IMPROVING METHODOLOGY*
in
Natural Language Processing

William C. Mann
USC Information Sciences Institute
Marina Del Rey, California

SCOPE

This is a position paper on understanding and improving the current styles and methods of scientific work in the application of computers to texts composed of elements from human languages, such as stories, dialogues and sentences. It deals only with kinds of research in which acoustic issues are secondary or absent. It is written specifically to precede discussion at the Workshop on Technical Issues in Natural Language Processing.

There are various orientations toward value that tend to get assumed rather than discussed at this point. They need not conflict, but some selectivity is necessary. Very roughly, there is an orientation toward understanding and scientific knowledge, and there is an orientation toward application and practical use. Many people regard understanding as a nearly-necessary prerequisite to practical accomplishment. That´s the view in this paper, so we therefore concentrate on scientific values without denying the others.

There is a great diversity of activities that are carried out by recognizable methods, for which serious questions of methodology could be raised. There are tool-building and laboratory setup activities. We do not build linear accelerators or observatories, but we put large efforts into tools anyway. There are speculative and exploratory activities that influence the course of later, more formal work. Choice of phenomena to study is an absolutely crucial one of these activities. There are administrative activities for which methods are important. Staffing and seeking funds are also vital. All of these anticipate and support the creation of specific results and are vital to success.

The activities that produce the knowledge that keeps the work going are of a different kind. IT IS THESE CONSUMMATORY ACTIVITIES THAT I FOCUS ON HERE, TO THE EXCLUSION OF ALL THE OTHERS.

CONSEQUENCES OF METHODOLOGY CHOICE

We are currently at a crucial stage in the development of methodology, since we have a significant history of experience, but a great deal of remaining flexibility. For better or for worse, the methodological choices made in the next few years by our present leaders are likely to be with us for a very long time. The formal result-producing style that we adopt is particularly crucial for two reasons - first, because it ends up being the least flexible set of precedents, perhaps with the exception of basic presuppositions, and second, because it produces a strong final filtering effect on the results. The adoption of a statistical hypothesis evaluation framework leads to different kinds of results. Likewise, our formal approach will produce its own kind of results and inherent limitations. So, we must pay careful attention to our current style.

My general attitude is that current methods can be very significantly improved, and that doing so will have a very high payoff with benefits far beyond the improvements to present and contemplated efforts. The methods currently in use are under-examined and poorly understood, and traditions are still weak enough to allow changes. There are attractive alternatives to many common practices.

PRESENT ADVANTAGES

Of the great diversity of approaches to language, the process approach represented at the workshop is uniquely capable. The two key methodological problems in the study of language over the last 2,500 years or so have been the problem of rigor and the problem of complexity. The problem of rigor in the use of natural language led to formal logics and to Godel. The problem of complexity has led to various strong reductions on the general phenomena, with tools such as the Osgood Semantic Differential, or paired-associate tests. Sequential-order phenomena and individual use of language tend to get badly obscured.

Process theory approaches the problem of rigor with methods by which process specifications are made very explicit. It approaches the problem of complexity with computers, that can hold and make use of very large numbers of processes at once. The compatability and effective coverage of large collections of hypotheses can now actually be tested.

These are exciting, reorienting advantages that make me prefer the process approach to any other, to hold high hopes for its success, and to want it to be built on good foundations.

WHAT MAKES A DIFFERENCE?

What do we want out of our methodology? Three characteristics of a methodology are particularly important:

reliability
efficiency
integrative power

Reliability encompasses all of those things that make experiments trustworthy at face value, including repeatability, clarity

----

of definition and freedom from various kinds of circumstantial effects that might be responsible for success. Efficiency addresses the effort required to achieve particular results. (You don't plan to do basic genetics studies on elephants; you may prefer fruit-flies as subjects.) It deals not only with the costs of performing the work, but with support costs as well. Integrative power involves the scope of the theories, what diversity of phenomena they cover, what subtheories they coordinate, what kinds of investigations they facilitate.

In order to discuss current practices we need some representative example. The one here is deliberately simple and not identified with a particular development effort. However it is composed of elements that seem to be widely used.

EXAMPLE OF A NATURAL LANGUAGE PROJECT

Step 1: Select a phenomenon: CONTRADICTION

Step 2: Select an input form: ENGLISH SENTENCES

Step 3: Select an output form: ENGLISH SENTENCES THAT CONTRADICT THE INPUT SENTENCES

Step 4: Design and draft a program in the local language: MEGALISP

Step 5: Debug on examples of opportunity, selected to exercise the code.

Step 6: Publish: "CONTRADICTION IN NATURAL LANGUAGE" by Leader and Worker.

SOME STRENGTHS IN CURRENT PRACTICE

We should hold on to the distinctive strengths of our methods in any changes we plan. These strengths are generally direct classic consequences of the use of computers to hold models:

Complexity of data and theory is easy to accommodate.

Time sequences and dependencies are preserved.

A diversity of hypotheses can be applied and tested for consistency in each experiment.

All of these have to do with integrative power, and on this dimension we are, at least potentially, in very good shape.

SOME WEAKNESSES

We have some serious problems. Here are some recurring problems with the FORM of the work:

1. Single experiments often take years to execute.

2. The activity is often treated as programming and program documentation rather than science. The consequences are generally that the data are poorly identified and poorly chosen, the status of the programs as theory is not clear, the business of making clear theoretical claims is neglected, and the relevance of the activity to existing theories that are not programs is never established. The remainder of science is thus cut off, and left wondering whether we are into science at all.

3. The attempt to perform a general transaction, such as Sentence:Contradiction, strongly limits the complexity of the input that gets actually addressed, with the result that significant phenomena are missed. The effects of prior context, speakers' goals, tacit mutual knowledge of speaker and hearer are often attenuated by the attempt to be general.

4. The unit of production is a system. Whole systems are difficult to disseminate and difficult to judge as scientific hypotheses, and are not generally understood or appreciated by non-programming scientists.

5. Coping with ad-hocness is a problem: The system runs the examples, but what else it will do is unclear, or, the degree of tuning to the examples is unclear, or, the representativeness of the examples is unclear, or, the rightness of the answers is only established intuitively.

We have problems with the CONTENT of the work. There are many problems, which may be a healthy condition, but I want to attend to just one that seems to be otherwise.

In the common notion, a natural language is a scheme of communication that people use. The fact that a language is used to communicate has strong consequences. For example, as languages change, their adequacy for communication must be maintained.

The communication properties of language are being ignored in a wide variety of approaches, including processing approaches. Often, it is outside of the paradigmatic scope of the studies.

Communication deals with changing correspondences between the knowledge of one individual or system and the knowledge of another. It is more than relations between strings and strings, or relations between strings and generators of strings (syntax). It is more than relations between strings and a world or a data base (semantics). Communication involves two active processors, and an adequate theory of language will specify some consequences of that fact. By restricting the view to a single processor (or less), I suspect that we are cutting ourselves off from the organizing principles that produce the regularities that we are trying to study.

Some of the changes of style that I would suggest are implicit in the identifications of the problems cited above:

- Design clear data collection methods.

- State theoretical claims that are distinct from the programs. (The claims may still contain algorithms, of course.)

- Decommit from attempts to be general, except where an empirical demonstration of generality is included in the work.

- Shift from focus on systems to focus on algorithms.

- Do something to drastically shorten the period required to do single experiments.

Beyond these suggestions, the special advantages of case analysis should be considered.

CASE ANALYSIS AS THE BASIS FOR AN ALTERNATE PROCESSING METHODOLOGY

Case analysis as a basic scientific activity is an attractive alternative to the current methodology sketched above. How would it work?

STEPS IN A CASE-ANALYSIS-BASED DEVELOPMENT IN NATURAL LANGUAGE PROCESSING

Step 1: DATA ACQUISITION. Examples of real-world use of natural language are collected. Some are selected for detailed attention.

Step 2: PHENOMENON IDENTIFICATION: The data are annotated and scored for particular phenomena of interest. Data can be scored for several phenomena at once. Scoring is performed by people who understand the language and the circumstances of the data occurrence, and who are given explicit instructions on what to look for and how to annotate it. The result of this step is a Commentary on the data.

Examples:

a. Identify requests and judge whether they are fulfilled in running dialogue.

b. Identify repeated references to an object, action or idea in a document.

Step 3: CASE MODELING: Custom-build for this data, a new one-shot program that will take the data as input, and make entries into a simulated Hearer´s Memory. The program is the Model, and its "output" is its trace.

Step 4: MODEL EVALUATION: Compare the Commentary with the execution trace of the model. For each significant event identified in the Commentary, decide whether there was a correctly corresponding event in the model´s execution.

With suitable selections of phenomena for study, it is not hard to decide whether the program performed appropriately. However, a serious problem remains: a program for a single case can be entirely ad hoc. This is an advantage, in that it is certain beforehand that the program will run successfully, independent of the complexity of the phenomena. But the program may or may not have any long-term significance.

The program is composed of cooperating processes. Each process can be considered to be an over-specified hypothesis, over-specified because details such as the programming language are inessential to the corresponding functional claims about language.

VERIFICATION STEP: In order to meet the ad-hocness problem, these hypotheses must be verified by repeated application to a diversity of cases. The experiment steps cited above must be repeated, and their results compared. Inessential details (such as programming language and machine) may be changed, if desired, but the properties of the algorithms which form the basis for the theoretical claims of the work must be held constant.

The verified results are those algorithms that continue to work correctly, when their actions are judged against the Commentary, in model after model. These algorithms are the valuable ones both for practical application and for scientific knowledge.

ADVANTAGES OF CASE ANALYSIS

METHODOLOGY

Since the data acquisition step is first rather than nearly last, stronger claims can be made for the ability to model real-world phenomena. Having the data in hand is a strong guide to implementation.

Because phenomena identification is explicit, and proceeds from explicit instructions, the resulting theory has a clear operational interpretation. since it substitutes powerful hindsight for less-powerful anticipation.

There is better control on complexity and effort, since no claims are made for the generality of the whole systems that are built. The amount of data modeled can be controlled, and a diversity of data sources can be accommodated. There is strong control over the involvement of world-knowledge in models, since most of the particulars can be anticipated by looking at the data.

The method can also be controlled by choices about whether several phenomena will be modeled in a single model or several smaller models. The smaller models are simpler, but the single model exhibits the compatability of the parts and the consistency of the set of hypotheses.

This approach typically runs in a more data-driven, phenomena-responsive manner than a general system building approach. It avoids the situation in which system design is based on inadequate stereotypes of what might happen at the input. Programming can be more goal-directed as well, since the phenomena of interest have already been identified in the Commentary.

The problems of ad-hocness are treated explicitly, rather than being left to the suspicions of the journal readers. This facilitates representations of the degree and kinds of tests that the theories have had. (I suspect that for some current systems, many readers believe that they will only run the explanatory examples in the papers).

Finally, because of the close control and 20-20 hindsight of case analysis, more complex phenomena can be accommodated. In particular, communication between two non-identical human processors can be modeled.

AN ACTIVE EXAMPLE OF CASE MODELING METHODOLOGY

The Dialogue Process Modeling work at ISI is an active attempt to apply the ideas above with some embellishments, to real natural language processing problems. All of the recommendations are being used in identifiable ways. This work will be described in discussion at the conference as time permits.