



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

Andrei Popescu-Belis
ISSCO / TIM / ETI
University of Geneva

Workshop on Using Corpora in NLG and MT
MT Summit XI, Copenhagen, 11 September 2007

Towards Shared-Task Evaluation Campaigns (STECs) for NLG

- STECs are often key to making progress in a particular research domain
 - challenge: find an agreement among a community of researchers – on the selected problem and on common evaluation metrics
- **MT**: n-gram based evaluation metrics such as BLEU
→ revived the interest for common evaluations
 - due to low costs and “objectivity”
- How could **NLG** benefit from a similarly innovative metric, and how could such a metric be found?



Outline

- Why are NLG systems difficult to evaluate?
 - typology of NLP systems
 - two options for evaluation
- NLG evaluation compared to MT evaluation
- Focus on referring expressions
 - proposal to evaluate generation of REs in combination with reference resolution

Note on evaluation terminology

- Sparck Jones & Galliers
 - intrinsic evaluation
 - = assess the “quality” of output
 - extrinsic evaluation
 - = estimate “utility” of output for a given task
- ISO 9126
 - internal evaluation
 - = static properties of a system
 - external evaluation
 - = assess system behavior when it runs
 - evaluation in use
 - = assess performance of user + system

Typology of NLP Systems

- Based on place of language among input and/or output
 - Language as input = type A for 'analysis' or annotation
 - Language as output = type G for 'generation'
 - Combining the two = type AG
 - Interact with a human user to produce a result = type AGI for 'interactive'
- Evaluation of type A = distance-based comparison between the desired output and actual output
- Evaluation of type G and AG systems = the range of acceptable outputs cannot generally be circumscribed with enough precision
 - distance-based evaluation is less applicable
 - case of MT: a very small subset of all acceptable output samples is used as a reference

Case of “type G” systems

- Type G systems do not seem to be a homogenous group
 - (*no more than type A*)
 - difficult to define a single STEC for the whole G group
- Proposal
 - narrow the targeted application
 - e.g. generation of weather reports from standardized data, or (attribute selection for) generation of REs
 - reference data, including samples of the desired output
- Two options
 - use **distance-based metrics**: determine quality of an output from its distance to the samples of desired output
 - use **task-based metrics**: measure either
 - performance of human using the output for a given task, or
 - performance of another NLP system using the output
 - a simple quality metric is required for this second system

Distance-based metrics for NLG

- Compute a distance between candidate output (generated sentence), and (samples of) desired outputs
 - MT eval (e.g. BLEU), summarization eval (e.g. Rouge)
 - their accuracy is often challenged: how well do such metrics reflect “quality”?
- Distance-based evaluation metrics for generated text?
 - not fine-grained enough to capture significant differences
 - especially at sentence or sub-sentence level
 - need to average values over large amounts of data
- ASGRE task @ UCNLG+MT 2007
 - selection of descriptive attributes for referents within a set
 - three metrics: evaluate a candidate solution intrinsically (**uniqueness**, **minimality**) or with respect to a set of solutions elicited from human judges (**humanlikeness**)
 - ➔ this is distance-based evaluation
 - limited by the specificity and cost of the input data

Task-based metrics for NLG (1/2)

- Generating REs is the converse task of solving REs
 - co-reference resolution: group REs referring to same entities
 - reference resolution: construct links between each RE and the entity that it refers to
 - **evaluation metrics** exist for both tasks: distance between distributions of REs
- ASGRE task @ UCNLG+MT 2007 (**continued**)
 - two other metrics: evaluate a candidate solution with respect to an identification task assigned to subjects (**accuracy, speed**)
 - ➔ this is task-based evaluation with human subjects
 - limited by the performance and cost of the subjects

Task-based metrics for NLG (2/2)

- Automating task-based NLG evaluation
- Couple an NLG module to a resolution system
- Use obtained scores to measure NLG performance
 - not a co-reference resolution system
 - would encourage generation of “proper names” for each referent, and repeating them identically
 - but a reference resolution one
 - as in the ASGRE identification task for humans
 - retrieve from logic-based description of referents the correct entity referred to by each generated RE
 - + efficiency constraint (length penalty), to avoid too long/specific REs
- Performance of reference resolution system is not 100%
 - but if relative scores improve, it means NLG improves as well

Conclusion

- Two priorities for defining a STEC
 - specify which aspect of NLG is targeted
 - attempt to automate task-based evaluation in order to avoid preparing too much data
 - n-gram based distances do not seem very well adapted
 - perspectives to automate task-based evaluation by combining NLG with another module