THE SRI TIPSTER III PROJECT

Steven Maiorano
Office of Advanced Analytic Tools
Washington, D.C. 20505
E-mail: stevejm@ucia.gov
Telephone: 703-613-8755

Introduction[1]

The SRI TIPSTER Phase III research program focused on improving information extraction (IE) accuracy by further extending their Phase II [1] research into linguistic approaches to data extraction. At the same time, SRI's efforts emphasized portability, and the moving away from hand-crafting complicated rules toward enabling nonexpert users to generate automatically new domain patterns.

CPSL

One step towards ease-of-use by nonexperts was the development reported in Phase II [1] of SRI's FastSpec language which enabled greater facility in generating and modifying the syntactic and semantic patterns necessary for identifying pertinent data. This was a motivating factor for the establishment of the Common Pattern Specification Language (CPSL) Working Group devoted to formulating a CPSL in order that different IE systems could share pattern libraries.[2]

Open Domain System

A central goal of SRI's open domain work was to allow users to develop their own customized IE systems called Fastlets for different topics of interest. SRI analyzed a large corpus of business news and developed a basic ontology of concepts. This ontology provides the basis for specifying patterns in a wide range of business news applications. Overall, the concepts and patterns cover 150 most common verbs and nominalizations in the Wall Street Journal. The patterns are currently being implemented in SRI's IE system, FASTUS.

As part of a collateral effort, SRI built Fastlets for 23 of the 47 TREC-6 [2] topics which were used in two joint TREC-6 submissions with GE. In the first, SRI re-ranked the top 2000 documents returned by the GE routing query. In the second submission GE used the topic-relevant text identified by the SRI Fastlets for query expansion. The results were encouraging and indicate that the Fastlet approach to document detection is feasible.

Merging

The previous FASTUS merging algorithm was simple in that it merged newly extracted data structures from a document with earlier extracted ones starting with the most recent. The results were good, but error analysis showed that improving this process would help increase overall system accuracy. SRI's approach was to apply machine learning (ML) techniques in order to learn merging strategies automatically -- an essential capability for enabling nonexperts to port the system to a new domain.

SRI employed two supervised learning techniques -- decision tree and maximum entropy. Both approaches were highly successful at classifying merges, but neither technique improved the overall accuracy of the end-to-end system. It is thought that this is due to the fact that correct merges improve the system's score much more than bad merges hurt it. Therefore, although the supervised-learning-based merges made more correct decisions than before, the incorrect blocking of a few correct merges reduced performance. SRI also experimented with unsupervised and weakly supervised learning of merging strategies. These results were reported at AAAI-98. [3]

Coreference Resolution

SRI's coreference module in FASTUS at the start of TIPSTER III performed close to the best published results for handling third person pronouns and reflexives. SRI worked on porting coreference improvements to the TIPSTER III system, and, at the same time, began work on coreferential elements to be handled for the first time; bare nominals,

---

[1] This material has been reviewed by the CIA. That review neither constitutes CIA authentication of information nor implies CIA endorsement of the author's views.

[2] SRI's Doug Appelt has proven the feasibility of CPSL by developing TextPro, a Mac-based IE system written in CPSL. (www.ai.sri.com/~appelt)

possessed nominals, and indefinites. It was thought that overall system degradation corresponds to the lack of semantic and discourse information available to the FASTUS processing modules and that WordNet might be a possible source knowledge to provide this information.

WordNet

In another related effort, SRI performed experiments in utilizing WordNet (WN) as a knowledge source for IE. SRI developed mechanisms for automatically adding relations on top of WN with respect to nominalizations. There are nominalizations in WN but they are not linked to the word sense of the verb that they nominalize. Because information such as "manager" is an AGENT nominalization of "manage" is useful to the open domain approach, SRI developed routines to identify the appropriate word sense. They also analyzed how WN might improve coreference resolution. Although the potential payoff of using WN is still unclear, it is clear that utilization of WN is not a simple matter of plug-and-play.

References

[1] Hobbs, J. R., Appelt, D., Bear, J., Israel, D., Kameyama, M., Kehler, A., Stickel, M., Tyson, M.; *"SRI's TIPSTER II Project,"* in Proceedings Tipster Text Program (Phase II) September 1996; Morgan Kaufman Publishers, Inc.; San Francisco, CA.

[2] Bear, J., Israel, D., Petit, D., Martin, D.;*"Using Information Extraction to Improve Document Retrieval,"* in Voorhees, E. and Harman, D. editors, The Sixth Text REtrieval Conference (TREC-6); NIST Special Publication 500-240, 1998.

[3] Kehler, A.; *"Learning Embedded Discourse Mechanisms for Information Extraction,"* in Applying Machine Learning to Discourse Processing (Papers from the 1998 AAAI Spring Symposium, Technical Report SS-98-01); AAAI Press; Menlo Park, CA; 1998.