Evaluation of NLG: Some Analogies and Differences with Machine Translation and Reference Resolution

Andrei Popescu-Belis
ISSCO / TIM / ETI
University of Geneva
Bd. du Pont-d'Arve 40
1211 Geneva 4, Switzerland
andrei.popescu-belis@issco.unige.ch

Abstract

This short paper first outlines an explanatory model that contrasts the evaluation of systems for which human language appears in their input with systems for which language appears in their output, or in both input and output. The paper then compares metrics for NLG evaluation with those applied to MT systems, and then with the case of reference resolution, which is the reverse task of generating referring expressions.

1 Challenges in NLG Evaluation

Defining shared-task evaluation campaigns (STECs) is often the key to making progress in a particular domain, thanks to the convergence of several research teams. However, the definition of STECs requires an acceptable agreement, among a community of researchers, on the relevance of the selected problem to the domain, as well as on common evaluation metrics that indicate progress on this task.

In the domain of Natural Language Generation (NLG), recent proposals have started meeting the challenge of STEC definition (Belz and Kilgarriff, 2006), few years after a new metric for Machine Translation (MT) evaluation (Papineni et al., 2001) had revived the interest for common evaluations, thanks to its low application costs, which in turn led to significant improvement of MT systems, and especially statistical ones. So, an important question is: how could NLG benefit from a similarly innovative metric, and how could such a metric be found?

This short paper offers an explanation of the difficulty to evaluate NLG systems based on a typology of natural language processing (NLP) systems, and draws from this typology some suggestions for NLG evaluation (Section 2). Then, NLG evaluation is compared to MT evaluation (Section 3). Finally, the focus is set on referring expressions (REs), which have been used in the task proposed at the 2007 UCNLG+MT workshop, and which might help providing an indirect measure of NLG "quality" by combining the generation of REs with reference resolution (Section 4).

2 A Typology of NLP Systems and Its Relation to Evaluation

Some approaches to evaluation distinguish *intrinsic* from *extrinsic* methods (Sparck Jones and Galliers, 1996), i.e. methods that try to assess the "quality" of an output vs. methods that estimate its "utility" for a given task. Other approaches distinguish *internal* from *external* evaluation, and then evaluation *in use* (ISO/IEC, 2001): internal methods look at static properties of a system while external ones assessing its behaviour when it runs.

These types of evaluation are not equally well adapted to the various types of NLP systems. A useful typology of NLP tasks can be based on the role of language among the input and/or output to a system (Popescu-Belis, 2007). One can distinguish systems that have language as input (type A for 'analysis'), systems that have language as output (type G for 'generation'), systems that combine the two (type AG), and systems that must interact with a human user to produce a result (type AGI, with I for 'interactive').

Type A systems typically produce some form of

annotation of linguistic input data. Even if the correct annotation of some reference data set cannot always be determined with full certainty, evaluation of type A systems generally involves a distance-based comparison between the desired output and the actual output of the system.

Distance-based evaluation is much less applicable to type G and AG systems, for two reasons. The main one is that the range of acceptable outputs cannot generally be circumscribed with enough precision, given the very large variability of language-based output. (The proposed solution for MT is to use a very small subset of all acceptable output samples.) The second reason is that type G systems are not a homogenous group—no more than type A systems—which makes it difficult to define a single STEC for the whole G group. (Similarly, for the AG group, researchers focus in fact on tasks such as MT and summarization, with separate evaluation techniques.)

These considerations suggest the road to follow for the evaluation of type G systems: first narrow the targeted application (e.g. focus on generation of weather reports from standardized numeric data, or on generation of referring expressions) in order to be able to prepare reference data, which should include one or more samples of the desired output for each input. From here, it is possible to use:

1. distance-based metrics, by extrapolating the "quality" of a system's output from its distance to the samples of the desired output;

2. task-based metrics, by measuring either the performance of a human using the system's output to accomplish a given task, or the performance of another NLP system using the NLG output, provided a simple quality metric exists for this second system.

3 Evaluating NLG like MT and Summarization

Evaluation metrics that compute a distance between a candidate output, such as a generated sentence, and the samples of desired outputs have been applied with some success to MT evaluation (e.g. BLEU (Papineni et al., 2001)), and also to summarization evaluation (e.g. ROUGE (Lin, 2004)), although their accuracy has been challenged (Callison-Burch et al., 2006)[1]. The distance between generated sentences or expressions can be computed using n-gram similarity, word error rate, or other techniques.

Depending however on the type of input data selected for a STEC in NLG, it is quite likely that distance-based evaluation metrics are not fine-grained enough to capture significant differences between the outputs of two NLG systems, especially at the sentence or sub-sentence level—in particular because distance-based metrics need a large amount of data to stabilize their scores.

The GRE task proposed for the 2007 UCNLG+MT Workshop (Belz and Reiter, 2006; Gatt, 2007) focussed on the generation of referring expressions, or rather on the optimal selection of descriptive attributes from the logic-based description of a set of referents. Each candidate solution was compared to a set of solutions elicited from human judges—such a comparison follows the distance-based metrics mentioned above. This potentially successful STEC is nevertheless limited by the specificity of the input data, and by the cost of eliciting reference responses from human judges.

4 Task-based Evaluation: Combining NLG with Reference Resolution

The design of an NLG STEC based on referring expressions (REs) need not however be limited to distance-based evaluation metrics. An idea is to observe that generating REs is the converse task of "solving REs", which can mean two things. *Coreference resolution* deals with the grouping of the REs from a text which refer to the same entities (Hirschman, 1997), while *reference resolution* aims at constructing links between each RE and the (computer representation of the) entity that it refers to (Popescu-Belis and Lalanne, 2004).

Reliable evaluation metrics exist for both tasks (Vilain et al., 1995; Popescu-Belis and Robba, 1998; Popescu-Belis et al., 2004), and they are expressed as a distance to the correct distribution of REs that can be easily annotated by human judges, with high reliability (Passonneau, 2004).

The proposal is thus to couple an NLG module to a resolution system, and use the scores obtained by

[1] 'Accuracy' often means that the computed distance reflects well the "absolute" quality assessments done by human judges.

the resolution system to measure NLG performance. Which of the two tasks, co-reference or reference resolution, would be more appropriate? It is likely that co-reference would be less appropriate, as this would encourage the NLG system (or rather its authors) to generate "proper names" for each referent, and to repeat them identically throughout the generated text, which is neither natural nor usable by humans.

To reflect genuine NLG quality, the NLG system should rather be coupled to a reference resolution system, which will attempt to retrieve, from a logic-based description of the referents (available to both systems) the correct entity referred to by each generated RE. An efficiency constraint (or length penalty) should be added to avoid the NLG system producing too long and specific REs. Of course, the performance of the reference resolution system is not 100% even on human-generated REs, so the scores must be considered from a relative point of view only. That is, if the same reference resolution system scores better on the output of NLG system #1 than on the output of system #2, then the first system is "better" than the second one. This metric can be applied automatically as often as needed, for instance to measure progress or to compare systems. The reference resolution scores on REs generated by humans (i.e. the "perfect" REs) could then serve as an upper bound and comparison point for the automated NLG systems.

References

Anja Belz and Adam Kilgarriff. 2006. Shared-task evaluations in HLT: Lessons for NLG. In *INLG'06 (4th International Conference on Natural Language Generation)*, pages 133–135, Sydney, Australia.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation in NLG. In *EACL'06 (11th Conference of the European Chapter of the ACL)*, pages 313–320, Trento, Italy.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *EACL 2006 (11th Conference of the European Chapter of the ACL)*, pages 249–256, Trento, Italy.

Albert Gatt. 2007. Input and output specification for the shared task in referring expressions generation. Technical report, University of Aberdeen, UK, http://www.csd.abdn.ac.uk/research/evaluation/training/spec.pdf.

Lynette Hirschman. 1997. MUC-7 coreference task definition 3.0. Technical report, MITRE Corp., 13 July 1997.

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

Chin-Yew Lin. 2004. ROUGE: a package for automatic evaluation of summaries. In *ACL'04 Workshop on "Text Summarization Branches Out"*, pages 74–81, Barcelona, Spain.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. Research Report RC22176 (W0109-022), IBM T.J.Watson Research Center, 17 Sept. 2001.

Rebecca J. Passonneau. 2004. Computing reliability for coreference annotation. In *LREC 2004 (4th International Conference on Language Resources and Evaluation)*, volume IV, pages 1503–1506, Lisbon, Portugal.

Andrei Popescu-Belis and Denis Lalanne. 2004. Reference resolution over a restricted domain: References to documents. In *ACL'04 Workshop on Reference Resolution and its Applications*, pages 71–78, Barcelona, Spain.

Andrei Popescu-Belis and Isabelle Robba. 1998. Three new methods for evaluating reference resolution. In *LREC'98 Workshop on Linguistic Coreference*, Granada, Spain.

Andrei Popescu-Belis, Los Rigouste, Susanne Salmon-Alt, and Laurent Romary. 2004. Online evaluation of coreference resolution. In *LREC 2004 (4th International Conference on Language Resources and Evaluation)*, volume IV, pages 1507–1510, Lisbon, Portugal.

Andrei Popescu-Belis. 2007. Le rôle des métriques d'évaluation dans le processus de recherche en tal. *T.A.L. (Traitement Automatique de la Langue)*, 47(2):25.

Karen Sparck Jones and Julia Rose Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review*. LNAI 1083. Springer-Verlag, Berlin / New York.

Mark Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *6th Message Understanding Conference (MUC-6)*, pages 45–52, Columbia, MD.