A Hearer-oriented Evaluation of Referring Expression Generation *

Imtiaz H. Khan, Kees van Deemter, Graeme Ritchie, Albert Gatt, Alexandra A. Cleland
University of Aberdeen, Aberdeen, Scotland, United Kingdom
{i.h.khan,k.vdeemter,g.ritchie,a.gatt,a.cleland}@abdn.ac.uk

Abstract

This paper discusses the evaluation of a Generation of Referring Expressions algorithm that takes structural ambiguity into account. We describe an ongoing study with human readers.

1 Introduction

In recent years, the NLG community has seen a substantial number of studies to evaluate Generation of Referring Expressions (GRE) algorithms, but it is still far from clear what would constitute an optimal evaluation method. Two limitations stand out in the bulk of existing work. Firstly, most existing evaluations are essentially speaker-oriented, focussing on the degree of "human-likeness" of the generated descriptions, disregarding their effectiveness (e.g. Mellish and Dale (1998), Gupta and Stent (2005), van Deemter et al. (2006), Belz and Kilgarriff (2006), Belz and Reiter (2006), Paris et al. (2006), Viethen and Dale (2006), Gatt and Belz (2008)). The limited number of exceptions to this rule indicate that the differences between the two approaches to evaluation can be substantial (Gatt and Belz, 2008). Secondly, most evaluations have focussed on the semantic content of the generated descriptions, as produced by the Content Determination stage of a GRE algorithm; this means that linguistic realisation (i.e. the choice of words and linguistic constructions) is usually not addressed (exceptions are: Stone and Webber (1998), Krahmer and Theune (2002), Siddharthan and Copestake (2004)).

Our aim is to build GRE algorithms that produce referring expressions that are of optimal benefit to a hearer. That is, we are interested in generating descriptions that are easy to read and understand. But the readability and intelligibility of a description can crucially depend on the way in which it is worded. This happens particularly when there is potential for misunderstanding, as can happen in the case of attachment and scope ambiguities.

Suppose, for example, one wants to make it clear that all radical students and all radical teachers are in agreement with a certain idea. It might be risky to express this as *'the radical students and teachers are agreed'*, since the reader[1] might be inclined to interpret this as pertaining to all teachers rather than only the radical ones. For this reason, a GRE program might opt for the longer noun phrase *'the radical students and the radical teachers'*. But because this expression is lengthier, the choice involves a compromise between comprehensibiliity and brevity, a special case of a difficult trade-off that is typical of generation as well as interpretation of language (van Deemter, 2004).

We previously reported the design of an algorithm (based on an earlier work on expressions referring to sets (Gatt, 2007)), which was derived from experiments in which readers were asked to express their preference between different descriptions and to respond to instructions which used a variety of phrasings (Khan et al., 2008). Here we discuss the issues that arise when such an algorithm is evaluated in terms of its benefits for readers.

2 Summary of the algorithm

In order to study specific data, we have focussed on the construction illustrated in Section 1 above: potentially ambiguous Noun Phrases of the general form *the Adj Noun$_i$ and Noun$_j$*. For such phrases, there are potentially two interpretations: *wide scope* (Adj modifies both Noun$_i$ and Noun$_j$) or *narrow scope* (Adj modifies Noun$_i$ but not Noun$_j$).

Our algorithm starts from an unambiguous set-theoretic formula over lexical items (i.e. words

* This work is supported by a University of Aberdeen Sixth Century Studentship, and EPSRC grant EP/E011764/1.

[1] In this paper, we use the word reader and hearer interchangeably.

Proceedings of the 12th European Workshop on Natural Language Generation, pages 98–101,
Athens, Greece, 30 – 31 March 2009.

have already been chosen), and thus has to choose between a number of different realisations. The possible phrasings for the wide scope meaning are: (1) *the Adj Noun*$_1$ *and Noun*$_2$, (2) *the Adj Noun*$_2$ *and Noun*$_1$, (3) *the Adj Noun*$_1$ *and the Adj Noun*$_2$, and (4) *the Adj Noun*$_2$ *and the Adj Noun*$_1$. For narrow scope, the possibilities are: (1) *the Adj Noun*$_1$ *and Noun*$_2$, (2) *the Noun*$_2$ *and Adj Noun*$_1$, (3) *the Adj Noun*$_1$ *and the Noun*$_2$, and (4) *the Noun*$_2$ *and the Adj Noun*$_1$. For our purposes, (1) and (2) are designated as 'brief', (3) and (4) as 'non-brief' (that is, 'brevity' has a specialised sense involving the presence/absence of '*the*' and possibly *Adj* before the second *Noun*). Importantly, the 'non-brief' expressions are syntactically unambiguous, but the 'brief' NPs are potentially ambiguous, and hence are the focus of attention in this work.

Our algorithm is based on certain specific hypotheses (from the earlier experiments) which make crucial use of corpus data concerning the frequency of two types of collocations: the collocation between an adjective and a noun, and the collocation between two nouns. At a broader level, we hypothesise: *the most likely reading of an* NP *can be predicted using corpus data (Word Sketches (Kilgarriff, 2003))*. The more specific hypotheses derive from earlier work by Kilgarriff (2003) and Chantree et al. (2006), and were further developed and tested in our previous experiments. The central idea is that this statistical information can be used to predict a 'most likely' scoping (and hence interpretation) for the adjective in the 'brief' (i.e. potentially ambiguous) NPs. We define an NP to be *predictable* if our model predicts a single reading for it; otherwise it is *unpredictable*. Hence, all 'non-brief' NPs are predictable (being unambiguous), but only some of the 'brief' ones are predictable.

In a nutshell, *the model underlying our algorithm prefers predictable expressions to unpredictable ones, but if several of the expressions are predictable then brief expressions are preferred over non-brief.*

3 Aims of the study

We want to find out whether our generator makes the best possible choices (for hearers) from amongst the different ways in which a given description can be realised. But although our algorithm uses sophisticated strategies for avoiding noun phrases that it believes to be liable to misunderstanding, misunderstandings cannot be ruled out, and if a hearer misunderstands a noun phrase then secondary aspects such as reading (and/or comprehension) speed are of little consequence. We therefore plan first to find out the likelihood of misunderstanding. For this reason, we will report on the degree of accuracy, as a percentage of times that a participant's understanding of an expression that we label as predictable fails to match the interpretation assigned by our model. Additionally, we shall statistically test two hypotheses:

Comprehension Accuracy 1: Predictable expressions are more often interpreted in agreement than in disagreement with the model.

Comprehension Accuracy 2: There is more agreement among participants on the interpretation of predictable expressions than of unpredictable expressions.

We will not only test the comprehensibility of the expressions generated by our algorithm, but their readability and intelligibility as well. This is necessary because the experiments which led to the algorithm design considered only certain aspects of the hearer's reaction to NPs (e.g. metalinguistic judgements about a participant's *preferences*) and we wish to check these comprehensibility/brevity facets from a different, perhaps psycholinguistically more valid, perspective. It is also necessary because avoidance of misunderstandings is not the only decisive factor: if several of the expressions are predictable then our algorithm chooses between them by preferring brevity. But why is brief better than non-brief? Taking readability and intelligibility together as 'processing speed', our third hypothesis is:

Processing speed: Subjects process predictable brief expressions more quickly than predictable non-brief ones.

Confirmation of this hypothesis would be a strong indication that our algorithm is on the right track, particularly if the degree of accuracy (see above) turns out to be high. Processing speed is a complex concept, but we could decompose it as 'reading speed' and 'comprehension speed', permitting us to examine reading and comprehension separately. We intend to see what evidence there is for the following additional propositions, which will be tested solely to aid our understanding.

Reading Speed:
RS1: Subjects read predictable brief NPs more quickly than unpredictable brief ones.
RS2: Subjects read unpredictable brief NPs more quickly than predictable non-brief ones.
RS3: Subjects read predictable brief NPs more quickly than predictable non-brief ones.

Comprehension Speed:
CS1: Subjects comprehend predictable brief NPs more quickly than unpredictable brief ones.
CS2: Subjects comprehend predictable non-brief NPs more quickly than unpredictable brief ones.
CS3: Subjects do not comprehend predictable non-brief NPs more quickly than predictable brief ones.
(Remember that, in our restricted set of NPs, a phrase cannot be both 'unpredictable' and 'non-brief'.) Rejection of any of these statements will not count against our algorithm.

4 Sketch of experimental procedure

Participants will be presented with a sequence of trials (on a computer screen), each of which consists of a lead-in sentence followed by a target sentence and a comprehension question that relates to the two sentences together. The target sentence might for example say *'the radical students and teachers were waving their hands'*. The comprehension question in this case could be *'Were the moderate teachers waving their hands?'*. As both the target sentence and the comprehension question make use of definite NPs (e.g. *'the moderate teachers'*), it is necessary to ensure any presuppositions about the existence of the referent set are met, without biasing the answer. For this reason, the target sentence is preceded by a lead-in sentence to establish the existence of the sets within the discourse (here, *'there were radical and moderate people in a rally'*).

Given this set-up we are confident that we can identify, from a participant's yes/no answer, whether the NP in the target sentence was assigned a narrow-scope or a wide-scope reading for the adjective. The computer will record the participant's response as well as the length of time that the participant took to answer the question. We will use Linger[2] for presentation of stimuli. Pilots suggest that the complexity of the trials makes it advisable to use *masked sentence-based* self-paced reading, in which every press of the space bar reveals the next sentence and the previous sentence is replaced by dashes.

The choice of nouns and adjectives (to construct NPs) is motivated by the fact that there is a balanced distribution of NPs in each of the following three classes. Wide scope class is the one for which our model predicts a wide-scope reading; narrow scope class is the one for which our model predicts a narrow-scope reading; and ambiguous class is the one for which our model fails to predict a single reading (Khan et al., 2008).

5 Issues emerging from this study

The design of this experiment raised some difficult questions, some quite unexpected:

1. The quality of the output of a generation algorithm might appear to be a simple and well-understood concept. However, output quality is multi-faceted, because an expression may be easy to read but difficult to process semantically, or the other way round. A thorough output evaluation should address both aspects of quality, in our view.

2. If both reading and understanding are addressed, this raises the question of how these two dimensions should be traded off against each other. If one algorithm's output was read more quickly than that of another, but understood more slowly than the second, which of the two should be preferred? Perhaps there is a legitimate role here for metalinguistic judgments after all, in which participants are asked to express their preference between expressions (see Paraboni et al. (2006) for discussion)? An alternative point of view is that these questions are impossible to answer independent of a realistic setting in which participants utter sentences with a concrete communicative purpose in mind. If utterances were made in order to accomplish a concrete task (e.g., to win a game) then *task-based* evaluation would be possible.

3. Even though this paper has not focussed on details of experimental design and analysis, one difficulty is worth mentioning: given the grammatical options between which the generator is choosing, only three types of situations are represented: a description can be brief and predictable (e.g. using 'the old men and women' to convey wide scope, since the adjective is predicted by our algorithm to have wide scope), brief and unpredictable (e.g. 'the rowing boats and ships' for wide scope, given

[2] http://tedlab.mit.edu/~dr/Linger/

a prediction of narrow scope), or non-brief and predictable (e.g. 'the old men and the old women' for wide scope). It might appear that there exists a fourth option: non-brief and unpredictable. But this is ruled out by our technical sense of 'non-brief': as noted earlier, 'non-brief' NPs do not have the scope ambiguity. Because of this "missing cell", it will not be possible to analyse our data using an ANOVA test, which would have automatically taken care of all possible interactions between comprehensibility and brevity. A number of different tests will be used instead, with Bonferroni corrections where necessary.

6 Conclusion

Human-based evaluation is gaining considerable popularity in the NLG community. Whereas evaluation of GRE has mostly been speaker-oriented, the present paper has explored a plan for an experimental hearer-oriented evaluation. The main conclusion is that hearer-based evaluation is difficult because the quality of a generated expression can be measured in different ways, whose results cannot be assumed to match. One factor we have not examined is the notion of *fluency*: it is possible that our algorithm will sometimes choose a word order (e.g. 'the women and old men') that is relatively infrequent, and therefore lacking in fluency. Such situations might lead to longer reading times.

References

A. Belz and A. Kilgarriff. 2006. Shared-task evaluations in HLT: Lessons for NLG. In *Proceedings of the 4th International Conference on Natural Language Generation*, pages 133–135.

A. Belz and E. Reiter. 2006. Comparing automatic and human evaluation of NLG systems. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 313–320, Trento, Italy, 3-7 April.

F. Chantree, B. Nuseibeh, A. de Roeck, and A. Willis. 2006. Identifying nocuous ambiguities in requirements specifications. In *Proceedings of 14th IEEE International Requirements Engineering conference (RE'06)*, Minneapolis/St. Paul, Minnesota, U.S.A.

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 NLG*.

A. Gatt. 2007. *Generating Coherent References to Multiple Entities*. Ph.D. thesis, University of Aberdeen, Aberdeen, Scotland.

S. Gupta and A. Stent. 2005. Automatic evaluation of referring expression generation using corpora. In *Proceedings of the Workshop on Using Corpora for Natural Language Generation*, pages 1–6.

I. H. Khan, K. van Deemter, and G. Ritchie. 2008. Generation of referring expressions: Managing structural ambiguities. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-8)*, pages 433–440, Manchester.

A. Kilgarriff. 2003. Thesauruses for natural language processing. In *Proceedings of NLP-KE*, pages 5–13, Beijing, China.

E. Krahmer and M. Theune. 2002. Efficient context-sensitive generation of referring expressions. In K. van Deemter and R. Kibble, editors, *Information Sharing: Reference and Presupposition in Language Generation and Interpretation, CSLI Publications*, pages 223–264.

C. Mellish and R. Dale. 1998. Evaluation in the context of natural language generation. *Computer Speech and Language*, 12(4):349–373.

I. Paraboni, J. Masthoff, and K. van Deemter. 2006. Overspecified reference in hierarchical domain: measuring the benefits for readers. In *Proceedings of the Fourth International Conference on Natural Language Generation(INLG)*, pages 55–62.

C. Paris, N. Colineau, and R. Wilkinson. 2006. Evaluations of NLG systems: Common corpus and tasks or common dimensions and metrics? In *Proceedings of the 4th International Conference on Natural Language Generation*, pages 127–129.

A. Siddharthan and A. Copestake. 2004. Generating referring expressions in open domains. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics Annual Conference (ACL-04)*.

M. Stone and B. Webber. 1998. Textual economy through close coupling of syntax and semantics. In *Proceedings of the Ninth International Workshop on Natural Language Generation*, pages 178–187, New Brunswick, New Jersey.

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*, pages 130–132.

K. van Deemter. 2004. Towards a probabilistic version of bidirectional OT syntax and semantics. *Journal of Semantics*, 21(3):251–281.

J. Viethen and R. Dale. 2006. Towards the evaluation of referring expression generation. In *Proceedings of the 4th Australasian Language Technology Workshop*, pages 115–122, Sydney, Australia.