Last Words

That's Nice . . . What Can You Do With It?

Anja Belz*
University of Brighton

A regular fixture on the mid 1990s international research seminar circuit was the "billion-neuron artificial brain" talk. The idea behind this project was simple: in order to create artificial intelligence, what was needed first of all was a very large artificial brain; if a big enough set of interconnected modules of neurons could be implemented, then it would be possible to evolve mammalian-level behavior with current computational-neuron technology. The talk included progress reports on the current size of the artificial brain, its structure, "update rate," and power consumption, and explained how intelligent behavior was going to develop by mechanisms simulating biological evolution. What the talk didn't mention was what kind of functionality the team had so far managed to evolve, and so the first comment at the end of the talk was inevitably "nice work, but have you actually done anything with the brain yet?"[1]

In human language technology (HLT) research, we currently report a range of evaluation scores that measure and assess various aspects of systems, in particular the similarity of their outputs to samples of human language or to human-produced gold-standard annotations, but are we leaving ourselves open to the same question as the billion-neuron artificial brain researchers?

Shrinking Horizons

HLT evaluation has a long history. Spärck Jones's *Information Retrieval Experiment* (1981) already had two decades of IR evaluation history to look back on. It provides a fairly comprehensive snapshot of HLT evaluation at the time, as much of HLT evaluation research was in the field of IR. One thing that is striking from today's perspective is the rich diversity of evaluation paradigms—user-oriented and developer-oriented, intrinsic and extrinsic[2]—that were being investigated and discussed on an equal footing

* NLTG, University of Brighton, Lewes Road, Brighton BN2 4GJ, UK. E-mail: A.S.Belz@brighton.ac.uk.

1 To which the answer was, in effect, "the brain isn't big enough yet to be used." The original aim of the CAM-Brain Project was to evolve behavior as complex as that of a kitten (the brain was going to control a robotic kitten, the "Robokoneko"). The functionality reported for the modules the brain was composed of was on the level of the XOR-function (de Garis et al. 1999). To date, no functionality appears to have been reported for the brain as a whole.

2 **User-oriented** evaluations (covered by ISO standards 9126 and 14598 on software evaluation) look at a set of requirements (available computational, financial, and other resources, acceptable processing time, maintenance cost, etc.) of the user (embedding application or person) and assess how well different technological alternatives fulfill them. **Developer-oriented** evaluations focus on functionality (just one component in the ISO standards) and seek to assess the quality of a system's (or component's) outputs. The user-oriented vs. developer-oriented distinction concerns evaluation purpose. Another common distinction is about evaluation methods: **intrinsic** evaluations assess properties of systems in their own right, for example, comparing their outputs to reference outputs in a corpus, whereas **extrinsic** evaluations assess the effect of a system on something that is external to it, for example, the effect on human performance at a given task or the value added to an application (Spärck Jones 1994).

in the context of academic research. At the same time Spärck Jones described a lack of consolidation and collective progress, noting: "there is so little control in individual tests and so much variation in method between tests that interpretations of the results of any one test or of their relationships with those of others must be uncertain" (page 245).

These days, HLT research has many more subfields, most of which devote substantial research effort to evaluation. We have far more established evaluation techniques and comparative evaluation is the norm. In fact, virtually all HLT subfields now have some form of *competitive* evaluation.[3] But it seems we have achieved comparability at the price of diversity. The range of evaluation methods we employ has shrunk dramatically. Not only is virtually all evaluation in HLT research now developer-oriented and intrinsic, but, even more narrowly, most of it is a version of one of just three basic intrinsic techniques: (i) assessment by trained assessors of the quality of system outputs according to different quality criteria, typically using rating scales; (ii) automatic measurements of the degree of similarity between system outputs and reference outputs; and (iii) human assessment of the degree of similarity between system outputs and reference outputs.[4]

What is noticeable by its absence is any form of extrinsic evaluation. Application purpose—of the embedded component or end-to-end system—is not part of task definitions, and we do not test how well components or systems fulfill (some aspect of) the application purpose.

Tasks in Need of an Application

Because application purpose does not figure in it, the intrinsic evaluation paradigm treats tasks as generic even though this may not always be appropriate. Kilgarriff warned against treating the word sense disambiguation (WSD) task as generic right at the start of the SENSEVAL evaluation competitions:

> [...] a task-independent set of word senses for a language is not a coherent concept. [...] Until recently, WSD researchers have generally proceeded as if this was not the case: as if a single program—disambiguating, perhaps, in its English-language version, between the senses given in some hybrid descendant of Merriam-Webster, LDOCE, COMLEX, Roget, OALDCE and WordNet—would be relevant to a wide range of NLP applications. (Kilgarriff 1997, page 107)

WSD may still be the most notorious "task in need of an application" (McCarthy and Navigli 2007), but the case of WSD points to a more general issue: in intrinsic evaluations, the output representation formalism (e.g., tag set, syntactic formalism) is fixed in the form of gold-standard reference annotations, and alternative representations are not subject to evaluation. There is evidence that it may be worth looking at how different representations perform. For example, Miyao et al. (2008) found significant differences between different parse representations given the same parser type when evaluating their effect on the performance of a biomedical IR tool. The intrinsic set-up makes it impossible to perform such evaluations of alternative representations, because this

3 Some examples are the NIST-run DUC document summarization evaluation campaign (now part of TAC), the NIST-run Open MT evaluation campaign (MT-Eval), and the academic-led SENSEVAL/SEMEVAL WSD evaluations, among many others.

4 All three techniques have been used in competitive evaluations: *i* and *iii* have been used, for example, in DUC; *ii* in MT-Eval, DUC, and SENSEVAL/SEMEVAL. By far the most common technique to be found in individual research reports is *ii*, although *iii* and related types of user assessments are also used.

requires an external—extrinsic—point of reference, as is provided by an embedding system like the IR tool in Miyao et al.'s work.

If we don't include application purpose in task definitions then not only do we not know which applications (if indeed any) systems are good for, we also don't know whether the task definition (including output representations) is appropriate for the application purpose we have in mind.

A Closed Circle

Whereas in analysis tasks evaluation typically measures the similarity between system output representations and gold-standard reference representations, in tasks where the output is language (e.g., MT, summarization, data-to-text generation), system outputs are compared to human-produced reference texts, or directly evaluated by assessors. Methods for evaluating these evaluation methods, or "meta-evaluation" methods, look in particular at the reference outputs and similarity measures they involve. In analysis, where there are single target outputs, and similarity measures are a matter of counting matching brackets or tags, we can't do much more than assess inter-annotator agreement and perform error analysis for reference annotations. In generation, it is the similarity measures that are scrutinized most. Metrics such as BLEU and ROUGE were conceived as surrogate measures[5] (the U in BLEU stands for 'understudy'). Surrogate measures in science in general need to be tested in terms of their correlation with some reliable measure which is known to be a true indicator of the condition or property (e.g., karyotyping for chromosomal abnormalities) for which the surrogate measure (e.g., serum testing for specific protein types) is intended to be an approximate indicator. In HLT, we test (surrogate) automatic metrics in terms of their correlation with human ratings of quality, using Pearson's product-moment correlation coefficient, and sometimes Spearman's rank-order correlation coefficient (Lin and Och 2004). The stronger and more significant the correlation, the better metrics are deemed to be. The human ratings are not tested.

In this set-up, clearly, there is no way in which the results of human quality assessment can ever be shown to be wrong. If human judgment says a system is good, then if an automatic measure says the system is good, it simply confirms human judgment; if the automatic measure says the system is bad, then the measure is a bad one, its results are disregarded, and the system is still a good system. This is a classic closed-circle set-up: It isn't falsifiable, and it doesn't include a scenario in which it would be concluded that the initial theory was wrong. The problem lies with treating what is but another surrogate measure—human quality ratings—as a reliable, objective measure of quality. We may be justified in not accepting contradicting metric scores as evidence against human quality judgments and humanlikeness assessments, but perhaps we should pay attention if such intrinsic measures are contradicted by the results of user-performance and other extrinsic experiments. For example, in a comparison of graphical representations of medical data with textual descriptions of the same data, Law et al. (2005) found that, whereas in intrinsic assessments doctors rated the graphs more highly than the texts, in an extrinsic diagnostic performance test they performed better with the texts than the graphs. Engelhardt, Bailey, and Ferreira (2006) found that subjects rated

5 This term is more commonly used in biology where it refers to a laboratory measurement of biological activity within the body that indirectly indicates the effect of treatment on disease state. For example, CD4 cell counts and viral load are examples of surrogate markers in HIV infection. ⟨http://cancerweb.ncl.ac.uk/omd⟩.

over-descriptions as highly as concise descriptions, but performed worse at a visual identification task with over-descriptions than with concise descriptions. In a recent set of evaluation experiments involving 15 NLG systems, the eight intrinsic measures tested (although correlating strongly and positively with each other) either did not correlate significantly with the three extrinsic measures of task performance that were also tested, or were negatively correlated with them (Belz and Gatt 2008). In parsing, Miyao et al. (2008) performed an extrinsic evaluation of eight state-of-the-art parsers used as part of a biomedical IR tool. The effect parsers had on IR quality revealed a different system ranking than the WSJ-Corpus based F-scores reported for the same parsers elsewhere.

Unreliable Evidence?

There is some indication that human quality judgments and measurements of similarity with human-produced reference material may not be able to live up to the role they are currently assigned. We know that agreement between annotators is notoriously difficult to achieve, particularly at the more subjective end of the spectrum (see, for example, Reidsma and op den Akker 2008). Stable averages of human quality judgments, let alone high levels of agreement, are hard to achieve, as has been observed for MT (Turian, Shen, and Melamed 2003; Lin and Och 2004), text summarization (Trang Dang 2006), and NLG (Belz and Reiter 2006). In fact, the large variance typical of human quality judgments can result in higher agreement between automatic metrics and human judges than among the human judges (Burstein and Wolska 2003; Belz and Reiter 2006).

Despite the evidence, it is hard to shake the assumption that similarity to "how humans do it" is an indicator of quality, that the more similar HLT system outputs are to human outputs the better they are. In fact, to some it is a matter of a priori fact that humans cannot be outperformed in HLT by machines:

> [...] no-one (surely) would dispute that human performance is the ultimate criterion for automatic language analysis.
>
> To draw an analogy with another area of computational linguistics, namely machine translation, it would not make sense to claim that some MT system was capable of translating language A into language B *better* than the best human translators for that language-pair: skilled human performance logically defines an upper bound for machine performance. (Sampson and Babarczy 2003, page 63)

Clearly, there do exist low-level HLT tasks that machines can perform faster and more accurately than humans (e.g., concordance construction, spell checking, anagram finding). But there is some evidence that there are more complex HLT tasks for which this is the case, too. In genre classification, even ad hoc systems have matched human performance (Jabbari et al. 2006). In NLG, domain experts have been shown to prefer system-generated language to alternatives produced by human experts (Reiter et al. 2005; Belz and Reiter 2006). In WSD, Ide and Wilks have pointed out that "claimed and tested success rates in the 90%+ range are strikingly higher than the inter-annotator agreement level of 80%+, and to some this is a paradox"; they conclude that the only explanation that seems to fit the data is that the average tested person is not as good as the tested systems at this task (Ide and Wilks 2006, page 52).

Limited Conclusions

So, current HLT evaluation practices involve a limited number of basic evaluation techniques capable of testing for a limited range of system characteristics; because

they do not involve a system-external perspective we can't test systems for suitability for application purpose and we can't effectively meta-evaluate evaluation procedures; instead, we have to rely heavily on human judgments and annotations that we know to be unreliable in many cases. In addition, we tend to evaluate systems on a single corpus (failing to make use of one way in which an extrinsic perspective could be introduced into an intrinsic evaluation set-up). In this situation only very limited conclusions can be drawn from evaluations. When large companies with corporate lawyers are among the participants of an HLT competition, this fact must be made explicit in a prominently displayed formal disclaimer:

> The data, protocols, and metrics employed in this evaluation [...] should not be construed as indicating how well these systems would perform in applications. While changes in the data domain, or changes in the amount of data used to build a system, can greatly influence system performance, changing the task protocols could indicate different performance strengths and weaknesses for these same systems. (Disclaimer, NIST Open MT Evaluation 2005[6])

Prevailing evaluation practices guide the development of an entire field of research; flagship shared-task evaluation competitions such as MT-Eval, DUC, SEMEVAL, and CoNLL are regarded as determining the state of the art of a field—should we not expect such evaluations to give some indication of "how well [...] systems would perform in applications"?

Towards a More Extrinsic Perspective

The explanation routinely given for not carrying out extrinsic evaluations is that they are too time-consuming and expensive. There clearly is a need for radical innovation in this area, and industry involvement and crowd-sourcing may provide ways to offset cost. But there are things we can do now, even with limited budgets. For example, automatic extrinsic evaluations are possible, and avoid the cost of human participants: Kabadjov, Poesio, and Steinberger (2005) tested an anaphora resolver embedded in a summarization system; Miyao et al. (2008) evaluated a parser embedded within an IR system. Even evaluation experiments involving human subjects do not have to come with an exorbitant price-tag: REG'08, a competition in the field of referring expression generation which had very minimal funding, included a task-performance experiment in which the speed and accuracy with which subjects were able to identify intended referents was tested (Gatt, Belz, and Kow 2008).

Perhaps the most immediately feasible way to bring an extrinsic perspective into current HLT evaluation practices is to combine methods for *extrinsic validation* with current intrinsic techniques. What makes extrinsic evaluation infeasible in many cases is not the cost of a single experiment, but the fact that the experiment has to be repeated for every data set and for every set of systems, and that the cost of the experiment is the same every time it is run. In contrast, extrinsic validation involves one-off validation procedures for evaluation metrics, reference material, and training data. Because they are one-off experiments that form part of the development of evaluation methods and data resources, they can be achieved at a much lower cost than extrinsic evaluation

6 http://www.nist.gov/speech/tests/mt/2005/doc/mt05eval_official_results_release_20050801_v3.html.

methods that are directly applied to systems. Extrinsic validation can potentially take many different forms; the following are three examples:

1. Extrinsic meta-evaluation of evaluation metrics: Evaluation methods, in particular automatic metrics, are evaluated in terms of their correlation with user-performance and other application-specific evaluations.

2. Extrinsic evaluation of human-produced reference material: The quality of reference material is assessed by testing it directly in user-performance/application-specific evaluations. Intrinsic evaluation techniques using the reference material can then be weighted in favor of more highly scoring material.

3. Extrinsic evaluation of training data: Training data is annotated with information from extrinsically motivated experiments, for example, reading speed or eye-tracking information, or scores obtained in other user-performance/application-specific evaluations. Training procedures can then be weighted in favor of more highly scoring material.

Conclusions

Science and technology is littered with the remains of intrinsic measures discarded when extrinsic measures revealed them to be unreliable indicators.[7] The billion-neuron brain researchers pursued an intrinsic measure (size) without testing it against the corresponding extrinsic measure (improved functionality), and ended up with an artificial brain that was very, very large, but was of little actual use (and certainly didn't fulfill its declared application purpose of controlling a robot kitten).

In HLT we are currently enthusiastic about evaluation to the point where it is hard to get a paper into ACL or COLING that doesn't have evaluation results; at the same time we consider tables of metric scores on single data sets a meaningful form of evaluation. If we think that, say, the purpose of a parser is to place brackets in text where a human annotator thinks they should go, then we're doing fine. If we think it is to facilitate high-quality IR, NER, and similar tasks, then we need to evaluate extrinsically, with reference to a range of application contexts (if the tool is intended to be a generic one) or a specific application context (if it is a specialized tool); then we need to stop picking the low-hanging fruit, and instead put our energies into solving the hard problem of how to situate evaluation in a context:

> One of the main requirements for future NLP evaluations is thus to approach these in a comprehensive as well as systematic way, so that the specific tests done are properly situated, especially in relation to the ends the evaluation subject is intended to serve, and the properties of the context in which it does this. (Sparck Jones 1994, page 107)

Putting greater emphasis on the extrinsic perspective in HLT research will result in improved checks and balances for the evaluation methods we apply; it will enable us to

7 Some examples: (i) The drug cholestyramine successfuly reduces blood cholesterol levels, but was found to have no impact on overall mortality rate (Le Fanu 1999, page 341); (ii) cardiac ac/decleration and raised blood acidity were thought to be indicators of an increased risk of cerebral palsy, but the introduction of foetal monitoring had no effect on the incidence of cerebral palsy, which remained constant (Le Fanu 1999, pages 255–258); (iii) commonly used in adverts by financial institutions and as the basis for investor decisions, past performance has been shown not to be a reliable indicator of the future performance of mutual funds (Allen et al. 2003); (iv) total harmonic distortion was thought to be an overall indicator of amplifier quality until the 1970s when, following technological improvements in audio production, it was found not to correlate with expert listeners' assessments at improved levels.

make better predictions about how the methods we develop will perform when applied to the purpose we develop them for; and it will mean that we have a better answer when we are asked ". . . but what can you do with it?"

References

Allen, D., T. Brailsford, R. Bird, and R. Faff. 2003. A review of the research on the past performance of managed funds. ASIC REP 22, Australian Securities and Investment Commission.

Belz, A. and A. Gatt. 2008. Intrinsic vs. extrinsic evaluation measures for referring expression generation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 197–200, Columbus, OH.

Belz, A. and E. Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06)*, pages 313–320, Trento, Italy.

Burstein, J. and M. Wolska. 2003. Toward evaluation of writing style: Overly repetitious word use. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03)*, pages 35–42, Budapest.

de Garis, H., N. Eiji Nawa, A. Buller, M. Korkin, F. Gers, and M. Hough. 1999. ATR's artificial brain ('CAM-brain') project. In *Proceedings of the 1st Genetic and Evolutionary Computation Conference (GECCO'99)*, volume 2, page 1233, Orlando, FL.

Engelhardt, P., K. Bailey, and F. Ferreira. 2006. Do speakers and listeners observe the Gricean maxim of quantity? *Journal of Memory and Language*, 54:554–573.

Gatt, A., A. Belz, and E. Kow. 2008. The TUNA Challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Natural Language Generation Conference (INLG'08)*, pages 198–206, Salt Fork, OH.

Ide, N. and Y. Wilks. 2006. Making sense about sense (Chapter 3). In E. Agirre and P. Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*. Springer, Berlin, pages 47–74.

Jabbari, S., B. Allison, D. Guthrie, and L. Guthrie. 2006. Towards the Orwellian nightmare: Separation of business and personal emails. In *Proceedings of the Joint 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL'06)*, pages 407–411, Sydney, Australia.

Kabadjov, M. A., M. Poesio, and J. Steinberger. 2005. Task-based evaluation of anaphora resolution: The case of summarization. In *Proceedings of the RANLP'05 Workshop on Crossing Barriers in Text Summarization Research*, pages 18–25, Borovets, Bulgaria.

Kilgarriff, A. 1997. "I don't believe in word senses." *Computers and the Humanities*, 31:91–113.

Law, A. S., Y. Freer, J. Hunter, R. H. Logie, N. McIntosh, and J. Quinn. 2005. A comparison of graphical and textual presentations of time series data to support medical decision making in the neonatal intensive care unit. *Journal of Clinical Monitoring and Computing*, 19:183–194.

Le Fanu, J. 1999. *The Rise and Fall of Modern Medicine*. Abacus, London.

Lin, C.-Y. and F. J. Och. 2004. ORANGE: A method for evaluating automatic evaluation metrics for machine translation. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, pages 501–507, Geneva.

McCarthy, D. and R. Navigli. 2007. Semeval-2007 Task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval'07)*, pages 48–53, Prague.

Miyao, Y., R. Saetre, K. Sagae, T. Matsuzaki, and J. Tsujii. 2008. Task-oriented evaluation of syntactic parsers and their representations. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL'08)*, pages 46–54, Columbus, OH.

Reidsma, D. and R. op den Akker. 2008. Exploiting 'subjective' annotations. In *Proceedings of the COLING'08 Workshop on Human Judgements in Computational Linguistics*, pages 8–16, Manchester, UK.

Reiter, E., S. Sripada, J. Hunter, and J. Yu. 2005. Choosing words in computer-generated weather forecasts. *Artificial Intelligence*, 167:137–169.

Sampson, G. and A. Babarczy. 2003. Limits to annotation precision. In *Proceedings of the EACL'03 Workshop on Linguistically Interpreted Corpora (LINC'03)*, pages 61–89, Budapest.

Spärck Jones, K. 1981. Retrieval system tests 1958–1978 (chapter 12). In K. Spärck Jones, editor, *Information Retrieval Experiment*. Butterworth & Co, pages 213–255, Newton, MA.

Spärck Jones, K. 1994. Towards better NLP system evaluation. In *Proceedings of the Workshop on Human Language Technology (HLT'94)*, pages 102–107, Plainsboro, NJ.

Trang Dang, H. 2006. DUC 2005: Evaluation of question-focused summarization systems. In *Proceedings of the COLING-ACL'06 Workshop on Task-Focused Summarization and Question Answering*, pages 48–55, Prague.

Turian, J., L. Shen, and I. D. Melamed. 2003. Evaluation of machine translation and its evaluation. In *Proceedings of MT Summit IX*, pages 386–393, New Orleans.