Research Topics for Graduate Students

Under Continuous Update

Fazli Can
Computer Engineering Department
Bilkent University

March 20, 2008

1. Data Fusion: improving our Turkish IR results by combining the results of various matching functions. For example, see papers written by Aslam et al. from acm.org/dl, Ill. Inst. of Tech. group paper from JASIST 2004, Vol. 55, No. 10 issue. Also see "Combining the evidence of multiple query representations for IR"http://www.scils.rutgers.edu/~belkin/articles/ipm_trec2.pdf. Extend the 2007 work.
2. Finding typos in Turkish words: gozum is incorrect, but gözüm is correct. Activities include the algorithm development, measuring its performance and integrating it into our news portal query interface. Please check out zemberek and turk-nlp yahoogroups to prevent reinventions.
3. Orhan Pamuk: a stylometric analysis of the Orhan Pamuk novels using machine learning methods (SVM, NN-search, etc.). Possibilities include analyzing the voices of Celal and Galip in *Kara Kitap*, analyzing the narration styles of different novel characters in *Sessiz Ev*, and comparing the narration styles of Faruk Darvınoğlu in *Sessiz Ev* and *Beyaz Kale*.
4. Text Summarization and Question Answering: see DUC (Document Understanding Conference: http://duc.nist.gov/) and TAC (Text Analysis Conference: http://www.nist.gov/tac/tracks/). This project has already been assigned.
5. Using stylistic features for Turkish IR: using stylometry for IR (visit http://eprints.sics.se/view/ and look for the publications of Jussi Karlgren, e.g., his dissertations).
6. Üç İstanbul: hypertextual version of Mithat Kemal Kuntay's novel, some data mining analysis is possible. For some excerpts form the novel see http://www.derkenar.com/kitapkurdu/kuntay.asp, for its TV movie series version information see http://www.sinematurk.com/film.php?7342. For motivation see the hypertext version of Faulkner's *The Sound and The Fury* (http://www.usask.ca/english/faulkner). Another possibility is analyzing the correspondence between the novel and its movie version (see Reyhan Tutumlu's master thesis for some initial ideas, available at http://library.bilkent.edu.tr/, use catalog search to obtain its pdf copy).
7. Query clustering/expansion: Using the AOL log (about 20 million queries). A paper from the literature can be implemented and it can be used for inspiration. *
8. Distributed IR: Document partitioning, term partitioning, and additionally cluster partitioning. Then measure distributed IR performance (real or simulated). *
9. Detecting document duplicates: see Roberto J. Bayardo, Ma, Srikant WWW '07 Conf. paper ("Scaling up all pairs similarity search": http://portal.acm.org/citation.cfm?doid=1242572.1242591). Note that duplicated documents affect both the system storage efficiency and the retrieval effectiveness (user satisfaction). *
10. Efficient snippet generation for Web search engines: see Turpin, Tsegay, Hawking, Williams SIGIR '07 Conf. paper ("Fast generation of result snippets in web search": http://portal.acm.org/citation.cfm?doid=1277741.1277766). *

Please look at the 2006-2007 student presentations for understanding the nature of the projects and further possibilities.

* Projects [7, 9] are defined and will be supervised BY/with Sengor Altingovde.