Learning Information Extractors from Approximately Correct Patterns and Weakly Labeled Training Data

Mark Craven
craven@biostat.wisc.edu
Department of Biostatistics & Medical Informatics
University of Wisconsin
1300 University Ave.
Madison, WI 53711, U.S.A.

Our group has an ongoing research project that is aimed at making it easy and inexpensive to develop accurate information-extraction (IE) systems in new domains. Our research is focused both on developing new methods for information extraction [2, 3], and on applying these methods in biomedical applications [1, 3].

One common approach to developing an information-extraction system is to manually encode patterns or rules that can be applied to text sources in order to extract relation instances. Work in the field of information extraction over the last decade has shown that it is typically very expensive (in terms of an expert's time) to develop highly accurate IE systems through this hand-coding process. In recent years, there has been interest in using machine learning methods to induce IE modules from labeled training examples. The premise for this approach is that it is typically less expensive to assemble a set of labeled training examples than to directly hand-code extraction patterns. However, obtaining hand-labeled training examples is often itself a very expensive process.

Our approach to constructing IE modules is based on two key tenets. First, a small set of approximately correct extraction patterns can often be assembled with little cost. We initialize our IE models using such a set of approximately correct patterns, and then use machine learning methods to improve the accuracy of the models. Second, in many domains IE models can be learned from "weakly" labeled training data which is inexpensive to collect. Weakly labeled data consists of facts to be extracted (i.e., known relation instances) along with documents that *may* assert the facts. Conventional training data, in contrast, consists of documents precisely annotated with the facts that should be extracted from them. Weakly labeled training sets can be assembled by coupling existing information sources (knowledge bases, databases, or even simple lists or tables) with related documents. Our IE systems use shallow parsing techniques, and our learned models are able to represent regularities in grammatical structure and word statistics [3].

To develop and demonstrate our ideas, we have put together sets of weakly labeled training data for three target relations: subcellular locations of proteins, gene-disorder associations, and protein-protein interactions. We have used the first two of these data sets in recent publications [1, 3], and have carefully measured precision and recall on held-aside test sets to evaluate the accuracy of our learned models. We are currently using our IE models to populate a database of protein localization information, and to support the development of a model of cancer-related signaling and regulatory pathways.

References

[1] M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 77–86, Heidelberg, Germany, 1999. AAAI Press.

[2] M. Craven and S. Slattery. Relational learning with statistical predicate invention: Better models for hypertext. *Machine Learning*, 43(1-2):97–119, 2001.

[3] S. Ray and M. Craven. Representing sentence structure in hidden Markov models for information extraction. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, Seattle, WA, 2001. Morgan Kaufmann.