AI Magazine Volume 11 Number 2 (1990) (© AAAI)

Future Directions in Natural Language Processing

The Bolt Beranek and Newman Natural Language Symposium

Mark T. Maybury

The Workshop on Future Directions in NLP was held at Bolt Beranek and Newman, Inc. (BBN), in Cambridge, Massachusetts, from 29 November to 1 December 1989. The workshop was organized and hosted by Madeleine Bates and Ralph Weischedel of the BBN Speech and Natural Language Department and sponsored by BBN's Science Development Program. Thirty-six leading researchers and government representatives gathered to discuss the direction of the field of natural language processing (NLP) over the next 5 to 10 years. The intent of the symposium was "to make the conference and resulting volume an intellectual landmark for the field of NLP." This brief article summarizes the invited papers and strategic planning discussions of the workshop.

Semantics and Knowledge Representation

Robert Moore of SRI International began the workshop by considering adverbs that modified either the fact ("Strangely, John sang.") or manner ("John sang strangely.") of sentences describing events. He detailed manner and factual modifications of situation descriptions applied to Davidsonian semantics and situation semantics and argued that only the fact use was possible with copula constructions. Bill Woods of Harvard replied with a copula counterexample (for example, "John is strangely tall"). It was noted that quantification introduced significant problems (for example, "John polished all boots quickly" versus "John quickly polished all boots").

Next, James Allen (University of Rochester) argued for a logical form with built-in ambiguity to bridge the NLP–knowledge representation gap. He defined *vagueness* as overgeneralization in a type hierarchy (for example, a horse can be a mare, a colt, . . .). In contrast, *ambiguity* was ascribed to different senses (for example, different word senses, such as a peach pit and a pit in the ground). Allen called for more work on limited inference systems not based on completeness. Although not convinced of the concept of vagueness, Norm Sondheimer (GE Research) replied that commonsense reasoning was necessary and pointed to Patrick Hayes's work on liquids.

Building a Lexicon

Sue Atkins (Oxford University Press) discussed work in computational lexicography on Oxford's machine-readable dictionary (on CD-ROM). She presented examples of her proposed structure for lexical entries based on logical definitions, where an entity is identified as a member of a class (genus) with characteristics that distinguish it from its siblings (differentia). Mark Maybury (Rome Air Development Center and Cambridge University) pointed out that logical definition was but one technique; others include synonymic, antonymic, and etymological definition as well as exemplification, classification (subtypes), and constituency (subparts).

Continuing the discussion of the lexicon, Beth Levin (Northwestern University) presented her analysis of sound verbs, illustrating how lexical-semantic generalizations could be made across a variety of lexical entries. She presented a sound verb lexical template that identifies a sound's physical properties (for example, low, high, shrill), manner of production (for example, by vibration, by blowing, electronically), and selectional restrictions (for example, plus or minus animate, human, concrete). This template can then be used for the semiautomatic acquisition of lexical knowledge, the interpretation of unknown words, or the handling of novel uses of known words. James Pustejovsky (Brandeis University) pointed to his entity and qualia theory of lexical semantics and noted that "by allowing both verbs and nouns to shift in type, we can spread the semantic load on the lexicon more evenly, while capturing the ways that words can extend their meanings."

Beth Levin's presentation was followed by Bran Boguraev's (IBM Yorktown Heights) discussion of the use of machine-readable sources (for example, dictionaries and text corpora) to semiautomatically construct thesauri as well as knowledge bases (for example, building a generalization hierarchy from an online lexicon). Don Walker (Bellcore and the Association for Computational Linguistics [ACL]) replied that several ACL initiatives (text collection, text encoding as well as Mitch Marcus's treebank effort (sponsored by the Defense Advanced Research Projects Agency) and the consortium for lexical research) will soon make millions of words of corpora available online. ACL has received over 400 million words that it is currently evaluating and classifying.

The intent . . . was "to make the conference and resulting volume an intellectual landmark for the field of NLP."

Challenging Problems

Mark Steedman (University of Pennsylvania) delivered a paper on the use of intonation contours to constrain syntactic parsing. He discussed how functional composition (for example, modal plus infinitive, as in "John might eat . . .") and subject type raising found elegant solutions with his approach. Rusty Bobrow (BBN) found Steedman's focus on the integration of speech signal information and language processing encouraging. Group

0738-4602/89/$4.00 ©1990 AAAI

discussion then centered on the contrast between using knowledge sources to filter syntactic structures during parsing and using them to filter after parsing.

Steedman's paper was followed by a presentation entitled "Critical Challenges for NLP" by Madeleine Bates, Rusty Bobrow, and Ralph Weischedel. Their comments centered on the need for automatic language acquisition and the need to handle realistic (for example, ill-formed) input in application user interfaces. Weischedel discussed BBN's recent work on interfacing with heterogeneous back-end applications (for example, expert systems, databases, planners, simulations). Bates focused on dealing with language novel to a natural language system (for example, unknown lexemes), discussing the exploitation of statistics as well as learning from a variety of knowledge sources. Bobrow pointed out the positive consequences of treating sentence fragments and multisentence chunks as first-class language. Christine Montgomery (Language Systems Inc.) replied that we need to better understand events and situations, as well as default reasoning, to handle more realistic language phenomena and applications.

Discourse

Rebecca Passonneau (Unisys, Paoli Research Center) presented results of a study in focus of attention and the choice between "it" and the demonstrative "that" in referring expressions in text. By abstracting a number of properties away from the data (for example, syntactic characteristics, given-new distinctions), she developed a state-transition representation of the selection of "it" versus "that." Candy Sidner (DEC) noted that this property-sharing approach contrasted with past approaches that rank ordered lists of forward-looking centers. Although pleased with Passonneau's statistical approach, Sidner warned that results using this methodology needed to be substantiated with cross-language and cross-application evidence. In particular, Sidner noted that Passonneau's data from career-counseling sessions were biased because their characteristics (for example, interactive discourse, heavy use of personal pronouns you and I) are not common to all forms of discourse (for example, text). Bonnie Webber pointed to related research on the Italian use of lo (it) versus quello (that).

> *Their comments centered on the need for automatic language acquisition and the need to handle realistic . . . input in application user interfaces.*

Spoken Language Systems

Janet Pierrehumbert (Northwestern University) presented a multilevel model of prosodic structure, including levels of tunes, phonetic segments, syllables, metric feet, words, minor intonational phrases, and major intonational phrases. She discussed how prosody and intonation convey information about organization, attentional structure, and the speaker's intention. James Allen supported Pierrehumbert's comments and suggested coupled automata (for example, cascaded augmented transition networks) as a formalism for dealing with integration of multiple knowledge sources, although he indicated this type of processing scheme is far beyond the current state of the art.

Richard Schwartz (BBN) overviewed the advantages and disadvantages of several speech-recognition strategies (predictive coding, lattice techniques, and N-best approaches) and suggested combining the use of statistical grammars for phonetic scoring with NLP syntax and semantics for subsequent filtering. He suggested an N-best algorithm that "grows and prunes" at each stage of processing, claiming that it empirically reduced computation from $O(N^2)$ to about $O(\sqrt{N})$.

Directions in Natural Language Processing

The symposium closed with a group strategy discussion led by Ralph Weischedel concerning future technical directions and government investment in NLP. The key areas identified for application exploitation in the next five (plus or minus two) years were data extraction from text, information retrieval, machine-assisted translation, limited task instruction (computer-aided instruction), report generation (that is, natural language planning and realization), meaning to speech systems, help-advisory systems, intelligent forms, and document checking.

Several key technologies were indicated that could bear fruit within the near term, including research in phonology and morphology, constrained grammatical formalisms, knowledge representation formalisms, lexical semantics, and discourse models (for example, models of attention, intention, situations and events). Many pointed to the need for shared resources (for example, corpora, lexicons, tools) to advance the state of the art in the technology.

Conclusion

The workshop underscored the importance of developing natural language interfaces to real-world applications and the problems they entail (for example, ill-formed input, large domain-specific lexicons, commonsense reasoning). It was felt that these activities would motivate additional theoretical and practical developments (for example, automatic acquisition of linguistic knowledge). The workshop was successful in focusing on the need for increased communication between the computational linguistics and speech communities.

Mark T. Maybury received a B.A. in mathematics in 1986 at the College of the Holy Cross in Worcester, Massachusetts. In 1987, he completed a M.Phil. in speech and natural language at Cambridge University. He is completing his Ph.D. dissertation in absentia at the Cambridge University Computer Laboratory, focusing on the generation of coherent multisentential explanations from knowledge-based systems. Maybury is currently a captain in the United States Air Force, stationed at the Air Force Systems Command Center for Excellence in Advanced Command and Control Systems at Rome Air Development Center, Griffiss Air Force Base, New York.