ED 301 168 IR 013 508

AUTHOR Dyer, Michael; Read, Walter

TITLE A Sourcebook Approach to Evaluating Artificial Intelligence Systems.

SPONS AGENCY Office of Naval Research, Arlington, Va.

PUB DATE Apr 88

CONTRACT N00014-86-K-0395

NOTE 6p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988). This work is part of the Artificial Intelligence Measurement System (AIMS) of the Center for the Study of Evaluation at the University of California, Los Angeles.

PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Artificial Intelligence; *Computational Linguistics; *Databases; *Evaluation Criteria; Expert Systems; Guides; *Research Methodology; Vertical Organization

IDENTIFIERS *Natural Language Processing

ABSTRACT

"The Sourcebook," a database of exemplars of representative problems in natural language understanding, is being developed for use in evaluating natural language processing systems. A literature search of the computational linguistics literature was used to identify the exemplars, which are analyzed in terms of the general issues represented and grouped into categories of related problems to generate a hierarchical classification of the issues. At this point, the Sourcebook contains several hundred exemplars, which are estimated to represent 10% of the literature in computational linguistics, artificial intelligence, and cognitive science. Exemplars will continue to be added, and further elaboration of the classification scheme is planned. (3 references) (EW)

A Sourcebook Approach to Evaluating Artificial Intelligence Systems*

Michael Dyer
Walter Read
Artificial Intelligence Laboratory
UCLA Computer Science Department

Paper presented at the Annual Meeting of the
American Educational Research Association

for the Symposium
"How Smart are Smart Computers?
Alternative Approaches to the Evaluation of Artificial Intelligence Technology"

April, 1988, New Orleans

*This work reported here is part of the Artificial Intelligence Measurement Systems (AIMS) Project, which is supported in part by ONR contract number N00014-86-K-0395.

Evaluating Natural Language Systems

Recent years have seen a proliferation of computer systems for natural language processing (NLP). These include front ends to databases, expert systems and tutoring systems. Such systems generally come with a list of inputs (typically single sentences) that the system is claimed to 'handle'. The problem in judging these systems is that it is very difficult to tell from the examples just what claims are being made. If one of the examples includes an ellipsis, does that mean the system handles ellipsis in general? Or only certain kinds? What is ellipsis 'in general'? Are there different kinds of ellipsis that require different kinds of understanding?

Evaluating these claims requires that we know what inputs the system *should* handle and what it would mean to understand the input. Testing understanding is easier for applied systems since there is generally a specific task involved, e.g., accessing a database. But deciding what inputs should be handled is more difficult because there is no general agreement on what kinds of linguistic phenomena there are. Without a common classification of the problems in natural language understanding authors have no way to specify clearly what their systems do, potential users have no way to compare different systems and researchers have no way to judge the advantages or disadvantages of different approaches to developing NLP systems.

This paper reports progress in development of evaluation methodologies for natural language systems. This work is part of the Artificial Intelligence Measurement System (AIMS) project of the Center for the Study of Evaluation at UCLA.

Previous Work

These problems have been discussed for some time in computer science NLP work but there has been very little work in developing actual evaluative criteria. Woods (1977) discussed the taxonomic approach and pointed out some of its strengths and weaknesses. Guida and Mauri (1984, 1986) discuss a formal model which involves measuring the correctness of the understanding and averaging it over a weighted set of inputs. But this method assumes that we can describe a weighting for (categories of) inputs.

The Sourcebook

In developing evaluative criteria for NLP systems we had several goals in mind. First, the criteria used should be applicable over the broadest possible range of systems and still provide comparability of the systems. Second, the system shouldn't just be rated on a pass/fail count. It should outline areas of competence so that implementers can see where further work is needed in their system. They should be able to say "this approach handles types 1, 2 and 3 of ellipsis but not types 4 and 5 yet" rather than "this approach handles ellipsis". Third, the criteria used should be comprehensible to the general user and to researchers outside computational linguistics. We need to present the issues in such a way that the user can make judgments about the importance of different components of the evaluation. This means presenting the issues in terms of the general principles involved and giving concrete examples. This approach also allows us to bring in information from areas like education, psychology, sociology, law and literary analysis and enables researchers in those areas to contribute to the evaluation.

To this end, we are building a database of *exemplars* of representative problems in natural language understanding, mostly from the computational linguistics literature. Each exemplar includes a piece of text (sentence, dialogue fragment, etc.) a description of the conceptual issue represented, a detailed discussion of the problems in understanding the text and a reference to a more extensive discussion in the literature. The Sourcebook consists of a large set of these exemplars and a conceptual taxonomy of the types of issues represented in the database. The exemplars are indexed by source in the literature and by conceptual class of the issue so that the user can readily access the relevant examples. The Sourcebook provides a structured representation of the coverage that can be expected of a natural language system.

Rather than start with a particular theory of language, we began with a search of the computational linguistics literature. While no-one would claim that computational linguistics has discovered, let alone solved, every problem in language use, twenty-five years of research has covered a broad range of problems. Looking at language use computationally focuses attention on phenomena that are often neglected in more theoretical analyses. Building systems intended to read real text or interact with real users raises complex problems of interaction of linguistic phenomena. The exemplars are mostly taken from the literature although we have added examples to feel in gaps where we felt the published examples were incomplete. Because many of the published cases involved particular systems, the examples are often discussed in the literature in relation tc that system. In the exemplars, we analyze the example in terms of the general issue represented. Then the exemplars are grouped into categories of related problems. This will generate the hie· .rchical classification of the issues.

Continuing and Future Work

We have several hundred exemplars and we estimate that we have covered 10 per cent of the relevant literature (journals, proceedings volumes, dissertations, major textbooks) in computational linguistics, artificial intelligence and cognitive science.

We are continuing to add exemplars to the Sourcebook and are elaborating the classification scheme. We will be making the Sourcebook available to other researchers for comment and analysis.

References

Guida, G. & Mauri, G. (1984). A Formal Basis for Pcrformance Evaluation of Natural Language Understanding Systems. *Computational Linguistics, 10*, 15-30.

Guida, G. & Mauri, G. (1986). Evaluation of Natural Language Processing Systems: Issues and Approaches. *Proceedings of the IEEE, 74*, 1026-1035.

Woods, W. A., (1977). A Personal View of Natural Language Understanding. *SIGART Newsletter*, 17-20.

A Sample Exemplar

(1) The next day after we sold our car, the buyer returned and wanted his money back. (Allen, 1987, p. 346)
(2) The day after we sold our house, the escrow company went bankrupt.
(3) The day after we sold our house, they put in a traffic light at the corner.

Topic

Anaphoric reference - roles.

Discussion

In (1) the 'buyer' refers back to a figure in one of the roles in the 'selling a car' event. The system must search not only the direct possible antecedents (the 'selling') but must also consider aspects of the selling to resolve the reference. In (1), there is nothing specific to 'car' about resolving the reference. But in (2), finding the reference of 'the escrow company' involves looking past the general "buying" script and searching through aspects of selling specific to selling houses. There is a general problem here with controlling the amount of search while still looking deep enough. In (3), the system has to go from the house to the location to the street to the corner to understand the reference.

Reference

Allen, J. F. (1987). *Natural Language Understanding.* Menlo Park, CA: Benjamin/Cummings.