Automatic Evaluation of Referring Expression Generation Is Possible

Jette Viethen
Centre for Language Technology
Macquarie University
Sydney NSW 2109
jviethen@ics.mq.edu.au

Shared evaluation metrics and tasks are now well established in many fields of Natural Language Processing. However, the Natural Language Generation (NLG) community is still lacking common methods for assessing and comparing the quality of systems. A number of issues that complicate automatic evaluation of NLG systems have been discussed in the literature. [1]

The most fundamental observation in this respect is, in my view, that speaking about "evaluating NLG" as a whole makes little sense. NLG is not one task such as Syntax Parsing or Information Retrieval, but comprises many different subtasks. Just as the subtasks of NLU are evaluated separately using different metrics, corpora and competitions, the subtasks of NLG can only be evaluated individually. With its relatively clear defined task and input characteristics, referring expression generation (REG) is a subtask of NLG for which a shared evaluation scheme is conceivable. In this position paper, I therefore aim to take a solution-oriented look at the challenges of evaluating REG. Although it is unclear just how far any solutions for REG evaluation can be transferred directly to other NLG subtasks, progress in one task might help find solutions for others.

Gold standards: Natural language provides almost countless possibilities to say the same thing in a different way and even under the same external circumstances people use different descriptions for the same object. This variability of human language poses a huge difficulty in terms of what could be used as a gold standard corpus for the evaluation of any NLG task, including REG. It would be unfair to penalise a REG system for not delivering the exact referring expression contained in a corpus, when a large number of alternatives might be equally good or acceptable.

My position: A corpus for REG evaluation would have to contain a large number of descriptions for each referent, as opposed to just one solution per instance. It is unlikely that such a corpus can be drawn from naturally occurring text; the corpus would need to be constructed 'artificially'. This might be done by asking many online participants to provide descriptions for objects from scenes displayed on the screen.

Nevertheless, we will need to keep in mind that an evaluation corpus in NLG will never be really golden: a bad evaluation score might only be due to the 'bad luck' that the perfectly viable solutions a system delivers do not occur in the corpus.

What output do we expect? Three questions need to be answered with respect to the expectations we have of the output of a REG system. Firstly, we lack a definite *Goodness Measure* with which to assess the quality of a referring expression. Secondly, the *Linguistic Level* of the output of existing systems varies and it is not clear at which level we should evaluate. Most researchers are mainly interested in content determination, while some are concerned with the property ordering or even full lexical and syntactic surface realisation. A third question concerns *Solution Counts*: are we contented with one *good* referring expression for each referent, or do we expect a system to be able to produce all the possible descriptions for a referent used by humans.

My position: Psycholinguistic theories such as Grice's maxims of conversational implicature might provide an accurate model of speakers' behaviour when they refer. However, they do not

[1] A bibliography on recent literature relevant to the evaluation of referring expression generation and NLG can be found at http://www.ics.mq.edu.au/~jviethen/evaluation.

provide a straightforward way to reverse-engineer from these behavioural rules to practical guidelines for judging the actual referring expressions produced. A simple and feasible way to find a *Goodness Measure* for the output of REG systems would be to ask human participants not only to provide a description for the gold standard corpus, but also to rank different referring expressions for the same object.

It is clear that output at different *Linguistic Levels* cannot be evaluated using the same corpus and metrics. Before we enter a long and possibly fruitless discussion, we could get started by limiting ourselves to evaluation of REG systems only concerned with content determination. However, we should ensure the possibility to extend the corpus and metric to take word order and surface realisation into account with little extra effort.

If a *Solution Count* of one per referent is expected, the evaluation score can depend directly on the goodness rank of that solution in the corpus (if present at all). If more than one description is allowed, the number of descriptions provided and penalties for over-generation need to be incorporated in the evaluation metric to avoid 'blind' attempts at listing hundreds of descriptions.

Parameters: Most REG systems take a number of parameters such as preference orderings or cost functions over properties and objects, which can have a huge impact on the output. In view of the variability of human-produced referring expressions, it could be argued that algorithms should be allowed to use multiple parameter settings for an evaluation to produce different referring expressions. However, in some cases the parameters are so fine-grained that virtually any desired output can be engineered by carefully choosing the right settings.

My position: This means either that the parameter setting should be considered part of the algorithm proper allowing only one setting to be used throughout the whole evaluation, or that the evaluation metric must penalise systems for switching parameter settings during the evaluation.

A wide field with few players: Just as NLG is a huge field with many subfields, REG can be subdivided into different subtasks such as descriptions involving relations, incorporating object and property salience, or describing sets, and higher-level surface realisation tasks. This is compounded by the high domain-specificity of NLG systems in general. At the same time, the number of researchers in REG, as in most NLG subfields, is comparatively low.

My position: A competitive evaluation scheme for REG bears the potential to stifle research in this field by drawing the attention of the few people working in it to a race for slight percentage increases in a small subtask and domain, instead of advertising the advantages of working on the many untouched research questions.

To cater for evaluation of different subtasks of REG, the gold standard corpus needs to be subdividable and contain referring expressions of different kinds and different domains. To get started, it could be restricted to the most commonly considered types of referring expressions and subsequently extended.

Input Representation: Arguably, the problem of agreeing on the input for NLG is the biggest obstacle in the way towards automatic evaluation. Not only are input representations highly dependent on the application domain of a system, but in existing REG systems the design of the knowledge base from which the algorithm can draw the content for a description is usually tightly intertwined with the design of the algorithm itself. The amount and detail of information contained in the system input differs from case to case, as well as the form it takes: this can range from raw numerical data, over premeditated ontologies of domains, to natural newspaper text.

My position: In order to automatically evaluate REG systems, we have no other choice but to agree on the type of knowledge representation required for the domains covered in the evaluation corpus. As a minimum, the properties and relations of the objects in the different scenes that a system can draw from will need to be predetermined in a simple standard knowledge representation.

Conclusion: There are a number of challenges that have to be overcome in developing useful evaluation metrics for any NLG subtask. However, I am convinced that, for REG, automatic evaluation is possible and would be highly beneficial to the development of systems, if it is based on a large, divisible corpus of ranked descriptions and on basic agreements regarding input representation, parameters, and output expectations.