Content Determination in GRE: Evaluating the evaluator

Kees van Deemter and **Albert Gatt**
Department of Computing Science
University of Aberdeen
{kvdeemte, agatt}@csd.abdn.ac.uk

1 Introduction

In this paper, we discuss the evaluation measures proposed in a number of recent papers associated with the TUNA project[1], and which have become an important component of the First NLG Shared Task and Evaluation Campaign (STEC) on attribute selection for referring expressions generation. Focusing on reference to individual objects, we discuss what such evaluation measures should be expected to achieve, and what alternative measures merit consideration.

The measures mentioned above can be motivated as follows. Suppose a large number of utterance situations had been defined, where each situation contained a number of objects, one of which needed to be described by a referring expression. Suppose, furthermore, one had an infallible oracle which told us, for each of these situations, what was *the best* referring expression for that situation. **How could this oracle be used to evaluate the extent to which *other* referring expressions are similar to the one proposed by the oracle?** In reality, an infallible oracle is not available of course. What *is* available is a large corpus in which sixty-odd human subjects make their best stab at each of the utterance situations. We acknowledge that evaluation measures could handle the unavoidable differences between human subjects in different ways: focussing on the *average* Dice score of an algorithm (over all subjects as well as all the descriptions in the corpus) is evidently only one possibility. This, however, is a topic for another day: we will assume there to be one oracle, and we will take this oracle to be infallible. A number of other simplifications were made. In particular, all the referring expressions involved are thought of as sets of properties (rather than an ordered sequence of words). In other words, this is an evaluation of the *semantic content* of a referring expression. Gatt et al. (2007) proposed to use the Dice measure (Salton and McGill., 1983), defined as follows: Let $s(A, B)$ denote the degree of similarity between sets of properties A and B. Then $s(A, B)$ equals $2n/(\|A\| + \|B\|)$, where n is the cardinality of $A \cap B$ (and $\|X\|$ denotes the cardinality of X). The Dice measure is symmetrical, in that $s(A, B) = s(B, A)$, for all A and B. Also, $s(A, A)$ holds for every set A. Finally, a triangular kind of transitivity holds: if $s(A, B) = m$ and $s(B, C) = n$ then $s(A, C) \leq m + n$. All of this is, of course, as one would expect of a well-behaved similarity relation. In the following, however, we want to argue Dice is not the right measure for GRE in the long run. We shall do this by raising a number of questions about the notion of similarity that is relevant for GRE. We start with a relatively minor issue.

1.1 Do addition and deletion cost the same?

Dice punishes the omission of properties from the oracle more heavily than the addition of properties to it. Consider the case where the Oracle O produces the description $\{P, Q\}$. Suppose an algorithm A_1 produces the description

[1] Towards a Unified Algorithm for the Generation of Referring Expressions. See: http://www.csd.abdn.ac.uk/research/tuna/

$\{P\}$ (leaving one property out); Algorithm A_2 produces $\{P, Q, R\}$ (adding one property to the description proposed by the oracle). According to Dice, $s(A_1, O) = 2/3$, while $s(A_2, O) = 4/5$. The difference becomes smaller as the size of descriptions grows (but short descriptions are highly frequent, so they have a large influence of the average Dice score achieved by an algorithm). It might be thought that this is not so bad. Adding properties is, arguably, a smaller sin than leaving them out, because redundancy can be useful (Paraboni et al., 2007). In the present context, however, this seems irrelevant since Dice does its thing irrespective of whether the descriptions in question are fully, under- or overspecified (see §1.2 below). There is something to be said for replacing Dice by a version of edit distance (after making sure that all sets contain their elements in the same order), making addition and deletion equally costly. It might be best to do this in such a way that *substitutions* (which Dice punishes even more heavily than omissions, which seems difficult to motivate) are not viewed as combined deletion + addition, but perhaps as equally costly as each of the other operations.

1.2 Does discriminatory power matter?

Dice is completely blind towards the goal of a description. Let us keep matters simple by assuming, as is customary at the present stage of research in GRE, that *identification of the referent* is the only goal of a referring expression. (Our remarks can easily be generalised to the case where other communicative goals are taken into account, cf. Jordan and Walker (2005) .) Even this goal is disregarded by Dice. This is most easily seen when comparing two descriptions, one of which underspecifies its referent while the one one does not. For example, suppose the oracle O says $\{P, Q, R\}$, while the minimal description (i.e., the smallest set of properties identifying the referent) is $\{P, Q\}$. Now compare two algorithms: A_1 which produces precisely this minimal description, and A_2 which produces the description $\{P, R\}$, which (we assume) fails to identify the referent. Dice treats the two descriptions as equally similar to O's proposal. An obvious move would be to use underspecification as a second measure, additionally to Dice.[2] Alternatively, one could modify the Dice measure, punishing any algorithm for every time it deviates from the extent to which O has specified the referent (e.g., by underspecifying if O does not, or by fully specifying where O does not). But surely, this is tinkering with a flawed method! After all, the term underspecification (and its mirror image overspecification likewise) covers a multitude of sins, and one would want to take account of the degree to which a given description under- or overspecifies, for example as measured by the number of distractors that a description fails to remove. In other words, just counting properties (as done by the Dice measure) is only one way in which a description should be judged; another dimension is that of the degree of over- and underspecification.

1.3 Are all properties equidistant?

Taking the degree of over- and underspecification inherent in a description into account (as discussed under 2) might be deemed to be a bridge too far. Even so, it seems unnecessarily crude to assume that two atomic properties can only relate to each other by being equal or different. The properties ANIMAL and MAMMAL, for example, are different, yet they are closely related in many ways. For example, one subsumes the other. Surely this makes $\{striped, animal\}$ more similar to $\{striped, mammal\}$ than it is to $\{striped, mother\}$. It seems natural here to take a leaf out of the Information Retrieval book by viewing a description as a *vector*.[3] One way of doing this is as follows:

Suppose we represent descriptions not simply as sets of un-analysed properties but as sets of ⟨Attribute, Value⟩ pairs. Then each Attribute can be seen as a dimension, the points on which are sets (not numbers). We can then compare

[2] This is indeed the option taken by the organisers of the Shared Task, where in addition to Dice, automatic evaluation includes an estimate of whether a description is minimal and/or uniquely distinguishing.

[3] For a description of this and several other similarity measures, see `http://www.dcs.shef.ac.uk/ sam/stringmetrics.html`.

two descriptions by inspecting the Values they assign to a given Attribute. (For simplicity, we assume that each Attribute can have only one Value in a given description. If an Attribute has no Value in the description then it is regarded as semantically empty, i.e., coreferential with the domain as a whole.) For example, one description might be represented as $\{\langle$Type: Mammal$\rangle$, $\langle$Origin: Africa$\rangle$, $\langle$Gender: Female$\rangle\}$, another as $\{\langle$Type: Animal$\rangle$, $\langle$Origin: Africa$\rangle$, $\langle$Gender: Any $\rangle\}$.

We now need a way to decide how similar two Values of a given Attribute are. One simplistic (because exclusively extensional-semantic) approach is to use Dice once again, this time at the level of property denotations (i.e., at the level of the sets of objects for which a given Value holds true). Suppose the animals in the domain are $\{a_1, ..., a_{20}\}$, while the mammals are $\{a_1, ..., a_{15}\}$. Because both these denotations are sets, their similarity $s(animal, mammal)$ could be calculated as $(2.15)/35 = 6/7$ (twice the number of objects in the intersection of ANIMAL and MAMMAL, divided by the total number of objects). If the mothers in the domain are $\{a_1, a_2, a_3, , ..., a_{19}, a_{20}\}$ then the similarity $s(animal, mother)$ is much lower, at $(2.5)/25 = 2/5$. This would be one possible way in which Dice could be made to take similarity between properties into account. Alternatively, one could use the distance between properties in an ontology tree for this purpose.

2 Conclusion

We do not claim to have the answers to all the questions that we have raised. Moreover, we are aware that other, equally pertinent questions could be asked. (How, for example, might evaluation measures be used for evaluating fully realised Noun Phrases, instead of their semantic content only?) Hardest of all, a philosophical question comes up: How do we decide whether one evaluation measure is better than another? In other words, how does one evaluation an evaluation measure? Arguably, this can only be done by relating similarity measures to something else. In the case of Information Retrieval, this "something else" tends to be, ultimately, a measure of user satisfaction (e.g., as captured by precision and recall with respect to the set of documents that a user thinks relevant). In the case of GRE, one might simply ask human subjects "How similar are the descriptions X and Y in your opinion?", but the subject might retort "Similar in what respect?" In response, one could focus on the usefulness of a generated description to a reader/hearer (asking which similarity metric offers the best formalisation of the degree to which two descriptions are similar in their usefulness to a reader). One might ask subjects "How similar are the descriptions X and Y, in terms of their usefulness to a hearer?" Alternatively, one might measure the usefulness of descriptions for a particular task (in terms of the number of errors made by a human interpreter, for example), and derive a notion of similarity from the results. Either way, it appears that speaker-oriented evaluation (i.e., where the quality of a description is a function of its similarity to descriptions produced by human speakers or writers) cannot stand on its own, and must ultimately be connected with hearer-oriented evaluation (i.e., where the quality of the description is a function of its effect on a hearer or reader). – In short: similarity is in the eye (or the experience) of the beholder.

References

A. Gatt, I. van der Sluis, and K. van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. In *Proceedings of the 11th European Workshop on Natural Language Generation, ENLG-07*. To appear.

P. W. Jordan and M. Walker. 2005. Learning content selection rules for generating object descriptions in dialogue. *Journal of Artificial Intelligence Research*, 24:157–194.

I. Paraboni, K. van Deemter, and J.Masthoff. 2007. Generating referring expressions: making referents easy to identify. *Computational Linguistics*, 32(2).

G. Salton and M. J. McGill. 1983. *Introduction to modern information retrieval*. McGraw Hill, New York.