Building a semantically transparent corpus for the generation of referring expressions

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

Abstract

This paper discusses the construction of a corpus for the evaluation of algorithms that generate referring expressions. It is argued that such an evaluation task requires a semantically transparent corpus, and controlled experiments are the best way to create such a resource. We address a number of issues that have arisen in an ongoing evaluation study, among which is the problem of judging the output of GRE algorithms against a human gold standard.

1 Creating and using a corpus for GRE

A decade ago, Dale and Reiter (1995) published a seminal paper in which they compared a number of GRE algorithms. These algorithms included a Full Brevity (FB) algorithm which generates descriptions of minimal length, a greedy algorithm (GA), and an Incremental Algorithm (IA). The authors argued that the latter was the best model of human referential behaviour, and versions of the IA have since come to represent the state of the art in GRE. Dale and Reiter's hypothesis was motivated by psycholinguistic findings, notably that speakers tend to initiate references before they have completely scanned a domain. However, this finding affords different algorithmic interpretations. Similarly, the finding that basic-level terms in referring expressions allow hearers to form a psychological gestalt could be incorporated into practically any GRE algorithm.[1]

We decided to put Dale and Reiter's hypothesis to the test by an evaluation of the output of different GRE algorithms against human production. However, it is notoriously difficult to obtain suitable corpora for a task that is as semantically intensive as Content Determination (for GRE). Although existing corpora are valuable resources, NLG often requires information that is not available in text. Suppose, for example, that a corpus contained articles about politics, how would the output of a GRE algorithm be evaluated against the corpus? It would be difficult to infer from an article exactly which representatives in the British House of Commons are Liberal Democrats, or Scottish. Combining multiple texts is hazardous, since facts could alter across sources and time. Moreover, the conditions under which such texts were produced (e.g. *fault-critical* or not, as explained below) are hard to determine.

A recent GRE evaluation by Gupta and Stent (2005) focused on dialogue corpora, using MAPTASK and COCONUT, both of which have an associated domain. Their results show that referent identification in MAPTASK often requires no more than a TYPE attribute, so that none of the algorithms performed better than a baseline. In contrast to MAPTASK, COCONUT has a more elaborate domain, but it is characterised by a collaborative task, and references frequently go beyond the identification criterion that is typically invoked in GRE[2]. Mindful of the limitations of existing corpora, and of the extent to which evaluation depends on the corpus under study, we are using controlled experiments to create a corpus whose construction will ensure that existing algorithms can be adequately differentiated on an identification task.

[1] A separate argument for IA involves tractability, but although some alternatives (such as FB) are intractable, others (such as GA) are only polynomial, and can therefore not easily be dismissed on purely computational grounds.

[2] Jordan and Walker (2000) have demonstrated a significantly better match to the human data when task-related constraints are taken into account.

Proceedings of the Fourth International Natural Language Generation Conference, pages 121–123,
Sydney, July 2006.

2 Setup of the experiment

Like Dale and Reiter (1995), we focused on first-mention descriptions. However, we decided to include simple 'disjunctive' references to sets (as in 'the red chair and the black table'), in addition to conjunctions of atomic properties, since these can be handled by essentially the same algorithms (van Deemter, 2002). For generality, we looked at two very different domains. One of these involved artificially constructed pictures of furniture, where the available attributes and values are relatively easy to determine. The other involved real photographs of individuals, which provide a richer range of options to subjects. To date, data has been collected from 19 participants, and analysis is in progress.

Our first challenge was to make the experiment naturalistic. Subjects were shown 38 randomised trials, each depicting a set of objects, one or two of which were the targets, surrounded by 6 distractors (Figure 1). In each case, a minimal distinguishing description of the targets was available. Subjects were led to believe that they would be describing the targets for an interlocutor. Once a description was typed, the system removed from the screen what it took to be the referents.

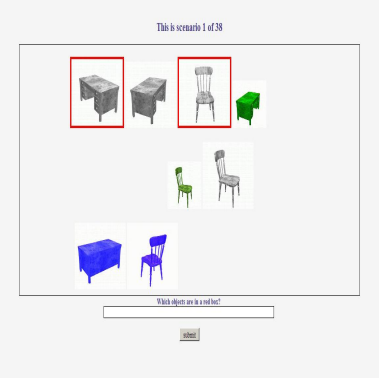

Figure 1: A stimulus example from the furniture domain.

Three groups performed the task in different conditions, namely: $\langle \pm FaultCritical \rangle$, where half the subjects in the $\langle +FaultCritical \rangle$ case could use location ('in the top left corner'). The $\langle +FaultCritical \rangle$ group was told: 'Our program will eventually be used in situations where it is crucial that it understands descriptions accurately. In these situations, there will often be no option to correct mistakes. Therefore, (...) you will not get the chance to revise (your description)'. By contrast, the $\langle -FaultCritical \rangle$ subjects were given the opportunity to revise their description should the system have got it wrong. Subjects in the $\langle -Location \rangle$ condition were told that their interlocutor could see exactly the same pictures as they could, but these had been jumbled up; by contrast, $\langle +Location \rangle$ subjects were led to believe that their addressee could see the pictures in exactly the same position.

The second main challenge was to create trials that would distinguish between all the algorithms. For instance, if trials involved only one attribute, say an object's TYPE (e.g., *chair* or *table*), they would not allow us to distinguish IA from FB, as both would always generate the shortest description. Subtler issues arise with *local brevity* (Reiter, 1990), an optimisation strategy which requires sufficiently complex trials to make a difference.

3 How to analyse the data?

Our semantically transparent corpus can be used for testing various hypotheses, for instance about when an algorithm should overspecify descriptions (e.g. more in $\langle +FaultCritical, +Location \rangle$ (Arts, 2004), and/or when the target is a set). Here, we focus on the issue raised in Section 1, namely, which of the algorithms discussed in Dale and Reiter (1995) matches human behaviour best.

The first problem is determining the relevant algorithms. The IA comes in different flavours, because its output depends on the order in which the different properties are attempted (commonly called the preference order). It is possible to consider *all* different IAs (trying every conceivable preference order), but this would increase the number of statistical hypotheses to be tested, impacting the validity of the results and requiring a Bonferroni correction. Instead, we are using a pretest to find the optimal version of IA, comparing only that version to the other algorithms.

The second question is how to assess algorithm performance. Since our production experiment does not yield a *single* gold standard (GS), an algorithm might match subjects better in one condition (e.g. $\langle +FaultCritical \rangle$), or perform better in one domain (e.g. furniture). Moreover, it might match subjects poorly overall due to sample variation, while evincing a perfect match with a single individual. Using both a *by-subjects* and a *by-items* analysis will partially control for sample

dispersion.

How should we calculate the *match* between an algorithm and a GS? Once again, there are two facets to this problem. Since we are focusing on Content Determination, each human description could be viewed as associating, with the relevant trial, a set of properties. Our approach will be to annotate each human description with the set of attributes it contains. However, the real data is often messy. For example, when one subject called an object 'the non-coloured table', and another called it 'the grey desk', both may be expressing the same attributes (i.e. TYPE and COLOUR). Also, while it is often assumed that the output of GRE is a definite noun phrase, this is not always the case in our corpus, which contains indefinite distinguishing descriptions such as *'a red chair, facing to the right'*, and telegraphic messages such as *'red, right-facing'*.

The second aspect to the problem concerns the actual human-algorithm comparison. Suppose the GS equals the output of one subject, and we are comparing two algorithms, x and y. Suppose our subject produced 'the two huge red sofas', which the GS associates with the set $\{sofa, red, large\}$. Suppose our algorithms describe the target as:

Output from x : $\{sofa, red, top\}$
Output from y : $\{sofa, red, large, top\}$

Which of these algorithms matches the GS best? Algorithm y adds a property (perhaps overspecifying even more than the GS). Algorithm x has the same length as the GS, but replaces one property by another. Several reasonable ways of assessing the differences can be devised, one of which is Levenshtein distance (which suggests preferring y over x, since the latter involves a deletion *and* an addition) (Levenshtein, 1966). We also intend to examine how often the GS over- or underspecifies where the algorithm does not.

4 Conclusion

Corpora can be an invaluable resource for NLG as long as the necessary contextual information and the conditions under which the texts in a corpus were produced are known. We believe that controlled and balanced experiments are needed for building semantically transparent resources, whose construction we have discussed. As shown in this paper, evaluation of algorithms against the number of gold standards obtained with such a corpus needs careful consideration.

Evaluation of GRE – and NLG systems more generally – would benefit from more investigation of the differences between readers and producers. In future work, we intend to follow up with a reader-oriented experiment in which we test the speed and/or accuracy with which the output of different GRE algorithms is understood by subjects. The dependent variables here will be non-linguistic (perhaps involving subjects clicking on pictures of presumed target referents). This illustrates a more general issue in this area, namely that corpora should, in our view, only be a starting point, with which data of different kinds can be associated.

5 Acknowledgments

Thanks to Ehud Reiter, Richard Power and Emiel Krahmer for useful comments. This work is part of the TUNA project (`http://www.csd.abdn.ac.uk/research/tuna/`), funded by the EPSRC in the UK (GR/S13330/01).

References

[Arts2004] A. Arts. 2004. *Overspecification in Instructive Texts*. Ph.D. thesis, Tilburg University.

[Dale and Reiter1995] R. Dale and E. Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 18:233–263.

[van Deemter2002] K. van Deemter. 2002. Generating referring expressions: Boolean extensions of the incremental algorithm. *Computational Linguistics*, 28(1):37–52.

[Gupta and Stent2005] S. Gupta and A. J. Stent. 2005. Automatic evaluation of referring expression generation using corpora. In *Proceedings of the 1st Workshop on Using Corpora in NLG, Birmingham, UK.*

[Jordan and Walker2000] P. Jordan and M. Walker. 2000. Learning attribute selections for non-pronominal expressions. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

[Levenshtein1966] V. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710.

[Reiter1990] E. Reiter. 1990. The computational complexity of avoiding conversational implicatures. In *Proceedings of the 28th ACL Meeting*, pages 97–104. MIT Press.