The Anaphor-Antecedent Match: Issues for Referring Expressions Generation

Diego Jesus de Lucena, Ivandré Paraboni

School of Arts, Sciences and Humanities - University of São Paulo (USP / EACH)
Av. Arlindo Bettio, 1000 São Paulo Brazil
{diego.si , ivandre } @ usp.br

Abstract. The design of referring expressions generation algorithms that produce descriptions closer to human performance may benefit from evidence on how human speakers actually interpret these descriptions. To shed light on this issue, in this paper we present an experiment to measure the time required for the interpretation of matching and non-matching anaphoric definite description pairs. Preliminary results suggest that non-matching descriptions, or those that share fewer attributes with the antecedent term, seem to be more difficult to interpret, an insight that may be potentially useful to the design of novel referring expressions generation algorithms.

Keywords: Natural Language Generation, Anaphora, Referring Expressions.

1 Introduction

The Generation of Referring Expressions - the task of providing linguistic labels for discourse objects - is central to the study and development of Natural Language Generation (NLG) systems. In particular, the computation of the semantic contents to be realised as definite descriptions (e.g., "Please press *the Play button*") known as the attribute selection (AS) task has been a very active research field in NLG [1,2,3], and it is the focus of the present work as well.

Attribute selection is the computational task of determining the appropriate contents of a definite description. By 'appropriate' it is generally understood that the selected attributes may be (at the very least) relevant in the sense that they help ruling out potential distractors (i.e., any discourse object other than the intended referent in the given context) whilst preventing false conversational implicatures as defined by H. P. Grice [6]. For example, consider the following domain entities and their properties adapted from [1]:

Example 1. A referential context from [1].

```
Entity1: (type, dog), (size, small), (colour, black)
Entity2: (type, dog), (size, large), (colour, white)
Entity3: (type, cat), (size, small), (colour, black)
```

In the above, we may for example refer to *Entity1* using a uniquely distinguishing description such as "the small dog". The reference to the *size* attribute rules out

Entity2, and the reference to the type attribute rules out *Entity3*. Similarly, *Entity2* could be described unambiguously as "the large dog" and *Entity3* simply as "the cat". Additional attributes are not necessary in this context, and they may indeed suggest a false conversational implicature. For instance, a description such as "the small *black* dog", in which the reference to *colour* is made redundant by the fact that we also refer to *size* may suggest that there is another 'black dog' in the context.

The work in [1] presents one of the best-known AS algorithms to date, the Dale & Reiter Incremental algorithm. The input to the algorithm is a set *C* representing a context with domain objects with their corresponding semantics (attribute-value pairs as in 'size-large'), and the target object *r* to be described by means of a definite (or indefinite) description. The goal of the algorithm is to compute a list of attributes *L* such that *L* denotes the target object *r* and no other distractor in *C*. In doing so, the description *L* is said to be unambiguous and, provided that all attributes in *L* are discriminatory[1], free from false conversational implicatures.

Attribute selection is however much more than simply guaranteeing uniqueness and avoiding false implicatures: in order to select the appropriate attribute set to make a particular referring expression, AS algorithms are also required to appear *plausible*, that is, producing descriptions that resemble those that a human speaker would produce under the same circumstances. In other words, for practical NLG applications it is not sufficient to produce an unambiguous, implicature-free description: the description is required to be as close as possible to what a human speaker would utter.

What actually drives attribute selection as performed by human speakers is an open research question. To shed light on this issue we will focus on the cognitive load of the AS task, presenting an experiment to collect evidence on how anaphoric expressions are *interpreted*, so that these results may be used in the (NLG) design of future AS algorithms (the algorithms themselves are not presently discussed.) Our underlying assumption is that certain attributes may be more difficult to interpret than others, and that for that reason they are best avoided by AS algorithms that intend to pay regard to humanlikeness.

2 Experiment Design

We designed an experiment to measure the time required for the interpretation of matching and non-matching anaphoric definite description pairs as performed by human readers. Subjects were exposed to a number of description pairs and were asked to decide whether each pair comprised a match or mismatch situation.

The purpose of our experiment is twofold. First, we would like to test whether interpreting a match (e.g., an anaphor and its antecedent term) is faster than interpreting a mismatch (e.g., an anaphor and a false candidate term.) Second, focusing on the mismatch cases, we would like to test whether the interpretation of more overlapping description pairs is faster than less overlapping ones. For example,

[1] However, in the Incremental approach there is no guarantee that every attribute in *L* will remain discriminatory. The (incremental) selection of an attribute may actually render a previous attribute redundant, a problem that is not dealt with in [1] for reasons of computational complexity.

"the black cat" and "the small black cat in the box" are matching descriptions that share one overlapping attribute (*colour*) besides the basic attribute *type*.

Subjects.26 native speakers of Brazilian Portuguese.

Procedure. Each of the subjects was shown 26 description pairs in random order on a computer screen, one pair at a time. Two description pairs were presented only for practice, and the reminder 24 descriptions make our research statements.

Subjects were instructed to keep their fingers touching the 'S' and 'N" keys on the computer keyboard (which stand for Portuguese 'yes' and 'no') during the entire experiment. After a number of practice examples, the experiment was started and had to be completed without interruptions. For each description pair (*r1*, *r2*), subjects were asked to decide whether *r1* and *r2* would match (i.e., whether an object described as *r1* could also be described as *r2*) by pressing the yes/no keys as soon as possible[2]. All interpretation times were recorded with millisecond precision using DMDX [5] from the moment in which the descriptions were displayed on screen until the subject pressed a valid yes/no key.

We would like to evaluate matching and mismatching description pairs of various degrees of similarity and length. To this end, we used descriptions conveying from 1 to 3 attributes (*colour*, *size* and *location*) besides the basic *type* attribute that forms the head noun of the description. The attribute *colour* was always present, and when a second attribute was to be included, this could be either *size* or *location* as in "the small black cat" or "the black cat in the box" (both alternatives were tested.) The *r1* description could convey from 1 to 3 attributes, whereas *r2* always had maximum length. By varying the contents of *r1*, we tested both matching description pairs conveying 1, 2, and 3 overlapping attributes, and also mismatches involving 1, 2 and 3 conflicting attributes.

Pairing Type (match/mismatch), *Attribute Set* (1-3 attributes) and *r1-length* (1-3 attributes) make our three experiment variables. Their meaningful combinations are summarized below (notice that *Attribute Set* cannot be larger than *r1-length*), accompanied by their corresponding instances in the experiment, making 24 research statements in total. Some statements occur more than others to take full advantage of the experiment setting, and also to reuse the data from [4] as we discuss later.

[2] As in this paper we do not discuss error analysis, wrong answers were simply computed as if taking longer to be interpreted than the correct answers.

Table 1. Research statements 1-8 and their corresponding instances in the experiment.

#	Pairing Type	Attrib.Set	r1-length	Instances
1	match	1	1	11,23
2	match	2	2	12,13,24,25
3	match	3	3	14,26
4	mismatch	1	1	3,7,15,19
5	mismatch	1	2	9,21
6	mismatch	2	2	4,5,8,16,17,20
7	mismatch	2	3	10,22
8	mismatch	3	3	6,18

Research questions. We intend to measure interpretation times for both matching descriptions, and for non-matching descriptions of different degrees of overlap. This gives rise to the following two research hypotheses (the notation *avg*(*x*) stands for the average interpretation time in statement #*x* as seen in previous Table 1.)

h1: *Identifying a pair of matching description is faster than identifying a non-matching pair:*

$$avg(1) < avg(4),\ avg(2) < avg(5,6),\ avg(3) < avg(7,8).$$

This hypothesis states that antecedent terms (as in matching description pairs) require less cognitive effort than false candidate terms (as in a mismatch.) The test will be carried out by comparing the interpretation times of matching description pairs of *r1-length* = 1, 2 and 3 (that is, statements of type 1, 2 and 3), with non-matching descriptions of corresponding length (statements 4,5,6,7 and 8.) We expect the average interpretation time of matching descriptions to be shorter.

h2: *Identifying a mismatch is faster when the descriptions have more overlapping attributes:*

$$avg(5) < avg(6),\ avg(7) < avg(8).$$

This hypothesis states that false candidates are easier to interpret when they share more information with the anaphor, that is, the greater the mismatch between the descriptions, the longer the interpretation time. This hypothesis was first hinted at by the experiment in [4], and we presently intend to provide further evidence on this issue by removing the referring expression pairs from the surrounding text (hence eliminating the effect of search) and focusing on reading times only. Notice however that the anaphoric relation between the antecedent term and the anaphor (when applicable) is kept intact.

This hypothesis will be tested by comparing the interpretation times of non-matching description pairs of *r1-length* = 2 and 3 in descriptions with more or less overlap. In the first case, we will compare the use of one and two overlapping attributes (statements 5 and 6), and in the second case we will compare the use of two and three overlapping attributes (statements 7 and 8.) We expect all average interpretation times to decrease as the descriptions overlap.

Materials. 26 description pairs in Brazilian Portuguese, taken from the data set in [4] in two different domains (descriptions of cats and cars.) Each description pair was displayed in two lines on a blank computer screen, with the *r1* description always on the top row.

3 Results

26 subjects completed the experiment. The results are statistically significant according to a Wilcoxon Signed-ranks test, and confirm both hypotheses *h1* and *h2*.

Table 2. Experiment results.

Hypothesis	N	Nsr	W	z	p
h1	26	26	165	2.09	0.0366
h2	26	13	63	2.18	0.0293

In the above we observe that matching an anaphor to its antecedent term is indeed faster than deciding that a given candidate does not match (*h1*). Moreover, deciding that a candidate term is not the antecedent is faster when the descriptions have more overlapping attributes (*h2*).

4 Discussion

We have presented an experiment on anaphora resolution to measure the time required to interpret matching and non-matching definite description pairs. According to our findings, non-matching descriptions, or those that share fewer attributes with the antecedent term, seem to be more difficult to interpret.

AS algorithms such as those proposed in [7] may only indirectly be considered to account for ease of interpretation at all. For example, some of these algorithms use an AS strategy that combines frequent and highly discriminating attributes. The results of our experiment, however, may suggest the design of AS algorithms that actually do make interpretation easier in a number of ways.

For instance, it may now be possible to define a cost function *d* representing the degree of overlap between a pair of descriptions, and then use *d* to rank the available attributes used by the algorithm. However, we notice that discriminatory attributes (which would have higher costs) are precisely those that are required to single out the intended referent, that is, balancing the costs of interpretation and discriminatory power may not be straightforward. We believe that more research on this issue is still required.

As future work we intend to take these insights into account in the design of novel AS algorithms following an attribute selection policy based on interpretation effort.

Acknowledgments. This work has been supported by CNPq and FAPESP.

References

1. Dale, R. and E. Reiter. Computational interpretations of the Gricean maxims in the generation of referring expressions. Cognitive Science (19) (1995)
2. Krahmer, E., S. van Erk and A. Verleg. Graph based generation of referring expressions. Computational Linguistics, 29(1), pp. 53-72 (2003)
3. Dale, R. and J. Viethen. Referring Expression Generation through Attribute-Based Heuristics. 12th European Workshop on Natural Language Generation, pp. 58-65 (2009)
4. de Lucena, D. J. and Ivandré Paraboni. The Design of an Experiment in Anaphora Resolution for Referring Expressions Generation. Recent Advances in Natural Language Processing (RANLP-2009) Borovets, Bulgaria, 14-16 September, pp. 225-229 (2009)
5. Forster, K. I. and J. C. Forster. DMDX: A Windows display program with millisecond accuracy. Behavior Research Methods, Instruments, & Computers 35, pp. 116–124 (2003)
6. Grice, H. P. Logic and Conversation. In P. Cole and J. L. Morgan (eds.) Syntax and Semantics, Vol. iii: Speech Acts. New York, Academic Press pp. 41-58 (1975)
7. de Lucena, Diego Jesus and Ivandré Paraboni. Combining Frequent and Discriminating Attributes in the Generation of Definite Descriptions. 11th Ibero-American Conference on Artificial Intelligence (IBERAMIA-2008) LNAI vol. 5290, pp. 252-261 (2008) Springer-Verlag Berlin Heidelberg.