Beyond DICE: measuring the quality of a referring expression

Kees van Deemter (k.vdeemter@abdn.ac.uk)
Department of Computing Science, University of Aberdeen
Aberdeen, UK

Albert Gatt (a.gatt@abdn.ac.uk)
Department of Computing Science, University of Aberdeen
Institute of Linguistics, University of Malta

Abstract

This paper discusses ways in which the similarity between the contents of two referring expressions can be measured. Similarity metrics of this kind are essential when expressions generated by an algoritm are compared against the ones produced by human speakers, for example as part of an experiment in which referring expressions are elicitated. We discuss arguments for and against different metrics, taking our departure from the well-known Dice metric.

Keywords: Similarity metrics; Dice metric; generation of referring expressions; evaluation; content determination

Introduction

Computational work on the generation of referring expressions (GRE) is often guided by automatic metrics for measuring the "quality" of the expressions produced by an algorithm. The metrics tend to be ultimately based on a comparison with human performance. One method is to elicit referring expressions from experimental participants, annotating them with semantic information to construct a "semantically transparent" text corpus (van Deemter et al. 2006). The quality of a generated referring expression is then measured as a function of its **similarity** to the relevant expressions in the corpus. Procedures of this kind have been an important component in a recent series of Shared Task and Evaluation Challenges (STECs) which focussed on referring expressions (Gatt and Belz 2008, following proposals in Gatt et al. 2007), which collectively constitute the most extensive attempt to date at linking computational GRE with psycholinguistic experimentation.

It is possible to think of a referring expression as expressing a set of semantic properties (e.g., being a male person). This is a reductive view, which abstracts away from syntactic structure and word choice (e.g. 'man', 'guy', 'chap', etc.), but it is an interesting view nonetheless, which leads one to focus on the semantic content of the expression, in line with much computational work on GRE (e.g. Dale 1989, Dale and Reiter 1995, Krahmer et al 2002). The same semantically-oriented perspective is also implicit in a number of psycholinguistic studies on reference which focus on what properties of a referent tend to be selected by speakers, and why (e.g. Pechmann 1989, Belke and Meyer 2002, Engelhardt et al. 2006).

When this view is taken, it is tempting to use an evaluation metric designed for measuring the similarity between *sets*. One such metric is the Dice metric, which measures the similarity between sets A and B as $s(A,B) = 2n/(\|A\| + \|B\|)$, where n is the cardinality of $A \cap B$ (and $\|X\|$ denotes the cardinality of X) (Salton and McGill 1983). The Dice metric is symmetrical, in that $s(A,B) = s(B,A)$, for all A and B. Also, $s(A,A) = 1$ 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 as one would expect of a similarity relation. Dice was used heavily in the STECs and related publications.

This paper asks how suitable *Dice* and related metrics are for the specific task of measuring the semantic quality of a referring expression, and what an ideal metric that takes the peculiarities of GRE into account would look like (cf. Van Deemter and Gatt 2007 for preliminary notes). Our observations are also relevant for most other evaluation metrics that are currently used, and which tend to resemble Dice closely, in that they compute the similarity between A and B as a function of the cardinalities of (at most) all possible set-theoretic combinations of A and B, namely A and B themselves, $A \cup B$, $A \cap B$, $A - B$ and $B - A$.[1] We focus here on reference to individual objects: reference to sets introduces additional problems, because a non-singleton set can be overconstrained to different degrees (depending on how many elements of the target set are incorrectly ruled out by a referring expression) that we do not have space to discuss.

Questions about metrics

Many corpus-based evaluations assume that a corpus represents a gold standard. It is viewed as an oracle which says, for each of a large number of situations, what *the best* referring expression is in this situation. This view allows us to address the main question we are interested in, namely how such an oracle should be worked into an evaluation metric. 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, each using their own linguistic style. Differences between subjects open up many interesting issues (e.g., calculating the average over all subjects' descriptions is one possibility; see e.g. Reiter and Sripada 2002 for discussion) but in this paper, we simplify by assuming there to be one oracle, which we take to be infallible.

Our research question will be broken down into three smaller questions:

[1] A well-known variant on Dice is the Jaccard metric, defined as $s(A,B) = \|A \cap B\|$ divided by $\|A \cup B\|$.

1. Is it better to add a property or to leave one out (or both)? 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 $\{P\}$ (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). Adding properties is arguably a smaller sin than leaving them out, because redundancy can be useful (to the hearer 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 point 2). 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) are not viewed as combined deletion + addition, but as equally costly as each of the other operations.

2. Does identification matter? Dice is a general metric for comparing sets. As such it is blind towards the goal of a description. GRE has often made strong simplifications, for example by focussing on one-shot (i.e., non-anaphoric) descriptions. More crucially, for present purposes, it has often focussed on situations where *identification of the referent* is the only goal of a referring expression. (Our remarks can be generalised to the case where other communicative goals are taken into account, cf. Jordan and Walker 2005). But even identification of the referent is disregarded by Dice. This is most easily seen when comparing two descriptions, one of which underspecifies its referent while the other 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 metric, additionally to Dice. Alternatively, one could modify the Dice metric, 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). More drastically, one could take account of the degree to which a given description under- or overspecifies, as measured by the number of distractors that a description fails to remove. In the set-theoretic spirit of Dice, the Dice metric could be revised by multiplying the original Dice coefficient by 1 minus the proportion of "incorrectly" treated domain elements.

3. Are all properties equidistant? The sets we are interested in comparing consist of properties. It seems crude to assume, as Dice suggests, 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. Surely this makes $\{striped, animal\}$ more similar to $\{striped, mammal\}$ than it is to $\{striped, mother\}$. Let us see whow one might take a leaf out of the Information Retrieval book by viewing a description as a *vector* (e.g. Simetrics 2007).

Suppose we represent descriptions not simply as sets of unanalysed 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 two descriptions by inspecting the Values they assign to a given Attribute. (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 {⟨Type: Mammal⟩, ⟨Origin: Africa⟩, ⟨Gender: Female⟩}, another as {⟨Type: Animal⟩, ⟨Origin: Africa⟩, ⟨Gender: Any ⟩}. We now need a way to decide how similar two Values of a given Attribute are. One interesting 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 is one possible way in which a revised metric could take similarity between properties into account. Alternatively, one could use the distance between properties in an ontology or taxonomy.

Evaluating the evaluator

There appears to be a strong case for adapting metrics such as Dice in such a way that the nature of the sets whose similarity is being assessed is taken into account. But how does one "prove" that one evaluation metric is better than another? In other words, How does one evaluate an evaluation metric? We suggest that this should involve studying how well the metric corresponds to an external (non corpus-based) evaluation criterion. We propose to do this by making use of a little-explored part of one of the STECs (Gatt et al. 2009), where human participants were asked to rate the adequacy of a set of referring expression in a set of utterance situations. We are planning to use this data to compare the Dice metric to some of the variant metrics suggested above, by analysing their correspondence to adequacy judgements produced by humans, to see which metric offers the best prediction as to which referring expressions are equally adequate.

Acknowledgement

Thanks are due to three anonymous reviewers for their helpful comments. The work presented here was supported by the EPSRC platfrom grant Affecting People with Natural Language (EP/E011764/1).

References

E. Belke and A. Meyer. Tracking the time course of multidimensional stimulus discrimination: analysis of viewing patterns and processing times during same-different decisions. *European Journal of Cognitive Psychology* 14(2): 237-266.

R. Dale. Cooking up referring expressions. In *Proceedings of 27th Annual Meeting of the Association for Computational Linguistcs (ACL-89).*

R. Dale and E. Reiter. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science* 19(8): 233-263.

P. E. Engelhardt, K. Bailey, and F. Ferreira. Do speakers and listeners observe the Gricean maxim of Quantity? *Journal of Memory and Language* 54: 554-573.

A. Gatt and A. Belz (2008). Attribute selection for referring expression generation: New algorithms and evaluation methods. In *Proceedings of the 5th International Conference on Natural Language Generation (INLG-08).*

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).*

A. Gatt, A. Belz and E. Kow (2009). The TUNA-REG Challenge 2009: Overview and Evaluation Results. In *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG-09).*

P. Jordan and M. Walker (2005). Learning content selection rules for generating object descriptions in dialogue. *Journal of AI Research* 24: 157-194.

E. Krahmer, S. van Erk, and A. Verleg (2002). Graph-based generation of referring expressions. *Computational Linguistics* 29(1): 53-72.

I. Paraboni, K. van Deemter and J. Masthoff (2007). Generating referring expressions: making referents easy to identify. *Computational Linguistics* 32(2): 229-254.

T. Pechmann (1989). Incremental speech production and referential overspecification. *Linguistics* 27: 89-110.

E. Reiter and Y. Sripada (2002). Should corpora texts be gold standards for NLG? In *Proceedings of the Second International Conference on Natural Language Generation (INLG-02).*

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

Simetrics 2007. Web page on string metrics, `http://www.dcs.shef.ac.uk/∼sam/stringmetrics.html`

K. van Deemter and A. Gatt (2007). Content Determination in GRE: Evaluating the Evaluator. In *Proceedings of the MT Summit XI: Language Generation and Machine Translation (UCNLG+MT).*

K. van Deemter, I. van der Sluis, and A. Gatt (2006). Building a semantically transparent corpus for the generation of referring expressions. In *Proceedings of the 4th International Conference on Natural Language Generation (INLG-04).*