Reference-based vs. task-based evaluation of human language technology

Andrei Popescu-Belis

IDIAP Research Institute
Av. des Prés-Beudin 20, PO Box 592
CH-1920 Martigny, Switzerland
andrei.popescu-belis@idiap.ch

Abstract

This paper starts from the ISO distinction of three types of evaluation procedures – internal, external and in use – and proposes to match these types to the three types of human language technology (HLT) systems: analysis, generation, and interactive. The paper explains why internal evaluation is not suitable to measure the qualities of HLT systems, and shows that reference-based external evaluation is best adapted to 'analysis' systems, task-based evaluation to 'interactive' systems, while 'generation' systems can be subject to both types of evaluation. In particular, some limits of reference-based external evaluation are shown in the case of generation systems. Finally, the paper shows that contextual evaluation, as illustrated by the FEMTI framework for MT evaluation, is an effective method for getting reference-based evaluation closer to the users of a system.

1. Introduction

The nature of the evaluation methods that can be applied to human language technology (HLT) systems depends on the type of such systems, and more specifically on the place of language among their inputs and outputs. This paper considers the three types of evaluation synthesized in the ISO/IEC 9126 and 14598 standards – internal, external, and in use – and attempts to match them to an I/O-based typology of HLT – analysis, generation or interactive systems.

We argue first that internal evaluation cannot significantly capture the quality of HLT systems (Section 2). Then, we show that 'analysis' systems are naturally submitted to reference-based external evaluation (Section 5), while for 'generation' systems reference-based and task-based evaluation have respective advantages and drawbacks, mainly a trade off between informativeness and cost (Section 6). We also pinpoint the potential risk of training a system for higher score on a specific metric, regardless of its overall quality (Section 7). For interactive systems, the only feasible evaluation appears to be the task-based one, which can be carried out in more or less realistic settings (Section 8).

Finally, we argue that adapting reference-based evaluation to the intended context of use of a system – as in the FEMTI guidelines for context-based MT evaluation – is a way to get reference-based evaluation closer to the conclusions of task-based evaluation, for a smaller cost (Section 9).

2. Types of evaluation according to ISO

The ISO/IEC standards for software evaluation, under the 9126 and 14598 series and then the SQuaRE framework (Azuma, 2001), have defined *software quality* as the "features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs" (ISO/IEC, 2001 : p. 11). According to ISO/IEC 14598-1 (1999 : p. 12, fig. 4) the software life cycle starts with an analysis of the user needs, determining a set of *external quality requirements*, which are then transformed into *internal* ones during the development phase. Once a system is implemented, it becomes possible to assess its internal quality (without running it) and the external quality (using the results of external metrics obtained by running the system as a black box), and finally its *quality in use*, i.e. the extent to which it really helps users fulfil their tasks (ISO/IEC, 2001 : p. 11). Quality in use is often expressed in terms of effectiveness, efficiency, user satisfaction and safety (ISO/IEC, 2004) while internal and external qualities belong in six categories: functionality, reliability, usability, efficiency, maintainability and portability.

According to ISO/IEC, quality in use does not follow automatically from external quality, as it is not possible to predict all the results of using the software before it is operational in its intended context of use. However, in many cases of HLT evaluation, it is the qualities under "functionality" that are the focus of evaluation.

3. An I/O typology of HLT systems

In order to study the most adapted evaluation techniques for HLT systems, we propose to classify the systems according to the occurrence of language in their input and/or output data: in the input to a system, in its output, or in both. Additionally, the system may or may not require an interaction with a human user in order to produce its global results.

Type A systems, for 'analysis' or 'annotation' have language as an input only – most often they perform classification of the linguistic material into a small number of categories (e.g. POS tagging, WSD, or even anaphora resolution). Type G systems, for 'generation', have language only as an output (e.g. generating weather reports from non-linguistic data). Type AG systems have both linguistic input *and* output (e.g. machine translation, automatic summarization, or question answering). Finally, type I systems, or more accurately type AGI, are language-based human-computer dialogue systems. This classification appears to be exhaustive and non-ambiguous, as shown elsewhere (Popescu-Belis, 2008) by analyzing the HLT domains and applications from two encyclopaedias of HLT and NLP (Dale, Moisl & Somers, 2000; Mitkov, 2003).

The correct result of a type A system can generally be defined by a unique ground truth or gold standard annotation, possibly accompanied by an estimate of its reliability, if human judges agree less than perfectly upon a gold standard. In the case of G or AG systems, it is however impossible to find a unique gold standard, or to enumerate all acceptable results, due to the variability of natural language. In this case, it is still possible to provide a sample of the set of acceptable results, produced by human subjects; or, given the output of a system, a human judge can decide whether it belongs or not to the ground truth, i.e. whether it is a "perfect" answer.

4. Types of evaluation vs. types of systems

Following the preliminary definitions above, the main point of this paper is to discuss whether some of the three ISO-based types of evaluation are better suited to some of the types of HLT systems described in the previous section. In principle, according to ISO, all types of HLT systems can (and should) undergo all types of evaluation, at different stages of their development lifecycles. However, this is clearly not feasible in the HLT community. More precisely, we will argue that the following rules characterize best practice in HLT evaluation:

- *internal* evaluation is not enough informative for HLT systems, as it cannot predict external and in use qualities;
- for type A systems, *external* evaluation using ground-truth data is informative and cost-effective;
- for type G and AG systems, there is a trade-off in informativeness vs. cost when switching from reference-based *external* evaluation to evaluation *in use* or *task-based*;
- for type I systems, only evaluation in use is informative enough, and can take place in more or less idealized conditions.

The first point is justified by the observation that the behaviour of nearly all HLT systems cannot be reliable predicted from their internal properties, unlike more deterministic software. Linguistic problem solving is most often based on heuristics that show no clear relation between internal properties and external performance: e.g., for a parser, the amount of syntactic rules is only marginally correlated with parsing accuracy or coverage. Of course, some generic qualities such as portability can be measured internally, but such qualities are seldom the focus of HLT evaluation, which generally focuses on functionality, i.e. the capacity to perform an intended linguistic function. In addition, *speed* is sometimes taken into account as well, but again it is generally not measured using internal metrics.

5. Evaluation of type A systems: importance of reference-based external metrics

The linguistic functionality of 'annotation' systems is most often measured by comparing their results to ground truth annotations produced by human judges. Such reference-based external metrics are generally expressed as (pseudo)distances between a system's response on some test data and the expected response or set of responses, as defined by human judges, and are generally computed automatically. Whether or not the set can be determined with enough precision is a problem related not to HLT, but to the study of the respective linguistic capacity in human subjects.

This of course does not exclude evaluation in use – in case such a specific use for the annotations was identified – but, in most cases, the results of reference-based evaluation are good indicators of performance in use, while being considerably cheaper to obtain, and more reliable in the sense that the measures can be repeated at will, with the same results on the same data.

6. Evaluation of type G/AG systems: reference-based vs. task-based evaluation

For HLT systems that generate linguistic output (type G or AG), reference-based evaluation can only be applied if one can determine a distance between the system's response and a set of ground truth responses that is potentially very large, has fuzzy borders, and is generally known only through a small set of samples that are collected from human subjects. The quality of a system's output, i.e. the distance to the set of acceptable responses, must either be judged directly by human evaluators, using or not the samples of acceptable responses, or inferred automatically from the distance to the samples. Therefore, while for type A systems the human judges define explicitly the set of acceptable responses, for type G/AG systems they merely verify mentally, using their linguistic competencies, whether a response belongs or not to this set, which is vastly larger in this case.

The design of reference-based automatic metrics for type G/AG systems has been formulated as a training problem (Soricut & Brill, 2004), often solved using machine learning. The distance to the samples and its average, when several samples are available, are often adjusted over training data to match human judgments of quality.

A typical example are machine translation (MT) systems, for which the BLEU metric (Papineni, Roukos, Ward & Zhu, 2001) estimates the quality of automatically translated sentences based on their similarity to up to four human-translated versions of the same source sentence (BLEU was manually optimized to match human judgments of adequacy and fluency). The limits of reference-based evaluation metrics for MT have been widely discussed (Culy & Riehemann, 2003; Callison-Burch, Osborne & Koehn, 2006), but the cost-effectiveness of these methods compensates their divergence from human judges in many cases, though as MT quality gets closer to human translators, the defects of reference-based metrics become more obvious (Popescu-Belis, 2003).

Task-based evaluation is the other option for assessing the quality of G/AG systems. This method appears to be more informative than reference-based evaluation as it measures "directly" the satisfaction of user needs (which is the very definition of quality in ISO terms) and considers all the quality aspects of a system, but comes at a significantly higher cost, as each measurement involves a large number of human subjects. Also, as each measurement has to be repeated when the system changes, task-based evaluation is less generic than reference-based evaluation.

Turning again towards MT evaluation as a case study, task-based evaluation was discussed by (White, Doyon & Talbott, 2000) among others, and has inspired a recently-proposed metric named HTER, which estimates the utility of MT output based on the human post-editing effort required to correct it (Snover, Dorr, Schwartz, Micciulla & Makhoul, 2006; Przybocki, Sanders & Le, 2006).

7. A risk of reference-based evaluation for type G/AG systems

As we have shown, reference-based metrics approximate the quality of the output from its "distance" to a small number of samples of desired output. When evaluators define such approximations in order to measure as accurately as possible output quality, providing data and software to compute the distances, these (pseudo)metrics are soon used by developers to improve their systems. Therefore, the metrics start being incorporated into the optimization criteria of the systems, especially those based on machine learning approaches. Hence, two potential problems may arise:

- if the metric is quite imperfect, training a system to improve scores will not improve its true quality (which can be assessed by independent metrics);
- although developers are not allowed to train their systems on *test* data (a fact that would invalidate the evaluation results), it is not impossible that they can train the system to obtain higher scores for a given evaluation metric, regardless of the training/test data.

A simple fix to both problems, still within the framework of reference-based evaluation, is to use several evaluation metrics instead of one, and consider that only concordant variations of all metrics represent significant variations of output quality. Another, more radical approach would be to use a previously unseen metric for official evaluation, although it is not likely that developers would accept such a challenge.

For instance, using again MT as a case study, BLEU scores are broadly improved if MT output is "smoothed" using a language model, regardless of the resulting meaning. To avoid this kind of tuning to BLEU, a solution can be to use several automated metrics, some of which are not n-gram based, as in the CESTA French evaluation campaign (Hamon, Popescu-Belis, Choukri, Dabbadie, Hartley, Mustafa El Hadi, Rajman & Timimi, 2006). The NIST TIDES campaigns in the USA also used internally several metrics – automatic and (for validation) human ones, although only BLEU scores were reported finally (NIST, 2006).

8. Evaluation in use for interactive systems

Type I systems do not produce directly a result based on input data, but require a series of interactions with a human user, in which language may appear in the input or output, and most often in both, for the general class of human-computer dialogue systems. Such systems have been called 'symbiotic' ones (King & Underwood, 2006), and the one-input/one-output view does not suit them: hence, reference-based evaluation metrics are difficult to apply to such systems, due to the large variety of possible input/output combinations at each step of the interaction. Therefore, type I systems are mainly evaluated using task-based approaches or evaluation in use, requiring human subjects to interact with the system (Dybkjær, Bernsen & Minker, 2004; Bevan, 2001). The toplevel parameters that are evaluated are:

- effectiveness, i.e. whether the task is accomplished or not;
- efficiency, i.e. how efficiently or quickly the task is accomplished;
- user-satisfaction;
- safety (seldom measured for HLT systems[1]).

The limits of this type of evaluation are its relatively higher cost with respect to reference-based evaluation (due to the use of human subjects) and the difficulty to generalize the obtained results to slightly different tasks or contexts of use.

The evaluation of interactive systems may use two slightly different approaches, depending on what level of generality is sought, and which human subjects are available. One can distinguish *task-based* evaluation from *evaluation in use*, defining the first one as evaluation using an idealized setting and generic subjects (or even another software interacting with the first one), while the second one is the evaluation in the final, intended context of use, with a sample of the final users. Task-based evaluation can be applied to research prototypes, while evaluation in use is reserved for end-user products.

Meeting browsers are a prototypical example of interactive systems which allow search and browsing of (large) multimedia recordings of meetings (instances of human dialogues) in order to find information that is relevant to the human users. Initial experiments in the evaluation of meeting browsers have defined reusable resources and metrics for task-based evaluation, and have shown the difficulties in reducing the variance of responses from human subjects (Popescu-Belis, Baudrion, Flynn & Wellner, 2008).

9. Context-based evaluation: between reference-based and evaluation in use

Our analysis has used two exhaustive typologies, one for evaluation methods (internal, external, in use) and the other for HLT systems (A, G/AG, I). A question arising at this point is the following one: where does one of the recent trends in the evaluation of HLT systems, namely *contextual evaluation*, belong in our analysis? This trend is best exemplified by the FEMTI guidelines for MT evaluation (Hovy, King & Popescu-Belis, 2002; Estrella, Popescu-Belis & Underwood, 2005) which emphasize the influence of the intended context of use of a system on the evaluation metrics used to assess its quality, i.e. a contextual quality model.

We hypothesize that contextual evaluation such as the FEMTI guidelines might offer a promising compromise between reference-based and task-based approaches, when neither approach is optimal. On the one hand, the

[1] An apocryphal example of safety evaluation (or lack thereof) is the proposal for MT known as "helicopters in Vietnam", which suggested to evaluate MT of technical documents (here, for helicopter maintenance) by the number of failures of the equipments repaired using translated documents.

methods contained in FEMTI-style guidelines cover both reference-based and task-based evaluation metrics, but at least in the case of MT systems, with a predominance of reference-based ones, related to external qualities, as FEMTI's generic quality model is based on the ISO toplevel external qualities. Therefore, using quality models inspired from FEMTI is a cost-effective approach to evaluation, if reference data and metrics can be found for each quality attribute that is evaluated.

On the other hand, unlike using a single reference-based metric, FEMTI argues that the set of evaluation metrics and their respective weights must be adapted to the intended context of use of the system, which is a significant step towards considering the human users of a system. The goal of FEMTI is in fact to generate evaluation plans that grasp the qualities of a system as close as possible to task-based evaluation, without the high costs and reduced generality of this type of evaluation. As shown within the EAGLES and ISLE projects (EAGLES Evaluation Working Group, 1996; Hovy, King & Popescu-Belis, 2002), the definition of a quality model should be based on an analysis of the intended use of the HLT system. This observation has inspired the FEMTI framework for MT evaluation but also user-based proposals for the evaluation of information retrieval systems (Sparck Jones, 2001; Chaudiron, 2004).

10. Conclusion

This paper has discussed the relationship between various types of evaluation and various types of HLT systems. While 'annotation' systems can be evaluated using mostly reference-based metrics and 'interactive' systems must be evaluated using task-based approaches, 'generation' systems are more challenging, as neither reference-based, nor task-based methods offer a satisfactory compromise between the cost of an evaluation (and hence its reproducibility) and its informativeness (the capacity to find the "real" qualities of a system). In this case, contextual evaluation exemplified by the FEMTI guidelines offers a principled way to use a set of reference-based metrics that is adapted to the intended tasks and users of a system.

11. Acknowledgments

The author acknowledges the support of the Swiss National Science Foundation's grants n. 200021-103318 and 200020-113604 for MT evaluation projects, and the IM2 National Center of Competence in Research.

12. References

Azuma M. (2001). SQuaRE: The Next Generation of the ISO/IEC 9126 and 14598 International Standards Series on Software Product Quality. *Proceedings of Escom 2001 (12th European Software Control and Metrics Conference)*, London, UK, pp. 337-346.

Bevan N. (2001). International Standards for HCI and Usability. *International Journal of Human-Computer Studies*, vol. 55, pp. 533-552.

Callison-Burch C., Osborne M. and Koehn P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. *Proceedings of EACL 2006 (11th Conference of the European Chapter of the Association for Computational Linguistics)*, Trento, Italy, pp. 249-256.

Chaudiron S. (2004). La place de l'usager dans l'évaluation des systèmes de recherche d'informations. In S. Chaudiron (ed.), *Évaluation des systèmes de traitement de l'information*, Paris, Hermès, pp. 287-310.

Culy C. and Riehemann S. Z. (2003). The Limits of N-Gram Translation Evaluation Metrics. *Proceedings of Machine Translation Summit IX*, New Orleans, Louisiana, USA, pp. 71-78.

Dale R., Moisl H. and Somers H. (ed.) (2000). *Handbook of Natural Language Processing*. New York, NY, USA, Marcel Dekker.

Dybkjær L., Bernsen N. O. and Minker W. (2004). Evaluation and Usability of Multimodal Spoken Language Dialogue Systems. *Speech Communication*, vol. 43, n° 1-2, pp. 33-54.

EAGLES Evaluation Working Group (1996). *EAGLES Evaluation of Natural Language Processing Systems*. Final Report Center for Sprogteknologi, EAG-EWG-PR.2 (ISBN 87-90708-00-8).

Estrella P., Popescu-Belis A. and Underwood N. (2005). Finding the System that Suits you Best: Towards the Normalization of MT Evaluation. *Proceedings of 27th ASLIB International Conference on Translating and the Computer*, London, UK, pp. 23-34.

Hamon O., Popescu-Belis A., Choukri K., Dabbadie M., Hartley A., Mustafa El Hadi W., Rajman M. and Timimi I. (2006). CESTA: First Conclusions of the Technolangue MT Evaluation Campaign. *Proceedings of LREC 2006 (5th International Conference on Language Resources and Evaluation)*, Genova, Italy, pp. 179-184.

Hovy E. H., King M. and Popescu-Belis A. (2002). Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, vol. 17, n° 1, pp. 1-33.

ISO/IEC (1999). *ISO/IEC 14598-1:1999 (E) -- Information Technology -- Software Product Evaluation -- Part 1: General Overview*, Geneva, International Organization for Standardization / International Electrotechnical Commission.

ISO/IEC (2001). *ISO/IEC 9126-1:2001 (E) -- Software Engineering -- Product Quality -- Part 1:Quality Model*, Geneva, International Organization for Standardization / International Electrotechnical Commission.

ISO/IEC (2004). *ISO/IEC TR 9126-4:2004 (E) -- Software Engineering -- Product Quality -- Part 3:Quality in Use Metrics*, Geneva, International Organization for Standardization / International Electrotechnical Commission.

King M. and Underwood N. (2006). Evaluating Symbiotic Systems: the Challenge. *Proceedings of LREC 2006 (Fifth International Conference on Language Resources and Evaluation)*, Genoa, Italy, pp. 2482-2485.

Mitkov R. (ed.) (2003). *The Oxford handbook of computational linguistics*. Oxford, UK, Oxford University Press.

NIST (2006). *NIST 2006 MT Evaluation Official Results*. National Institute of Standards and Technology, http://www.nist.gov/speech/tests/mt/mt06eval_official_results.html.

Papineni K., Roukos S., Ward T. and Zhu W.-J. (2001). *BLEU: a Method for Automatic Evaluation of Machine Translation*. Research Report, Computer Science IBM Research Division, T.J.Watson Research Center, RC22176 (W0109-022).

Popescu-Belis A. (2003). An experiment in comparative evaluation: humans vs. computers. *Proceedings of Machine Translation Summit IX*, New Orleans, Louisiana, USA, pp. 307-314.

Popescu-Belis A. (2008). Le rôle des métriques d'évaluation dans le processus de recherche en TAL. *T.A.L. (Traitement Automatique de la Langue)*, vol. 47, n° 2, pp. 25.

Popescu-Belis A., Baudrion P., Flynn M. and Wellner P. (2008). Towards an Objective Test for Meeting Browsers: the BET4TQB Pilot Experiment. In A. Popescu-Belis, H. Bourlard et S. Renals (ed.), *Machine Learning for Multimodal Interaction IV*, Berlin/Heidelberg, Springer-Verlag, pp. 108-119.

Przybocki M., Sanders G. and Le A. (2006). Edit Distance: A Metric for Machine Translation Evaluation. *Proceedings of LREC 2006 (5th International Conference on Language Resources and Evaluation)*, Genova, Italy, pp. 2038-2043.

Snover M., Dorr B., Schwartz R., Micciulla L. and Makhoul J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of AMTA 2006 (7th Conference of the Association for Machine Translation in the Americas)*, Cambridge, MA, USA.

Soricut R. and Brill E. (2004). A Unified Framework For Automatic Evaluation Using 4-Gram Co-occurrence Statistics. *Proceedings of ACL 2004 (42nd Annual Meeting of the Association for Computational Linguistics)*, Barcelona, Spain, pp. 613-620.

Sparck Jones K. (2001). Automatic language and information processing: rethinking evaluation. *Natural Language Engineering*, vol. 7, n° 1, pp. 29-46.

White J. S., Doyon J. B. and Talbott S. W. (2000). Determining the Tolerance of Text-Handling Tasks for MT Output. *Proceedings of Second International Conference on Language Resources and Evaluation (LREC'2000)*, Athens, Greece, vol. 1, pp. 29-32.