Bridging the Gap between Technology and Users: Leveraging Machine Translation in a Visual Data Triage Tool

Thomas Hoeft
Pacific Northwest National Laboratory
902 Battelle Blvd.
Richland, WA 99354

Nick Cramer
Pacific Northwest National Laboratory
902 Battelle Blvd.
Richland, WA 99354

M. L. Gregory
Pacific Northwest National Laboratory
902 Battelle Blvd.
Richland, WA 99354

Elizabeth Hetzler
Pacific Northwest National Laboratory
902 Battelle Blvd.
Richland, WA 99354

{thomas.hoeft;nick.cramer;michelle.gregory;beth.hetzler}@pnl.gov

1 Introduction

While one of the oldest pursuits in computational linguistics (see Bar-Hillel, 1951), machine translation (MT) remains an unsolved problem. While current research has progressed a great deal, technology transfer to end users is limited. In this demo, we present a visualization tool for manipulating foreign language data. Using software developed for the exploration and understanding of large amounts of text data, IN-SPIRE (Hetzler & Turner 2004), we have developed a novel approach to mining and triaging large amounts of foreign language texts. By clustering documents in their native language and only using translations in the data triage phase, our system avoids the major pitfalls that plague modern machine translation. More generally, the visualization environment we have developed allows users to take advantage of current NLP technologies, including MT. We will demonstrate use of this tool to triage a corpus of foreign text.

2 IN-SPIRE

IN-SPIRE (Hetzler et al., 2004) is a visual analytics tool developed by Pacific Northwest National Laboratory to facilitate the collection and rapid understanding of large textual corpora. IN-SPIRE generates a compiled document set from mathematical signatures for each document in a set. Document signatures are clustered according to common themes to enable information retrieval and visualizations. Information is presented to the user using several visual metaphors to expose different facets of the textual data. The central visual metaphor is a galaxy view of the corpus that allows users to intuitively interact with thousands of documents, examining them by theme.

Context vectors for documents such as LSA (Deerwester et al., 1990) provide a powerful foundation for information retrieval and natural language processing techniques. IN-SPIRE leverages such representations for clustering, projection and queries-by-example (QBE). In addition to standard Boolean word queries, QBE is a process in which a user document query is converted into a mathematical signature and compared to the multi-dimensional mathematical representation of the document corpus. A spherical distance threshold adjustable by the end user controls a query result set. Using IN-SPIRE's group functionality, subsets of the corpus are identified for more detailed analyses. Information analysts can isolate meaningful document subsets into groups for hypothesis testing and the identification of trends. Depending on the corpus, one or more clusters may be less interesting to users. Removal of these documents, called "outliers", enables the investigator to more clearly understand the relationships between remaining documents. These tools expose various facets of document text and document interrelationships.

3 Foreign Language Triage Capabilities

Information analysts need to sift through large datasets quickly and efficiently to identify relevant information for knowledge discovery. The need to sift through foreign language data complicates the task immensely. The addition of foreign language capabilities to IN-SPIRE addresses this need. We have integrated third party translators for over 40 languages and third party software for language identification. Datasets compiled with language detection allow IN-SPIRE to automatically select the most appropriate translator for each document.

To triage a foreign language dataset, the system clusters the documents in their native language

Proceedings of HLT/EMNLP 2005 Demonstration Abstracts, pages 2–3,
Vancouver, October 2005.

(with no pre-translation required). A user can then view the cluster labels, or peak terms, in the native language, or have them translated via Systran (Senellart et al., 2003) or CyberTrans (not publicly available). The user can then explore the clusters to get a general sense of the thematic coverage of the dataset. They identify clusters relevant to their interests and the tool reclusters to show more subtle themes differentiating the remaining documents. If they search for particular words, the clusters and translated labels help them distinguish the various contexts in which those words appear. Finding a cluster of document of interest, a particular document or set of documents can be viewed and translated on demand. This avoids the need to translate the entire document set, so that only the documents of interest are translated. The native text is displayed alongside the translation at all stages.

4 Evaluation

Since this is a prototype visualization tool we have yet to conduct formal user evaluations. We have begun field testing this tool with users who report successful data triage in foreign languages with which they are not familiar. We have also begun evaluations involving parallel corpora. Using Arabic English Parallel News Text (LDC 2004), which contains over 8,000 human translated documents from various Arabic new sources, we processed the English version in IN-SPIRE to view the document clusters and their labels. We also processed the Arabic version in Arabic according to the description above. The two screenshots below demonstrate that the documents clustered in similar manners (note that cluster labels have been translated in the Arabic data).

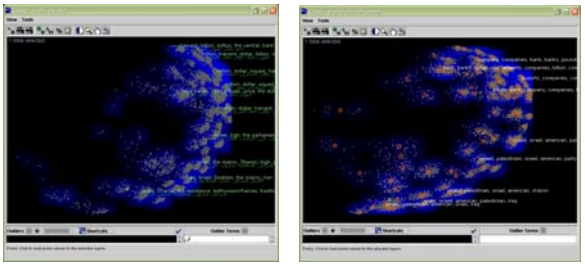

Figure 1: Galaxy view of the Arabic and English clusters and labels

To demonstrate that our clustering algorithm on the native language is an efficient and reliable method for data triage on foreign language data, we also pre-translated the data with CyberTrans and clustered on the output. Figure 3, demonstrates that similar clusters arise out of this methodology. However, the processing time was increased 15-fold with no clear advantage for data triage.

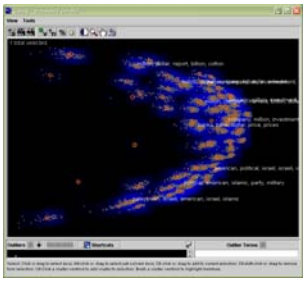

Figure 3: Galaxy view of the pre-translated Arabic to English clusters and labels

Initial user reports and comparisons with a parallel corpus demonstrate that our visualization environment enables users to search through and cluster massive amounts of data without native speaker competence or dependence on a machine translation system. Users can identify clusters of potential interest with this tool and translate (by human or machine) only those documents of relevance. We have demonstrated that this visualization tool allows users to derive high value from existing machine translation capabilities.

References

Bar-Hillel, Yehoshua, 1951. The present state of research on mechanical translation. *American Documentation* 2 (4), pp.229-237.

Hetzler, Elizabeth and Alan Turner. 2004. "Analysis Experiences Using Information Visualization," *IEEE Computer Graphics and Applications*, 24(5):22-26.

Deerwester, S., S.T. Dumais, T.K. Landauer, G.W. Furnas, R.A. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the Society for Information Science*, 41(6):391-407.

Linquistic Data Consortium. 2004. http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2004T18

Senellart, Jean; Jin Yang, and Anabel Rebollo. 2003. SYSTRAN Intuitive Coding Technology. MT Summit IX. New Orleans, Louisianna.