NIL: Attribute Selection for Matching the Task Corpus Using Relative Attribute Groupings Obtained from the Test Data

Raquel Hervás   Pablo Gervás
Natural Interaction based on Language (NIL)
Universidad Complutense de Madrid
c/ Profesor José García Santesmases s/n, 28040 Madrid,
Spain
raquelhb@fdi.ucm.es, pgervas@sip.ucm.es

Abstract

The entry presented by the NIL research group of the Universidad Complutense de Madrid adapts existing software previously developed for an application aimed at generating fluent texts for storytelling. The final entry is specifically geared towards system-human match over the challenge corpus. Attributes are selected based on an adaptation of Reiter and Dale's fast efficient algorithm for referring expression generation, using relative groupings of attributes obtained empirically from the training data to determine the order in which they are considered. The results for the development data indicate that the NIL entry is dealing adequately with issues of identification of referents. Results for system-human match are not so good, probably due to the fact that the corpus is a sample of possible references, rather than a selection of 'ideal' references.

1 Introduction

This document is a brief report on the entry for the first NLG Challenge on Attribute Selection for Referring Expressions Generation presented by the NIL (Natural Interaction based on Language) research group of the Universidad Complutense de Madrid. The entry was prepared by adapting existing software previously developed for a broader natural language generation application oriented towards the generation of fluent texts for storytelling. This document outlines the main characteristics of the challenge that were deemed relevant for the adaptation process, a brief sketch of the original software, a description of the adaptations carried out and how the training data were used in that process, and a report on the results obtained over the development data.

The data employed are based on the TUNA Corpus (van Deemter et al., 2006). They contain a domain representation listing all entities and their relevant attributes for a target referent and six distractors, paired with a human-authored description in terms of the attributes to be mentioned to describe the target referent. Two different domains are used, one consisting of digitally constructed furni-ture/household items; another consisting of real photographs of people.

The NIL entry for the challenge was prepared based on TAP (Text Arranging Pipeline), an ongoing software development project. TAP is a set of interfaces that define generic functionality for a pipeline of tasks oriented towards natural language generation, from an initial conceptual input to surface realization as a string, with intervening stages of content planning and sentence planning. TAP is intended for a storytelling application, which requires varying degrees of complexity at the level of discourse structures, referring expressions, lexical choice, aggregation, and surface realization. However, most of the material available was not relevant for the purposes of the challenge, so the NIL entry has been built based on a single module of the pipeline: the Reference Solver, in charge of building appropriate referring expressions for referents described conceptually in a context given by the particular occurrence of a mention to the referent within a discourse.

2 Description of Our Method

For the purpose of the challenge, the fragment of the code of the Reference Solver in charge of attribute selection was isolated, and an interface was provided so that it might be fed with the relevant challenge data in the form of a target and a set of distractors. Such data are indeed handled in the original Reference Solver, together with additional material required to decide whether pronominal or onomastic references are to be used, and whether the reference should be definite or indefinite. The subtask of attribute selection in the original Reference Solver is based on the algorithm described in (Reiter & Dale, 1992).

The original reference solver has been adapted to the task by refining it to meet those three of the four evaluation dimensions considered in this challenge that can be computed automatically from the data (identification, minimality, and system-human match using the Dice coefficient).

The identification evaluation was already met by the original method. The other two evaluation measures are adressed by tailoring appropriately the list of preferred attributes, to be considered in order of preference, as described in (Reiter & Dale, 1992). The attribute selection is guided by the order in which the available attributes are considered.

In the case of the minimality evaluation, the attribute order was decided following the Full Brevity algorithm from Dale (1989). The list of attributes that are considered for the distinguishing description of the target are ordered by their discriminatory power and used in that order until the target is univocally distinguished.

To configure the modified module to improve system-human match using the Dice coefficient, the training data was studied separately depending on the domain (furniture vs. people). Our idea was that not only the set of attributes in both domains was very different, but also that the psychological considerations taken into account for a person when referring to a piece of furniture or another person might be significantly different.

The results obtained (using the training data) for the Dice coefficient using the Reference Solver module adjusted to produce the minimal set of attributes for the expression are shown in Table 1. Considering the low resulting values for the Dice coefficient, we decided to concentrate on improving the Dice coefficient results at the price of sacrificing the minimality of the reference.

	Minimal	Dice
Furniture	100,00%	24,33%
People	100,00%	31,33%

Table 1: Results using the algorithm for the minimal expression for the training data

The six attributes employed in the furniture domain in different orders allow 6! = 720 possible permutations. Results were computed for all the possible permutations of attribute order, by generating the attribute selection corresponding to all the examples in the training corpus using that order. The average of the Dice coefficient results was calculated in each case.

The study of these results revealed a peculiarity of the way the quality of the results depended on the order of consideration of the attributes: it seemed to be dependant on the relative order in which certain ‘groups’ of attributes were considered, rather than the order of attributes in general. In other words, the results were almost the same for certain orders of groups of attributes, independently of the internal order inside these groups.

In the furniture domain the identified groups were [colour, type, size] and [orientation, x-dimension, y-dimension]. This distinction has some kind of psychological plausibility if we consider that one of the groups is more related with the spatial situation of the object, and the other with its own features. It seems possible that different people would feel more comfortable using one or another, depending on their general view of the world.

	Minimal	Dice
Furniture	0,00%	82,45%

Table 2: Best results obtained in the furniture domain

The best results obtained (using the training data) in the furniture domain are shown in Table 2. The order of the attributes was [type, colour, size, ori-entation, x-dimension, y-dimension].

The people domain uses 11 attributes (excluding the type which was not considered as a distinguishing attribute since all referents are persons), which allows for 11! = 39.916.800 possible permutations, too many to be explored exhaustively. Following the intuitions obtained from the furniture domain, we carried out several experiments creating different combinations of the given attributes. Best results were obtained by aggregating the attributes into groups depending on their significance for distinguishing one person from another. For example, to have beard or to wear glasses are usually more striking than to wear a tie (especially if the person is also wearing suit). Four groups were used in the experiments: [hasSuit, hasTie, hasShirt], [hasBeard, hasGlasses, hasHair, hairColour], [age] and [x-dimension, y-dimension, orientation].

	Minimal	Dice
People	42,72%	43,57%

Table 3: Best results obtained in the people domain

The best results obtained (using the training data) in the people domain are shown in Table 3. The order of the attributes used was [hasGlasses, hasBeard, hairColour, hasHair, hasSuit, hasTie, hasShirt, age, x-dimension, y-dimension, orientation]. In this domain, the values for the Dice coefficient and the values for the minimal expression were suprisingly similar. This may be due to the relative values between the number of attributes and the number of distractors considered in each domain. When there are more attributes than distractors, the intuitive reference seems to be more likely to match the minimal reference.

	Identification	Minimal	Dice
Furniture	100%	0,00%	75,21%
People	100%	33,82%	44,78%

Table 4: Results obtained for the development data

	Minimal	Dice
Furniture	100,00%	20,95%
People	100,00%	30,93%

Table 5: Results using the algorithm for the minimal expression for the development data

3 Development Set Results

The results over the development data obtained using the order of attributes that gave best results over the training data in each domain are given in Table 4. Minimal references might have been obtained with no problem (as shown in Table 5) had this been considered a priority. However, they result in significantly lower Dice coefficients for system-human match.

4 Conclusions and Future Work

The results obtained over the development data indicate that the NIL entry is dealing adequately with issues of identification of referents. Minimality has been discarded since it seems incompatible with system-human match. Dice coefficient results for our final method are rather poor when it comes to matching the specific expressions used in the corpus. We consider that better results might have been obtained if the corpus had been filtered to include only a subset of 'correct' references. Under these circumstances, we consider that to improve results in terms of similarity with a corpus would require an initial step of establishing a subcorpus of 'ideal references', and refining the software to obtain those.

As future work we have considered modelling particular styles of generating referring expressions, rather than working towards an ideal generic way. Such a solution would be useful in the context of a storytelling application, since it would provide the means of having different characters speaking with a different 'voice', or of having the same objects described in different ways when seen from the points of view of different characters.

Acknowledgements

This research is funded by the Spanish Ministry of Education and Science (TIN2006-14433-C02-01 project).

References

Dale, R. 1989. Cooking Up Referring Expressions. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver BC, June.

van Deemter, K., van der Sluis, I. & Gatt, A. (2006). Building a semantically transparent corpus for the generation of referring expressions. *Proceedings of the 4th International Conference on Natural Language Generation (Special Session on Data Sharing and Evaluation)*, INLG-06.

Reiter, E., Dale, R. (1992). A fast algorithm for the generation of referring expressions. Proceedings of the 14th conference on Computational linguistics, Nantes, France.