Evaluation of NLP systems

Coordinator: Bente Maegaard
Center for Sprogteknologi
DK-2300 Copenhagen S
E-mail:bente@cst.ku.dk

Computational linguistics as a science has had its evaluation methods since its early days: A concordance program can be evaluated according to its ability to find all occurrences, to list them properly, to have a flexible user interface etc., frequency programs may be evaluated according to their statistics, the possibility of lemmatisation, parsers are evaluated according to their efficiency etc. When we contemplate one component at a time and want a technical evaluation, we normally have no problem defining the evaluation criteria.

But once we get into more complicated systems: machine translation, dialogue systems, message understanding systems etc., the issue of evaluation becomes more complex. The most important reason for this is that a technical evaluation of the functionality of a component is no longer sufficient. First of all there are a number of components that have to function together, so the technical functionality of the whole system becomes an issue. Secondly, and more importantly, the performance of the whole system has to be assessed in itself. This performance cannot be calculated by the evaluation of the components.

In the United States, this insight has been used in a series of competitions (MUC, TREC) between research teams, where the system's performance on a specific task was the only element that counted. There are various other ways to evaluate research results, - the beauty of the algorithm, the methodology, the simplicity, the efficiency, the ability to explain linguistic facts, the ability to imitate human behaviour etc. The evaluation of research results, be they single components or whole systems, is one of the strands of the panel discussion.

When we turn to the application of research results, as in the type of language technology systems mentioned above: machine translation, dialogue systems for practical use etc., the real assessment is made by users in their application of the system. Is the system useful for the user? Is it fast? Reliable? How much of the job. e.g. translation, can it do in average? This raises questions about how this type of information can be measured. The assessment of language technology tools and its similarity with and/or difference from the evaluation of research results is the other strand of the panel discussion.

It is very important for the field to agree upon standard methods for evaluation and much work has been done in this field in recent years. The panellists will present their experience and their views and we hope to be able to provoke a discussion with the audience.

References

Galliers, J.R and K. Sparck Jones, 1993. Evaluating Natural Language Processing Systems, University of Cambridge, Cambridge (to appear in a printed version).

Grishman, R. and B. Sundheim, 1996. Message Understanding Conference - 6: A Brief History. COLING-96 Proceedings, Copenhagen.

Hendry, D.G. and T.R.G. Green, 1993. Spelling Mistakes: How well do correctors perform? Adjunct Proceedings of InterCHI'93.

Information Technology - Software product evaluation - Quality characteristics and guidelines for their use, 1991. ISO/IEC 9126, Geneva.

King, M., 1987. Machine Translation Today, Edinburgh Information Technology Series 2, Edinburgh University Press, Edinburgh.

TEMAA, A Testbed Study of Evaluation Methodologies: Authoring Aids, 1996. Final report, Center for Sprogteknologi, Copenhagen.