Information Extraction using Machine Learning Techniques

As suggested by the overview given above on IE methods for biomedical applications, there is a large and growing volume of research on the problem of automatically extracting factual information from text. Examples of specific subproblems to which learning methods have been applied include text classification, using a variety of techniques [1, 2, 3, 4, 5]; information extraction using hidden Markov models [6, 7] and related sequence models [8, 9]; learning models of individual web sites by taking advantage of site-specific layout structure [10, 11, 12, 13, 14]; using website-specific layout structure to improve extraction [15, 16] and text classification [17] systems; and "semi-supervised" learning from partially labeled data [18, 19, 20]. These research efforts, and many others, provide an indication of the current popularity and effectiveness of machine learning approaches to information extraction.

Many of these research results explore a single highly specific aspect of the problem of extracting information. A smaller number of end-to-end extraction systems have been built. Many of these are designed for extracting facts from on-line information sources on the web [21, 22, 10, 23, 24], including from on-line scientific papers [25, 26, 27]. Two of the investigators for this grant have been involved in enterprises of this sort.

The WHIRL project [28, 29] made use of statistical similarity metrics to collect and integrate facts extracted from several dozen web sites in several different domains. These collected facts were then used for a number of different purposes: in addition to providing an end-user query interface to the integrated database [23], experiments were performed in which datamining techniques (such as collaborative filtering [14] and classification learning [30]) were applied directly to automatically-extracted fact. Information extraction was performed using a number of techniques, including using machine learning techniques for automatically recognizing and parsing web pages with large, regular collections of information [31, 32, 14, 30]. The WHIRL software has also been applied to the problem of recommendation of technical papers [33]. The WHIRL software is freely distributed for academic purposes, and has been used by a number of external research groups [34, 35, 36].

The WebKB project [37], for which Mitchell was the principle investigator, also aimed at extracting information from the web. This project developed a database describing university faculty, students, courses, and research projects from approximately ten universities. The WebKB effort developed a variety of methods for automated classification of web pages (e.g., [19, 5] and as well as extraction of factual information from text (e.g., [38]). The approach taken in this WebKB project is to use machine learning methods to train a system to automatically locate web pages containing relevant information, and to then automatically extract the desired factual information from these pages. Later, a number of the principals of the WebKB project went on to form the CORA project [26], which used the same general approach to extract information from research publications encoded in postscript.

Subsequently, Mitchell and Cohen were involved (as Chief Scientist and Distinguished Research Scientist, respectively) in a startup called WhizBang Labs, which aimed at commercializing these techniques. WhizBang Labs successfully implemented several complex, multistep IE systems: for instance, a IE system for job postings was successfully sold to TMPW, the holding company for the Monster.com job board; an IE system for extracting corporate information was used to provide additional information for Dun and Bradstreet's corporate database; and an IE system for continuing education courses was by the United

States Department of Labor. Each of these systems was far larger in scale than earlier research prototypes, extracting information from up to several million sites (in the case of the corporate information system).

Other relevant prior work and preliminary results

Realizing the aims stated here requires addressing a number of difficult technical problems, over and above the core issues of information extraction and location proteomics discussed above. The team assembled for this grant thus includes expertise in several additional areas. One key area is information retrieval *[cite: Jamie's favorite pubs here]*: this will be crucial for identifying papers relevant to a scientists' need, and providing appropriate access to the information that is extracted (which is largely textual) . Also, since the extracted information is partly in the form of images, expertise with multimedia databases *[cite: Christos's favorite pubs here]* is important in order to provide useful and efficient access to extracted information. Finally, since the ultimate goal is to integrate the information extracted by this system with external databases, expertise in data integration [39, 40, 41, 42, 43, 29, 44] is important.

References

[1] David Lewis. Representation and learning in information retrieval. Technical Report 91-93, Computer Science Dept., University of Massachusetts at Amherst, 1992. PhD Thesis.

[2] David D. Lewis, Robert E. Schapire, James P. Callan, and Ron Papka. Training algorithms for linear text classifiers. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 298–306, Zurich, Switzerland, 1996.

[3] William W. Cohen and Yoram Singer. Context-sensitive learning methods for text categorization. In *Proceedings of the 19th Annual International ACM Conference on Research and Development in Information Retrieval*, pages 307–315, Zurich, Switzerland, 1996. ACM Press.

[4] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of the European Conference on Machine Learning*, 1998.

[5] Kamal Nigam, Andrew K. McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39((2/3)):1–32, 2000.

[6] S. Ray and M. Craven. Representing sentence structure in hidden markov models for information extraction. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Seattle, WA, 2001.

[7] D. M. Bikel, R. L. Schwartz, and R. M. Weischedel. An algorithm that learns what's in a name. *Machine Learning*, 34:211–231, 1999.

[8] Andrew McCallum, Dayne Freitag, and Fernando Pereira. Maximum entropy markov models for information extraction and segmentation. In *Proceedings of the International Conference on Machine Learning (ICML-2000)*, Palo Alto, CA, 2000.

[9] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning (ICML-2001)*, Williams, MA, 2001.

[10] Craig A. Knoblock, Steven Minton, Jose Luis Ambite, Naveen Ashish, Pragnesh Jay Modi, Ion Muslea, Andrew G. Philpot, and Sheila Tejada. Modeling web sources for information integration. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Madison, WI, 1998.

[11] Ian Muslea, Steven Minton, and Craig Knoblock. A hierarchical approach to wrapper induction. In *Proceedings of the Third International Conference on Autonomous Agents (Agents'99)*, Seattle, WA, 1999.

[12] D. Freitag and N. Kushmeric. Boosted wrapper induction. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000)*, Austin, TX, 2000.

[13] N. Kushmeric. Wrapper induction: efficiency and expressiveness. *Artificial Intelligence*, 118:15–68, 2000.

[14] William W. Cohen and Wei Fan. Web-collaborative filtering: Recommending music by crawling the web. In *Proceedings of The Ninth International World Wide Web Conference (WWW-2000)*, Amsterdam, 2000.

[15] Jamie Callan and Teruko Mitamura. Knowledge-based extraction of named entities. In *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM'02)*, McLean, VA, 2002.

[16] David M. Blei, J. Andrew Bagnell, and Andrew K. McCallum. Learning with scope, with application to information extraction and classification. In *Proceedings of UAI-2002*, Edmonton, Alberta, 2002.

[17] William W. Cohen. Improving a page classifier with anchor extraction and link analysis. In *Advances in Neural Processing Systems 15*, Vancouver, BC, 2002. To appear.

[18] E. Riloff and R. Jones. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, 1999.

[19] Avrin Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 1998 Conference on Computational Learning Theory*, Madison, WI, 1998.

[20] Michael Collins and Yoram Singer. Unsupervised models for named entity classification. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP99)*, College Park, MD, 1999.

[21] Alon Y. Levy, Anand Rajaraman, and Joann J. Ordille. Querying heterogeneous information sources using source descriptions. In *Proceedings of the 22nd International Conference on Very Large Databases (VLDB-96)*, Bombay, India, September 1996.

[22] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, and J. Widom. The TSIMMIS approach to mediation: Data models and languages (extended abstract). In *Next Generation Information Technologies and Systems (NGITS-95)*, Naharia, Israel, November 1995.

[23] William W. Cohen. Reasoning about textual similarity in information access. *Autonomous Agents and Multi-Agent Systems*, pages 65–86, 1999.

[24] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to construct knowledge bases from the world wide web. *Artifical Intelligence*, 118((1-2)):69–113, 2000.

[25] Steve Lawrence, C. Lee Giles, and Kurt Bollacker. Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6):67–71, 1999.

[26] A. K. McCallum, K. Nigam, J. Rennie, , and K. Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval Journal*, 3:127–163, 2000.

[27] Researchindex: The NECI scientific literature digital library. http://www.researchindex.com, 2002.

[28] William W. Cohen. WHIRL: A word-based information representation language. *Artificial Intelligence*, 118:163–196, 2000.

[29] William W. Cohen. Data integration using similarity joins and a word-based information representation language. *ACM Transactions on Information Systems*, 18(3):288–321, July 2000.

[30] William W. Cohen. Automatically extracting features for concept learning from the web. In *Machine Learning: Proceedings of the Seventeeth International Conference*, Palo Alto, California, 2000. Morgan Kaufmann.

[31] William W. Cohen and Wei Fan. Learning page-independent heuristics for extracting data from web pages. In *Proceedings of The Eigth International World Wide Web Conference (WWW-99)*, Toronto, 1999.

[32] William W. Cohen. Recognizing structure in web pages using similarity queries. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)*, Orlando, FL, 1999.

[33] C. Basu, H. Hirsh, W. W. Cohen, and C. Nevill-Manning. Technical paper recommendation: A study in combining multiple information sources. *Journal of Artificial Intelligence Research*, 14:231–252, 2001.

[34] Sarah Zelikovitz and Haym Hirsh. Improving short-text classification using unlabeled background knowledge to assess document similarity. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML-2000)*, Palo Alto, CA, 2000.

[35] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the Fifth ACM International Conference on Digital Libraries*, 2000.

[36] T. Chinenyanga and N. Kushmerick. Expressive retrieval from XML documents. In *Proceedings of the 24th ACM International Conference on Research and Development in Information Retrieval (SIGIR-2001)*, New Orleans, USA, 2001.

[37] Cmu world wide knowledge base (Web-¿KB) project. http://www.cs.cmu.edu/ webkb, 2002.

[38] Dayne Freitag. Information extraction from html: application of a general machine learning approach. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Madison, WI, 1998.

[39] A. Monge and C. Elkan. The field-matching problem: algorithm and applications. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, August 1996.

[40] Eric Sven Ristad and Peter N. Yianilos. Learning string edit distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):522–532, 1998.

[41] M. Bilenko and Ray Mooney. Learning to combine trained distance metrics for duplicate detection in databases. Technical Report Technical Report AI 02-296, Artificial Intelligence Lab, University of Texas at Austin, 2002. Available from http://www.cs.utexas.edu/users/ml/papers/marlin-tr-02.pdf.

[42] I. P. Felligi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Society*, 64:1183–1210, 1969.

[43] W. E. Winkler. The state of record linkage and current research problems. Statistics of Income Division, Internal Revenue Service Publication R99/04. Available from http://www.census.gov/srd/www/byname.html, 1999.

[44] William W. Cohen, Henry Kautz, and David McAllester. Hardening soft information sources. In *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining*, pages 255–259, 2000.