Multidocument, Multilingual, and Multimodal Information Extraction for Real World Applications

Mark T. Maybury

The MITRE Corporation
202 Burlington Road
Bedford, MA 01730

maybury@mitre.org
www.mitre.org/resources/centers/it

This keynote addresses current and future challenges in terminology and knowledge engineering focusing on multidocument, multilingual and multimodal information extraction. With some reports[1] that humanity creates more than an exabyte (10^{18} bytes) of unique information each year, the imperative for tools to mitigate the size, heterogeneity, and complexity of knowledge collections is a priority. After exemplifying this grand challenge in typical real world analytic environments, we briefly review the state of the art in information access. We note that automated systems exist that can return documents relevant to a particular subject with around 80% precision but low recall. Automated document query incorporating relevance feedback has achieved near human performance. Extraction of named entities (Hirschman 1998) is over 90% accurate and extraction of relations among entities in specific domains is about 70-80% accurate. Also, documents can be summarized to about 20% of their source size without information loss, which can save users 50% of their original task time. Finally, prototype systems can respond to a simple factual questions by returning answers from relevant documents with about 75% accuracy.

After this overview, we describe two terminology and information extraction activities we are presently engaged in, notably semi-automated terminology induction from machine readable dictionaries in the context of the Alembic system (Aberdeen et al 1995) and ontology induction and conceptual browsing. We then describe a range of applications that exploit terminology management including global infectious disease monitoring (MiTAP) (http://tides2000.mitre.org, Damianos et al 2002), topic detection and tracking in time and space (GeoNODE) (Hyland et al 1999), multimodal topic extraction (BNN) (Maybury 1997), biology terminology mining (KDD Cup 2002), and extraction for expertise management (Expert Finder, XperNet). In this final area we illustrate how terminology extraction has been applied to corporate knowledge management (Morey et al. 2000). In particular, we describe the creation and evaluation of Expert Finder (Mattox et al 1998, 1999), an expert skill finder that exploits the intellectual products created within an enterprise to support automated expertise classification. We also describe XperNet addresses the problem of detecting extant or emerging areas of human expertise without a priori knowledge of their existence. We conclude noting that as a community we can make the most rapid progress using corpus-based evaluation. This entails creating challenge problems with supporting data, sharing data, resources and tools, evaluating system performance on these problems, and comparing approaches.

[1] www.sims.berkeley.edu/how-much-info/summary.html

References

1. Aberdeen, A., Burger, J., Day, D., Hirschman, L., Robinson, P. & Vilain, M. 1995. MITRE: Description of the Alembic System as Used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference* (MUC-6), Columbia, MD, 6-8 November 1995, 141-155.
2. Damianos, L,. Day, D., Hirschman, L, Kozierok, R., Mardis, S., McEntee, T., McHenry, C., Miller, K., Ponte, J., Reeder, F., van Guilder, L, Wellner, B., Wilson, G., and Wohlever, S. 2002. Real Users, Real Data, Real Problems: The MiTAP System for Monitoring Bio Events. Proceedings of BTR 2002: Unified Science and Technology for Reducing Biological Threats and Countering Terrorism. Univ. of New Mexico, March 2002. http://www.mitre.org/support/papers/tech_papers_02/damianos_mitap/index.shtml
3. Hirschman, L. 1998. The Evolution of Evaluation: Lessons from the Message Understanding Conferences. *Computer Speech and Language*. 12: 281-305
4. Hyland, R., Clifton, C., and Holland, R. 1999. Geonode: Visualizing News in Geospatial Context. AFCEA Federal Data Mining Symposium. Washington, D.C.
5. KDD Cup 2002. http://www.biostat.wisc.edu/~craven/kddcup/index.html
6. Mattox, D., Smith, K., and Seligman, L. 1998. Software Agents for Data Management. In Thuraisingham, B. *Handbook of Data Management*, CRC Press: New York. 703-722.
7. Mattox, D., Maybury, M. and Morey, D. 1999. Enterprise Expert and Knowledge Discovery. International Conference on Human Computer International (HCI 99). 23-27 August 1999. Munich, Germany. 303-307.
8. Maybury, M. T. (ed.) 1997. *Intelligent Multimedia Information Retrieval*. Menlo Park: AAAI/MIT Press. (http://www.aaai.org/Press/Books/Maybury-2
9. Morey, D.; Maybury, M. and Thuraisingham, B. editors, Fall 2000. *Advances in Knowledge Management: Classic and Contemporary Works*. Cambridge: MIT Press.