Communications of the ACM, 35(12):80–81, December 1992.

Natural Language Understanding for Information-Filtering Systems

Ashwin Ram
College of Computing
Georgia Institute of Technology
Atlanta, Georgia 30332-0280
E-mail: ashwin@cc.gatech.edu

Traditional approaches to information filtering that rely on the occurrence of a given set of keywords to identify possibly relevant texts are limited by the fact that they involve no actual understanding of the input texts. Recent research in artificial intelligence, and in natural language understanding in particular, has resulted in technologies that can help in the design of *intelligent* information filtering systems. Such systems determine the interestingness or relevance of a text based on an analysis of the relationship of the content of the input to the user's interests. Consider the following simple example from our PIES system:

> The Chicago Cubs played the Atlanta Braves last night at Wrigley Field. Bret Saberhagen pitched for the Cubs, while Tom Glavine started for the Braves. Saberhagen had 3 strikeouts and allowed 2 earned runs, one coming off a solo homerun by Ron Gant in the 3rd inning.

PIES starts with a model that represents the interestingness and relevance values of different concepts as specified by a user. This model is used to guide an interest-based information filtering and extraction process. Suppose that PIES is reading this story from the point of view of a reader who is a fan of the Atlanta Braves, is interested in baseball plays and scores, and knows Bret Saberhagen personally. PIES first pre-processes the knowledge structures used to represent baseball games, pruning away parts that are unlikely to be of interest. For example, in this story the location of the game is probably not interesting, since it does not relate to the user's interests in any way. Thus there is no point in looking for, parsing, and representing the location information in this story.

Next, PIES uses the pruned knowledge structures to process the story. This process extracts information from the story that matches the representational elements it is looking for. This is done in a manner similar to most knowledge-based natural language understanding systems [e.g., 2, 3, 6]. At this point, the

story representation may still contain uninteresting facts. For example, if the Cubs pitcher was someone other than Saberhagen, the name of the pitcher would no longer be interesting. This determination can only be made after the story has been processed since if the system were to prune the pitcher information from its knowledge structures ahead of time, it would not be able to represent a situation in which Saberhagen (in whom the user is interested) was the pitcher. Thus, the final step is one in which PIES postprocesses the understood story, pruning away elements that are unlikely to be interesting to the user.

The pruning process is based the program's knowledge of the user, as well as on a user-supplied interest threshold that is used to represent the current inclinations of the user. For example, if the interest threshold is high, the user wants to see only those aspects of the story that are likely to be very interesting. In such a situation, the program prints out the following synopsis of the above story:

- The Chicago Cubs played the Atlanta Braves.
- Bret Saberhagen had 3 strikeouts and allowed 2 earned runs.
- Ron Gant had a home run in the 3rd inning against Bret Saberhagen.

A lower interest threshold may be used if the user has more time to browse through the details of the story. In this situation, more details of this story would be included in the synopsis. While PIES is a prototype system intended for research purposes, similar approaches have been scaled up to produce natural language understanding systems that can identify and extract relevant information in real-world application domains. For example, FERRET is a full text, conceptual information retrieval system that uses a partial understanding of its texts to provide greater precision and recall than keyword search techniques [4].

In the longer term, we see the eventual merging of information filtering systems with artificial intelligence systems that can represent, reason about and effectively manage their own information-seeking behavior (e.g., the recent workshop on Intelligent Information Retrieval [1, p. 233–573]. A key requirement in such systems is their ability to build and reason about explicit representations of the desired information—their *knowledge goals* [5, 7]. Such systems might, for example, remember queries and continue to notify users of incoming texts; generate more detailed queries based on an analysis of the specifications of user interests and their relationship to input texts; learn what the user does or does not find interesting and update their user models automatically; choose to filter and extract information, not explicitly because a user asked them to, but to learn more about some domain in support of their own problem solving actions; or use multiple strategies to look for, retrieve, filter, or infer information that might be relevant [e.g., 7].

References

1. Birnbaum, L. and Collins, G. Eds. *Machine Learning: Proceedings of the Eighth International Workshop*. Morgan Kaufman (Chicago, Ill. June 1991).

2. DeJong, G.F. An overview of the FRUMP system. In *Strategies for Natural Language Processing*, W.G. Lehnert and M.H. Ringle, Eds., chap. 5, Lawrence Erlbaum, Hillsdale, N.J., 1982, 149–276.

3. Lebowitz, M. Memory-based parsing. *Artif. Intell. 21*, 1983, 363–404.

4. Mauldin, M.L. Information retrieval from natural language text. Ph.D. thesis, Carnegie-Mellon University, Pittsburgh, Pa., Aug. 1989. Tech. Rep. CMU-CMT-90-122.

5. Ram, A. Knowledge goals: A theory of interestingness. In *Proceedings of the Twelvth Annual Conference of the Cognitive Science Society*, (Cambridge, Mass., July 1990), pp. 206–214.

6. Ram, A. A theory of questions and question asking. *J. Learning Sci. 1*, (3, 4), 1991, 273–318.

7. Ram, A. and Hunter, L. A goal-based approach to intelligent information retrieval. In *Machine Learning: Proceedings of the Eighth International Workshop* (Chicago, Ill., June 1991).