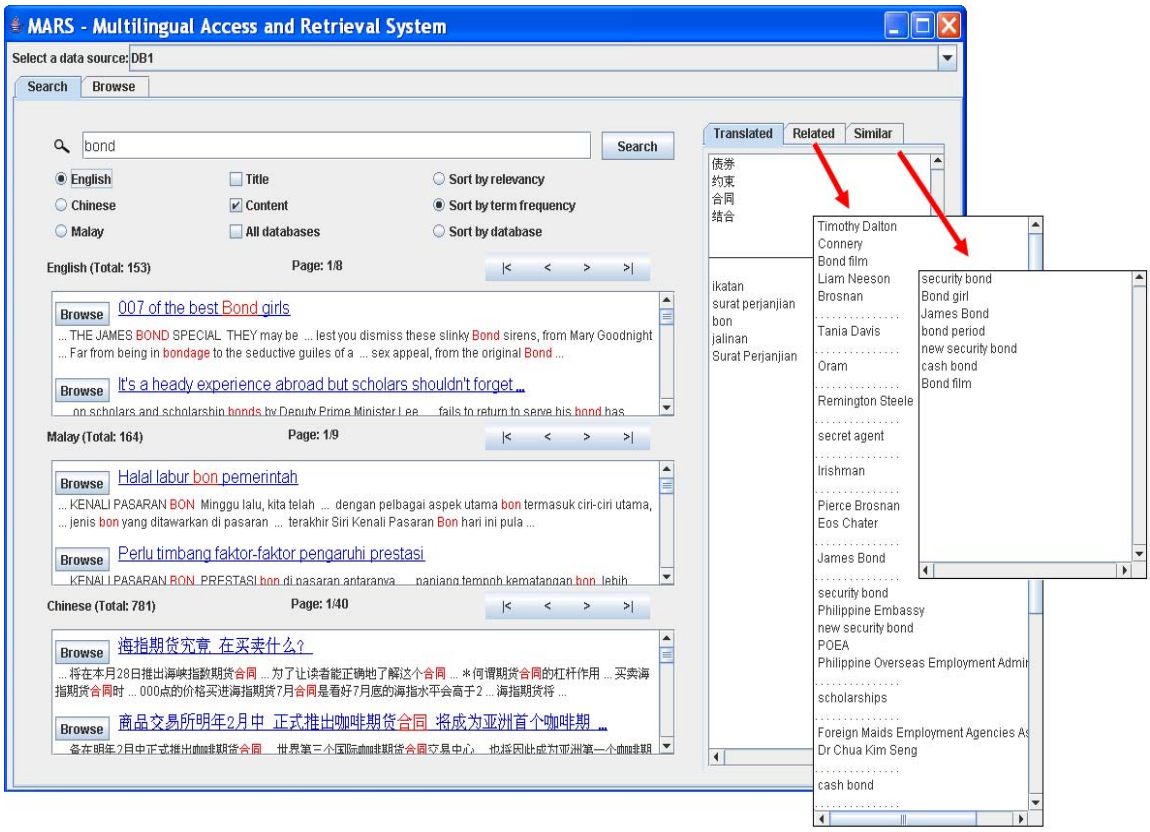

Fig. 2 Search mode in the MARS System

The right hand side of the browse panel provides the detail information about the selected cluster using three sub-panels, i.e. top, middle and bottom. The top panel displays a list of extracted terms from the selected cluster. User may narrow down the list of interested terms by using the search-text column on top. By clicking on a term in the list, its translations in other languages, if any, will be displayed in the middle sub-panel and the document containing the term will be listed in the bottom sub-panel. The "Search" buttons next to the term translations provide a short-cut to jump to the search mode with the corresponding term translation being cut and pasted over. Last but not least, user may simply click on any document listed in the bottom sub-panel to read the content of the document and its aligned documents in a pop-up window.

4.2 Search Mode

Search mode provides a means for comprehensive information retrieval. Refer to the figure 2, user may enter query in any of the selected languages to search for documents in all languages. The main difference is that query translation is done via bilingual terms extracted via the term alignment technology discussed earlier. For each retrieved document, documents with similar content in the other languages are also provided to supplement the searched results. This enables documents which are potentially relevant to the users be retrieved as some of these retrieved documents may not contain the translated terms at all.

On top of the query translation, other information such as related terms and similar terms to the query are shown at the tab panel on the right. Related terms are terms that correlate statistically with the query term and they are arranged by cluster, separated by dotted line in the list. Similar terms are longer terms that contains the query term in itself. Both the related terms and the similar terms provide user additional hints and guides to improve further queries.

5 Conclusion

The MARS system is developed to enable user to better navigate and search information from multilingual comparable corpora in a user-friendly graphical user interface. Query translation and document retrieval is enhanced by utilizing the in-domain bilingual terminologies and document alignment acquired from the comparable corpora itself, without limited by dictionaries and keyword matching.

Currently, the system only support simple query. Future work will improve on this to allow more general query.

References

Ernesto William De Luca, and Andreas Nurnberger. 2006. *A Word Sense-Oriented User Interface for Interactive Multilingual Text Retrieval*, In Proceedings of the Workshop Information Retrieval, Hildesheim.

M. Ranieri, E. Pianta, and L. Bentivogli. 2004. *Browsing Multilingual Information with the MultiSemCor Web Interface*, In Proceedings of the LREC-2004 Workshop "The amazing utility of parallel and comparable corpora", Lisban, Portugal.

Fatiha Sadat, Masatoshi Yoshikawa, Shunsuke Uemura. 2003. *Learning bilingual translations from comparable corpora to cross-language information retrieval: hybrid statistics-based and linguistics-based approach*, In Proceedings of the 6th international workshop on Information Retrieval with Asian Languages, vol. 1: pp. 57-64.

Raghavendra Udupa, K. Saravanan, A. Kumaran, Jagadeesh Jagarlamudi. 2008. *Mining named entity transliteration equivalents from comparable corpora*. In Proceedings of the 17th ACM conference on Information and knowledge management.

Tao Tao, and ChengXiang Zhai. 2005. *Mining comparable bilingual text corpora for cross-language information integration.* In Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining.

Tuomas Talvensaari, Jorma Laurikkala, Kalervo Jarvelin, Martti Juhola, Heikki Keskustalo. 2007. *Creating and exploiting a comparable corpus in cross-language information retrieval*. ACM Transactions on Information System (TOIS), vol. 25(1): Article No 4.

Thuy Vu, Aiti Aw, Min Zhang. 2008. *Term extraction through unithood and termhood unification.* In Proceedings of the 3rd International Joint Conference on Natural Language Processing (IJCNLP-08), Hyderabad, India.

Thuy Vu, Aiti Aw, Min Zhang. 2009. *Feature-based Method for Document Alignment in Comparable News Corpora.* In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09), Athens, Greece.