University of Massachusetts Amherst

Vol. 4, No. 11 • Fall 2002

Alum Matters

A newsletter for alumni of the Department of Computer Science

Breaking through the language barrier

RECENT YEARS HAVE SEEN AN EXPLOSION in the amount of electronic and on-line information. Although much of the content is in English, the amount of non-English resources and the number of non-English speaking Internet users are growing steadily. Increased global awareness, multi-national collaboration, and interest in foreign media sources have driven a need for Cross-Language Information Retrieval (CLIR) systems that span language boundaries. These systems allow the submission of a query in one language (e.g. Spanish) and retrieval of documents in other languages (e.g. Arabic). The Center for Intelligent Information Retrieval (CIIR) was one of the first groups to explore this area and is still at the forefront of advances in CLIR research. Lisa Ballesteros (Ph.D. 2001), former student of Bruce Croft and now Clare Boothe Luce Assistant Professor of Computer Science at Mount Holyoke College, continues to investigate techniques that enable access to and management of multilingual media.

Ballesteros's research focuses on language independent approaches exploiting readily available lexical resources. She developed a cross-language retrieval system based on approximate translation via machine-readable dictionary, augmented with statistical techniques for reducing the effects of translation ambiguity. Despite a relatively unsophisticated translation resource, Professor Ballesteros was one of the first to report a cross-language method that almost completely eliminated the effect of translation error from the retrieval process. "Dictionaries are more readily available than are other lexical resources, but there are still some language pairs for which no bilingual dictionaries exist," says Ballesteros. To address this issue, she became the first to publish results showing that transitive translation, which uses a third "pivot" language as an intermediate in the translation process, is a viable approach, producing acceptably accurate translation and subsequent retrieval. Her most recent work is investigating the use of statistical stemming for Arabic monolingual and cross-language retrieval. This is a language independent approach that measures the relatedness of stem-class pairs via their co-occurrence in fixed text windows.

Ballesteros

"CLIR tools can facilitate routine tasks of groups such as multi-nationals, information services, and government organizations (for instance, the Patent Office). These organizations routinely hire banks of multilingual speakers and translators to cull through vast amounts of foreign-language text to locate important information," says Ballesteros. "Furthermore, the events of September 11 have increased interest on the part of security organizations that are looking for ways to automatically identify items of interest in foreign language collections." Demand is also increasing for CLIR techniques to manipulate multimedia data including voice, images, and video. The growth of mobile accessory use has also led to challenging research questions regarding the development of mobile personal information retrieval systems. Such a system would exploit GPS information and other multimedia data to fulfill a specific information need given the user's current location. For example, a tourist may need to find a drug store that is nearest her current location. Unlike traditional IR systems in which results are generated using a global approach, mobile IR systems will need to focus on local context.

"The information technologies of the future will need more than the ability to cull through massive amounts of information," says Ballesteros. "They must also offer a wide range of tools for information analysis and manipulation; tools that facilitate discovery and organization, and that support the many new ways in which people interact with information."

Professor Ballesteros is spending her sabbatical at the University of Sheffield and at UMass, Amherst. She will be working on two main projects in the coming year. The first is developing corpus-based metrics that can be used to guide parameter selection for statistical stemming of different languages and corpora. The second is to develop a technique that relies on statistical analysis of contextual cues to infer translations for out-of-vocabulary (OOV) words. OOV words generally include special vocabulary such as technical terms that typically do not occur in bilingual dictionaries. An OOV solution would be useful for a wide range of tasks, including lexical acquisition and mobile personal IR, where a person may be unfamiliar with the correct foreign language term or expression to use in a specific context.