# 10 The future of computational linguistics

MADELEINE BATES AND RALPH M. WEISCHEDEL

## 10.1 Introduction

One of the most delightful features of a small symposium is that it allows for protracted discussions in which many people participate. Ample time for discussion was built into the symposium schedule throughout, but we allocated a special two-hour slot to challenge ourselves to identify the most significant problems capable of being solved in a five- to ten-year period. That they be solvable in that time frame challenges us beyond what we can see, but not beyond what we can reasonably extrapolate. That their solution be significant takes the discussion beyond questions of purely academic interest.

Furthermore, at the suggestion of one of the government representatives, we asked what applications should drive research (much as the application of natural language interfaces to database drove research in the 1970s and 1980s).

All attendees, including representatives of various governmental agencies, participated in this discussion. To keep our thoughts large, we construed natural language processing (NLP) as broadly as possible, freely including such areas as lexicography and spoken language processing.

To direct the discussion without focusing it too tightly, we set forth the following questions:

1. What are the most critical areas for the next seven (plus or minus two) years of natural language processing? ("Critical" is taken to mean that which will produce the greatest impact in the technology.)
2. What resources are needed (such as people, training, and corpora) to accomplish the goals involved in that work?
3. What organization is needed (e.g., coordinated efforts, international participation) to accomplish those goals?
4. What application areas and markets should open up in response to progress toward those goals?

The first question is addressed in Section 10.2 below. Questions 2 and 3 are combined in Section 3, and the final question is covered in Section 10.4.

https://doi.org/10.1017/CBO9780511659478.012 Published online by Cambridge University Press

## 10.2 Critical areas for work in NLP

Because the meeting was conducted as a brainstorming session, the reader should not interpret any of the suggestions as representing a unanimous view, or even a majority view. The following suggestions are presented in the order in which they were originally proposed:

*1. Grammars.* Most important are grammars that cover a broad range of language as measured by corpora of collected texts, not just a broad set of examples. These grammars should specifically facilitate the integration of speech and discourse. The group felt a strong need for broad coverage grammars in constrained formalisms that could be shared among organizations.

*2. Automatic methods to derive models of syntax, semantics, and probabilities; self-adapting systems.* Many domains have either idiosyncratic subgrammars or idiosyncratic lexical entities (such as part numbers consisting of two letters and three digits). In order to obtain useful systems, it will be necessary to move strongly away from hand-crafted knowledge bases such as grammars and semantic rules. Automatic methods of inferring knowledge bases (or of adapting previously existing knowledge bases) will have a large payoff for both theoretical and practical work. Methods that work with smaller, rather than larger, bodies of original data are to be preferred, but very little is currently known about the amount of data that will be required for various automatic methods to work effectively.

*3. Knowledge representation of sound structure.* This would be useful in incorporating prosodic information into the understanding process. For further information on this topic, see the chapter by Janet Pierrehumbert in this volume.

*4. Integration of speech and natural language.* Many participants felt that the development of practical speech interfaces is essential for the eventual success of NLP in the marketplace. Others felt that speech alone would not necessarily bring about either financial or technical success. Nonetheless, speech is an important part of language use and will certainly continue to be an important field of research.

*5. Methods for combining analytic and stochastic knowledge.* Both theoretical linguistics and computational linguistics have tended to focus on symbol manipulation mechanisms. Although there is no mathematical reason for these mechanisms to be devoid of statistical models, they almost invariably have been. Surely, some linguistic phenomena are rare whereas others are common. Even grammaticality and ungrammaticality are relative, not absolute. An appropriate hybrid of stochastic and knowledge-based techniques is likely to be more powerful than either alone.

*6. Ecological study of language.* This is the study of what people actually say in diverse contexts. For example, a mother unconsciously uses different vocabulary, syntax, and probably other linguistic attributes when talking to a one-year-old child than she does when talking with a six-year-old child; a person giving orders

https://doi.org/10.1017/CBO9780511659478.012 Published online by Cambridge University Press

uses different language than a person requesting information; a person trying to clarify a misunderstanding with another person will probably use different language than a person trying to clarify a similar misunderstanding with a machine. Very little is known about how to characterize these language differences, or about how to use those differences in building language understanding systems.

*7. Methodology.* Until quite recently, normal research technique revolved around in-depth study of some examples, and the development of an account of those examples with no analysis of how frequently or real those examples were compared to any other set. Rather than placing such heavy reliance on intuition, of late there has been a strong movement toward the analysis of corpora, that is, language that is collected in real or at least realistic settings. This is clearly a major step forward for the field, and it should be encouraged. One outcome of this process is that there is increased emphasis on the empirical evaluation of systems for NL processing using large corpora and knowledge bases.

*8. Intention in discourse.* By understanding more about how intention is conveyed in human discourse, we may make it possible for computers to infer better the intentions of their users, thus making the computer systems more helpful and the users more satisfied.

*9. Measuring utility of NL systems.* This is an attempt to focus on the ultimate users of systems that involve NL processing, to determine just exactly what benefit they derive from the system's NL capabilities. In some areas this is fairly clear (processing N messages per day X% faster by machine than by humans will save $Y per year, and will change the error rate by Z%), but in others, such as NL interfaces to databases, it is much harder to express and to quantify.

*10. Ways to evaluate accurately and meaningfully the effectiveness of NL systems.* This is clearly related to the previous point, but it is not identical. There are aspects of NL systems that developers need to quantify and evaluate whether or not they directly map into perceived benefits by the users. Evaluation should become an integral part of every NL system development, whether it is to be a product or a research demonstration system.

*11. Knowledge representation for NLP.* This may be considered an old topic, because it has been around virtually since the beginning of NL processing, but participants felt that a particular area that promises high payoff in the near future is the study of representation and reasoning about events.

*12. Dialogue phenomena.* Large bodies of text exhibit discourse structure, but conversations between two (or more) parties exhibit dialogue structure and phenomena that are quite different from what occurs in text. Clarification subdialogues abound. Fragments of language are produced frequently. Interruptions are common. New subjects are introduced in a variety of ways. Will computer systems have to be as facile as humans in order to carry on spoken dialogues? Probably not, but right now we do not know how much computer systems will have to be able to do, or even how to characterize it.

*13. Interaction phenomena for the human-computer interface.* Many people feel

https://doi.org/10.1017/CBO9780511659478.012 Published online by Cambridge University Press

that trying to manipulate language of the sort that people produce for one another is not only too hard, but also unnecessary. People will almost certainly modify their use of language when they are faced with a much less than human machine, but we know very little of what form this new interaction will take.

*14. Tools for coping with and analyzing large amounts of data.* Examples might include tools for corpus collection, corpus analysis and sharing, lexicon development, and evaluation. As more and more groups are working with larger and larger amounts of data, the need for good software tools becomes acute, and the potential for cross-fertilization of research work becomes enormous.

*15. Partial understanding.* How can one represent information that is only partially understood? What effect does this have on other components of a fully integrated system, such as reasoning, or the user interface? Under what circumstances is partial understanding sufficient? When and how should it be clarified?

This listing is very different from what would have been produced by a similar group a decade ago, yet there are some interesting similarities as well. The desire for grammars with broader coverage, for knowledge representation to support the reasoning that is an integral part of language understanding, for the inclusion of speech, and for systems capable of handling dialogues in a theoretically based way has been with us for much longer than a single decade.

But the widespread emphasis on dealing with large amounts of data is quite new, as are the pushes toward evaluation and partial understanding, the inclusion of techniques from other disciplines (such as the way probabilities are used in speech processing), and the desire to collect data.

## 10.3 Resources and organizations needed

Several efforts are already underway to facilitate the availability of linguistic data. A Consortium for Lexical Research has been started at New Mexico State University under the auspices of the Association for Computational Linguistics. It will serve as a repository for lexical resources, including both data and software.

The Data Collection Initiative of the Association for Computational Linguistics will distribute material at cost (that is, at the cost of the media used for distribution). The raw text corpora in this collection are not proprietary, nor are the tools that are being built to operate on them, but perhaps some of the databases, complement structures, etc., may be proprietary.

Sue Atkins reported that Oxford University Press, SRI, the University of Oxford, the British Library, the University of Lancaster, and the University of Cambridge are supporting this consortium, which will build a British National Corpus, and, in addition, build a database (and ultimately a machine-readable dictionary). The consortium is trying to get government matching funds for the project.

In Japan, the Electronic Dictionary Research Project (EDR) is a multi-year,

https://doi.org/10.1017/CBO9780511659478.012 Published online by Cambridge University Press

joint government/industry project to collect large corpora, labeling them morphologically, syntactically, and semantically.

The following additional needs were expressed:

### *1. Collections of spoken language, containing both the speech and the transcribed language*

There are several corpora of spoken language data that are being collected under the DARPA spoken language systems program. These corpora are available from the National Institute of Standards and Technology (NIST). Anyone interested in learning more about these corpora should contact either David Pallett, NIST, Technology Building Room A-216, Gaithersburg, MD 20899, email at dave@ssi.ncsl.nist.gov., or Mark Liberman at the Linguistic Data Consortium, 215–898–0464, 1dc@cis.upenn.edu.

### *2. Correct, complete, well-represented pronunciations in a machine-readable dictionary*

The problems involved in defining such a dictionary are great, starting with the fact that speech and natural language people have different ideas of what constitutes a word.

### *3. A morphological analyzer for English and other languages*

This should be a standard that all can use.

### *4. A formalism for input to a speech synthesizer*

This should permit the transfer of more than lexical information; it should allow, for example, the constituent structure of the sentence to be represented, so the synthesizer can use this information for intonation.

## 10.4 Markets and applications

The question addressed was "What markets will be available in a 5-year time frame?" That is, what applied technology will be ready to transfer to product development within that time?

The initial suggestions put forth, in the order they were proposed, are:

*1. Document checking.* This should incorporate grammar and style checking that is as accurate and as easy to use as today's spelling checkers.

*2. Data extraction from text.* This assumes that the kind of information to be extracted from free text can be specified quite tightly, e.g., sufficient to be added to a relational database.

*3. Speech production.* This means a next-generation language and speech syn-

https://doi.org/10.1017/CBO9780511659478.012 Published online by Cambridge University Press

thesizer that is capable of starting with a meaning representation (not a word or phoneme string), producing a word string, and then synthesizing speech whose quality is better than the current widely used DECTALK synthesizer from Digital Equipment Corporation. It is also important that this technology be easily ported to new domains.

4. *Language generation.* This could be used for summaries, report generation (e.g., generate a paragraph of text from a structured database, such as the summary of a medical or financial record of a patient/client from the last twenty-four-hour period), explanation, status monitoring, etc.

5. *Languages other than English.* In this area, improved machine assisted translation is possible. In addition, there is a need for automating non-translation tasks involving a second language, or the combination of a first and second language.

6. *Over-the-telephone information services.* This may or may not involve speech recognition or understanding. It may involve language generation.

7. *Information retrieval.* This is particularly needed for bodies of free text, as in tax law documents. Both the input and output of the IR system may have to deal with virtually unconstrained language.

8. *Help/advisor systems.* One example of this is over-the-phone help systems, which could be used for home appliance diagnosis and repair, simple income tax reporting questions, local directions, etc. These systems may involve an expert system, and may use both images and speech in their output.

9. *Spoken language systems.* If achievable, it will result in automatic dictation systems, voice control of devices, and highly interactive systems that are very cost effective because of the savings in human labor, and a world very different from what we are accustomed to!

After considerable discussion, it was decided that items 1, 3, and 4 could probably get private funding because of the potential for short-term return on investment.

On the other hand, problems 2, 5, and 7, although representing an enormous need, also require significant technical advances, and thus are less likely to attract private funding, and therefore a higher priority for government funding.

https://doi.org/10.1017/CBO9780511659478.012 Published online by Cambridge University Press