Corpus-based evaluation of Referring Expressions Generation

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

1 Introduction

Corpus-based evaluation of NLP systems has become a dominant methodology. Typically, some metric is invoked to evaluate the results produced by a system against a 'gold standard' represented in the corpus. Despite growing recognition of the importance of empirical evaluation in NLG, resources and methodologies for evaluation of Generation of Referring Expressions (GRE) are in their infancy (but c.f. Viethen and Dale (2006)), although this area has been studied intensively since the publication of the Incremental Algorithm (IA) by Dale and Reiter (1995). This contribution describes some of the difficulties which inhere in any corpus-based evaluation exercise involving GRE, as well as a methodology to create a corpus aimed at overcoming these difficulties.

GRE is a **semantically intensive** task. Given an intended referent, a GRE algorithm searches through a knowledge base (KB) to find a combination of properties that uniquely identifies the referent. In order to apply the 'human gold standard' rationale of a corpus-based evaluation to this task, the corpus in question must satisfy at least the following:

1. **Semantic transparency**:
   (a) The domain knowledge of authors in the corpus must be known in advance, so that the algorithm is exposed to exactly the same knowledge. Deviations from such knowledge by humans must be clearly indicated.
   (b) If it is 'standard' GRE that is being evaluated, where output is a semantic or logical form, the corpus should permit the compilation of a normalised logical form from the human data (i.e., abstract away from variations in syntactic and lexical choice).
2. **Pragmatic transparency**:
   (a) If it is 'standard' GRE that is being evaluated, then the communicative intentions of authors in the corpus must be restricted as far as possible to the *identification* intention.
   (b) The communicative situation in which descriptions are produced must be controlled. For instance, a fault-critical situation might elicit more informative descriptions than a non-fault-critical one, which would affect the performance of algorithms in the evaluation.

The rest of this contribution describes our methodology to construct and annotate the TUNA Reference Corpus (TRC). Since its introduction in van Deemter *et al.* (2006a), the TRC has been completed, and consists of ca. 1800 descriptions with annotations about domain knowledge, semantics, and some aspects of communicative context.

2 A corpus for GRE

The TRC was constructed by eliciting descriptions of objects in a controlled experiment, conducted over the internet over a period of three months. The structure of the corpus is shown below, with reference to the experimental conditions manipulated.

	+FC		-FC		
domain	**sing**	**plur**	**sing**	**plur**	**total**
household	210	390	105	195	900
photographs	180	360	90	180	810

Subjects interacted with a computer system and referred to objects in domains where the precise combination of properties that was minimally required to identify the objects was known in advance. Two domains were used, one consisting of artificially constructed pictures of household items, the other of real photographs of people. It was made clear to subjects that they had to identify objects for the system, which in turn 'interpreted' their description and removed objects from the screen. Some of the subjects were placed in a fault-critical situation (+FC) and were told that the system was being tested for use in critical situations where errors could not be corrected; for the other, non-fault-critical situation (-FC), subjects were given the opportunity to correct the system's mistakes by clicking on the correct targets. Descriptions were to both singular and plural referents, and also varied in whether or not subjects could use locative expressions.

The corpus is fully annotated in an XML representation designed to meet the four desiderata outlined above; see (van Deemter et al., 2006b) for details. Description are paired with an explicit domain representation (entities and their attributes) which also indicates the communicative situation (±FC). Domain properties are tagged with an `ATTRIBUTE` tag, which takes a `name` and a `value`. The logical form of a description is indicated by means of a `DESCRIPTION` tag. An example of the annotation for the description *the small desk and the red sofa* is shown below.

```
<DESCRIPTION NUM='PLURAL'>
<DESCRIPTION NUM='SINGULAR'>
<DET value='definite'>the</DET>
<ATTRIBUTE name='size' value='small'>small</ATTRIBUTE>
<ATTRIBUTE name='type' value='desk'>desk</ATTRIBUTE>
</DESCRIPTION>
and
<DESCRIPTION NUM='SINGULAR'>
<DET value='definite'>the</DET>
<ATTRIBUTE name='colour' value='red'>red</ATTRIBUTE>
<ATTRIBUTE name='type' value='sofa'>sofa</ATTRIBUTE>
</DESCRIPTION>
</DESCRIPTION>
```

Using the `DESCRIPTION` tag, a logical form can be compiled by the recursive application of a finite set of rules. Thus, `ATTRIBUTE`s within a `DESCRIPTION` are conjoined; sibling `DESCRIPTION`s are disjoined. Attribute names and values are normalised to match those in the domain, irrespective of the wording used by an author. For example, the above annotation is compiled into $(small \wedge desk) \vee (red \wedge sofa)$.

3 GRE Evaluation

We have used the corpus to conduct an evaluation of the IA against some earlier algorithms, whose perceived shortcomings the IA was designed to address (Gatt et al., In preparation). Logical forms compiled form human-authored descriptions were compared to those generated by an algorithm within the same domain.

Because domain properties are known, human-algorithm comparisons can be based on various metrics, for example, (dis-)similarity of sets of attributes using metrics such as some version of of edit distance or the Dice coefficient. Moreover, the design of an evaluation study can vary. For instance, it is possible to compare an algorithm to a single subject in the corpus, or to an average of all descriptions in the corpus. Overall, a corpus built in line with the requirements outlined in this paper will provide the possibility of more refined algorithm evaluations compared to those conducted in the past. We plan to make this corpus available to the research community in the near future.

References

R. Dale and E. Reiter. 1995. Computational interpretation of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(8):233–263.

A. Gatt, I. van der Sluis, and K. van Deemter. In preparation. Assessing algorithms for the generation of referring expressions, using a semantically and pragmatically transparent corpus.

K. van Deemter, I. van der Sluis, and A. Gatt. 2006a. Building a semantically transparent corpus for the generation of referring expressions. In *Proc. INLG-06.*

K. van Deemter, I. van der Sluis, and A. Gatt. 2006b. Manual for the tuna corpus: Referring expressions in two domains. Technical report, University of Aberdeen.

J. Viethen and R. Dale. 2006. Algorithms for generating referring expressions: Do they do what people do? In *Proc. INLG-06.*