13.4 Evaluation of Broad-Coverage Natural-Language Parsers

Ezra Black
Interpreting Telecommunications Laboratories, ATR, Kyoto, Japan

13.4.1 State of the Art

A *parser* for some natural language (English, Portugese, etc.) is a program that *diagrams* sentences of that language---that supplies for a given sentence a correct grammatical analysis, demarcating its parts (called *constituents*), labeling each, identifying the part of speech of every word used in the sentence, and usually offering additional information, such as the *semantic class* (e.g., Person, Physical Object) of each word and the *functional class* (e.g., Subject, Direct Object) of each constituent of the sentence. A *broad-coverage parser* diagrams *any* sentence of some natural language, or at least agrees to attempt to do so.

Currently the field of broad-coverage natural-language parsing is in transition. Rigorous, objective and verifiable evaluation procedures have not yet become established practice, although a beginning has been made. Until recently, objective evaluation essentially was not practiced at all, so that even the author of a parsing system had no real idea how accurate, and hence how useful, the system was. In 1991 the Parseval system for *syntactically* evaluating broad-coverage English-language parsers was introduced [BAF+91,HAB+91], and the next year seven creators of such parsers applied Parseval to their systems, all using the same test data [BGL93].

However, the Parseval evaluation routine is an extremely coarse-grained tool. For one thing, most of the information provided by a parse is not taken into account. But more importantly, the level of agreement on the particulars of linguistic description is fairly superficial among the creators of Parseval, and a fortiori among parsing-system authors who could or would not be included in the Parseval planning sessions. Consequently, parsers are evaluated by Parseval at a high remove from the actual parses being judged, and in terms rather foreign to their own vocabulary of linguistic description.

Currently there are plans to extend Parseval into the *semantic* realm, via Semeval, an approach to evaluation modeled on Parseval (see [Moo94]). But there is *more, not less* disagreement among professionals regarding the proper set of semantic categories for text, the various word senses of any given word, and related semantic issues, than there is about constituent boundaries. So Semeval can be expected to turn out even rougher-grained than Parseval.

13.4.2 Improving the State of the Art

The methodology of objective, rigorous, and verifiable measurement of performance of individual parsing systems is known, albeit by only a minority of practitioners. Key features of this methodology are the use of:

1. *separate* training and test sets;
2. test data from *new* documents only;
3. *large* test sets;
4. responsible *public access* to the test process;
5. *objective* criteria of evaluation;
6. the statement, in advance, of *all* acceptable analyses for a test item;
7. test runs on a *variety* of test materials to match the sort of claims being made for the system; and
8. at least a *twice-yearly* run of a full range of public tests.

A slow transition is now taking place within the field towards the recognition of the value, and even the necessity, of rigor of the above sort within evaluation. This kind of testing is necessary anyway for effective parsing-system development, as opposed to the onerous activities associated with testing via *compromise-based* tools such as Parseval, Semeval, or others. It may never be possible to compare *all* broad-coverage parsers of a given language in terms of a *common coin* of linguistic analysis. Instead, practitioners will probably want to opt for highly accurate and rigorous performance statistics on their own systems alone, rather than extremely coarse-grained scores obtained from comparing their systems with others on the basis of laborious and even dubious technical compromise.

Another progressive development has been the appearance since 1992 of parsing systems which parse previously-unseen text without referring to a set of grammar rules, by processing, statistically or logistically, a *treebank* or set of sentences parsed correctly by hand by competent humans [Bla93]. These systems are in theory directly comparable, and can employ more rigorous correctness criteria---e.g., exact match of the treebank parse---than can Parseval.

13.4.3 Future Directions

The remainder of the 1990s will probably see two major trends in this area. First should be a move toward the sort of rigor discussed above, when individual systems are evaluated either just to let the system developer himself or herself know the rate at which and the manner in which the system is improving over time, or else for the purpose of cross-system comparisons on a given document, where this is possible (see above). Second should be a move away from evaluating parsing systems in linguistic terms at all, i.e., away from judging the parses output by a system simply on their merits as parses. This move would be *toward* evaluating a parser on the basis of the *value added* to a variety of *client systems*. These would be bona fide, fully-developed AI systems of one sort or another, with a need for a parsing component. This as opposed to tasks conceived artificially, simply for the purposes of providing a *task* to support evaluation. Examples might be pre-existing systems for speech synthesis, speech recognition, handwriting recognition, optical character recognition, and machine translation. In this case the evaluation of a broad-coverage parsing system would come to be based on its performance over a *gamut* of such applications.