NIL-UCM: Most-Frequent-Value-First Attribute Selection and Best-Scoring-Choice Realization

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

1 Introduction

The NIL entry for the challenge has been constructed upon the general architecture for developing Natural Language Generation systems provided by the TAP project (Gervás, 2007). TAP (Text Arranging Pipeline) is a set of interfaces that define generic functionality for a pipeline of tasks oriented toward natural language generation, from an initial conceptual input to surface realization as a string, with intervening stages of content planning and sentence planning.

The TAP architecture considers three basic stages: content planning, sentence planning and surface realization. Of these, the first stage is not relevant to the challenge tasks. The configuration choices applied to the other two stages to adapt them to the challenge tasks are described below.

2 NIL-UCM-MFVF Entry for Task 1

The NIL-UCM-MFVF for Task 1 applies a Most-Frequent-Value-First method for Attribute Selection. Of the five evaluation dimensions considered in this challenge (Dice, MASI, accuracy, minimality and uniqueness), this method has been designed to address explicitly only three: Dice, MASI and uniqueness. Minimality was abandoned in view of results in previous challenges (Hervás and Gervás, 2007) that showed good minimality results tended to produce low Dice scores. We have also opted for not using *accuracy* evaluation to fit the performance of our system, since the corpus contains a wide range of style of reference and we are interested in providing our system with only a subset of these that ensure correct identification.

2.1 Most-Frequent-Value-First Attribute Selection

The selection algorithm employed is an adaptation of the algorithm described in (Reiter and Dale, 1992). The original algorithm has been modified to allow for a dynamically changing *list of preferred attributes*, which determine the particular order in which attributes are considered to generate the distinguishing expression. This list is constructed dynamically for each reference by computing the probability of occurrence in the corpus of the particular attribute-value pairs associated with the referent, and using those probabilities to rank them into a specific list of preferred attributes. The idea is that attributes should be considered in a particular order depending highly on their values. For example, in the people domain we have observed that almost the 100% of the target entities that have beard (attribute has value `1`) are referred using the attribute `hasBeard`, but when this attribute has value `0` it is never used. For the `hasHair` attribute, the opposite seems to be the case (mentioned only when lacking).

The training data was studied to obtain the probability of occurrence of an attribute given a certain value for it. This probability was calculated using Formula 1:

$$prob_{val_i} = \frac{\sum appsValueInAttSet}{\sum appsValueInTarget} \qquad (1)$$

For each possible value of each of the attributes of the domains, the sum of the appearances of this value in the `ATTRIBUTE-SET` elements (*appsValueInAttSet*) and the sum of the appearances of this value in the attributes of all targets (*appsValueInTarget*) are calculated. The division of these two values is the probability of mentioning an attribute when it has a specific value.

		Dice	MASI	Accuracy	Uniqueness	Minimality
Train.	**Furniture**	79,18%	56,95%	41,69%	100%	0%
	People	69,71%	42,41%	22,99%	100%	0%
	Both	74,80%	50,23%	34,81%	100%	0%
Dev.	**Furniture**	77,55%	53,97%	41,25%	100%	0%
	People	70,86%	42,59%	22,06%	100%	0%
	Both	74,48%	48,75%	32,43%	100%	0%

Table 1:Task 1 results for training and development data

Some examples of the results obtained are that the attribute `hasGlasses` is mentioned in the 60% of the situations when its value is `1`, and in the 0% of the situations when its value is `0`. On the contrary, the attribute `hasShirt` is almost never mentioned (0.8% when its value is `1` and 0% with value `0`).

The only exception in the algorithm is the `type` attribute for the people domain. As every entity in this domain is of type `person`, the attribute selector does not choose this attribute because no distractor is discarded by it. However, the experiments have shown us that in the corpus a lot of descriptions include the type `person` even when it is redundant. Following this idea, our algorithm always includes the `type` in the list of chosen attributes for the people domain.[1]

2.2 Obtained Results

Results obtained over the training and development data are shown in Table 1. As can be seen comparing both tables there are no surprises in the final results: the system gets similar results with both domains and with both the training and development data. These results confirm that the probability of appearance of an attribute depending on its value is more or less the same in the whole corpus.

3 NIL-UCM-BSC Entry for Task 2

The NIL-UCM-BSC for Task 2 applies a Best-Scoring-Choice approach to Realization.

The realization tasks of the 2008 GRE challenge required specific instantiations of the Referring Expression Generation, Syntactic Choice, and Lexicalization stages of the Sentence Planning module of TAP, and it draws on the SurReal (Gervás, 2006) surface realization module. SurReal provides a Java implementation of the surface realization mechanisms of FUF described in Elhadad (Elhadad, 1993), operating over a grammar which follows the notational conventions of the SURGE grammar in Elhadad (Elhadad and Robin, 1996), but it is not systemic in nature. It currently has much smaller coverage than the original, but quite sufficient to deal with the kind of realizations required for the challenge tasks.

3.1 Realization Choices in the Corpus

An analysis of the domain was carried out to ascertain what the various alternatives required for realization were for the given corpus, both in terms of how to realize syntactically the different concepts and what alternative lexicalizations should be considered. With respect to linguistic variation in the form of expression we have distinguished between choices that give rise to different syntactic structures (which we consider as syntactic choices) and choices which give rise to the same syntactic structures but with different lexical items (which we consider as lexical choices).

With respect to the *Referring Expression Generation* stage, the following issues required specific decisions. The use of **determiners** is erratic. Some examples in the corpus use indefinite article, some use definite articles, and some omit the determiners altogether. The corpus shows many cases where **spatial expressions** describing the location of referents are used, many using different systems of reference (north-south vs. top-bottom). The use of **particular features of the object** in its description, as in "the desk with the drawers facing the viewer" or "the chair with the seat facing away". **Comparison**

[1]We have only recently discovered that the surprising difference between NIL-UCM results for the *people* and the *furniture* domains in the 2007 GRE challenge was the mostly due to our not having taken this issue into account at the time. The effect is noticeable only when the type attribute is redundant, as it is in the *people* domain.

with all or some of the distractors are also used, either as adjuncts describing their position relative to other distractors, as in "the blue fan next to the green fan", or as comparative adjectives used for particular attributes, as in "the largest red couch" (and even combinations of the two as in "the smaller of the two blue fans"). Finally, there are samples in the corpus of use of **ellipsis and ungrammatical expressions**. The mention of particular features and the use of comparison would involve operating on more data than are generated in task 1, and the current submission is aimed to interconnection with task 2 for addressing task 3. The issue of ungrammaticality is important since it implies that there is an upper limit to the possible scores that the system may achieve over the corpus under the circumstances, totally unrelated with the correctness of the generated expressions.

With respect to *Syntactic Choice*, some attributes show **more than one possible option** for syntactic realization. The number of alternatives varies from color ("grey chair - chair that is gray"), through beards ("with beard - with the beard - with whiskers - the bearded man - with a beard - with facial hair") to orientation (12 different syntactic alternatives for expressing orientation: back).

There are slight **variations** of *Lexical Choice* over the corpus, as in "sofa - couch - settee - loveseat", "ventilator - fan - windmill" or "man - guy - bloke" (for nouns) and "large - big" or "small - little" (for adjectives). Because it has a significant impact on the edit distance measure, it is also important to consider the existence of a large number of **misspellings** in the corpus. Finally, there are some **conceptual mismatches in annotation**, between the attribute set and the given realization in some cases ("purple - blue", "black and white - grey",...).

3.2 Best Scoring Choice Solution

The solution employed in the present submission for selecting among the features described above implements straight forward realization rather than choice, in the sense in which (Cahill, 1998) uses the terms for lexicalization. To implement real choice the system would have to consider more than one alternative for a specific feature and to select one of them based on some criteria. This has not been done in the present submission. Instead, a single alternative has been implemented for each feature, using it consistently across all samples. The selection of which particular alternative to implement has been done empirically to ensure the best possible score over the training corpus.

3.3 Results and Discussion

Results obtained over the training and development data are shown in Table 2.

		SE distance	Accuracy
Train.	**Furniture**	4,26	14,15%
	People	5,43	9,12%
	Both	4,8	11,82%
Dev.	**Furniture**	4,21	15%
	People	4,94	7,35%
	Both	4,54	11,48%

Table 2:Task 2 results for training and development data

An important point to consider with respect to the current submission is whether a solution implementing real choice would have obtained better results. Such a solution might have benefited from the information that can be extracted from the `ANNOTATED-WORD-STRING` to train a decision procedure on the various features. This has not been addressed in the present submission more for lack of time than lack of conviction on its merit.

Addressing explicitly some of the possible constructions that are described in section 3.1 may also have a positive effect on the results.

4 NIL-UCM-FVBS Entry for Task 3

The NIL-UCM-FVBS entry for Task 3 applies a combination of the Most-Frequent-Value-First method for Attribute Selection and the Best-Scoring-Choice approach to Realization.

The modular architecture of TAP has allowed easy integration for Task 3 of the solution for attribute selection described in section 2, and the solution for realization described in section 3.

4.1 Results and Discussion

Results obtained over the training and development data are shown in Table 3. Comparing both sets of results there are no surprises in the final results: the system gets similar results with both domains and

with both the training and development data. These results confirm that the probability of appearance of an attribute depending on its value is more or less the same in the whole corpus.

		SE distance	Accuracy
Train.	**Furniture**	5,03	5,03%
	People	6,11	5,47%
	Both	5,53	5,24%
Dev.	**Furniture**	5,06	3,75%
	People	6,24	1,47%
	Both	5,60	2,70%

Table 3:Task 3 results for training and development data

The results obtained are a bit lower than the ones obtained by both the attribute selection and realization submodules separately. This is not an unexpected result. Bad choices produced in the attribute selection are propagated through the realization, resulting in accumulated errors in the final evaluation.

However, there are additional shortcomings that arise from considering the general goal of task 3 as a composition of task 2 over task 1. The reduction of the types of expression produced by human subjects to a set of attributes involves in some cases a certain loss of information. This is particularly the case when the human-produced expressions involve attributes for which additional information is provided. This can be seen if the `ANNOTATED-WORD-STRING` is compared with the actual attribute set generated for some of the human-produced expressions. For instance, the corpus contains examples in which the `hasBeard` attribute has a nested attribute that indicates the beard is white. Other examples provide color information on pieces of clothing worn. This information is lost to the realization stage if the data have to go through task 1, which reduces the available format to a set of individual unstructured attributes.

Considering a version of task 3 that allowed full realization directly from input data as considered for task 1, with no requirements on the stages of intermediate representation to be employed in the process, may result in a richer range of realizations, and possibly in improved performance with respect to human evaluation.

In more general terms, it seems that the corpus does contain adequate data for informing system performance at the level of sentence planning subtasks such as lexical choice or syntactic choice. Nevertheless, some of the variations in the corpus, such as the free use of determiners or the flexibility that subjects exhibit in the way they refer to the images do introduce a certain "noise". Instances of these occur when human-produced descriptions involve intense forms of ellipsis, and agrammatical ordering of attributes. Some of these might be reduced if a refined version of the corpus were produced with more control on the experimental settings, to ensure that subjects either described the elements as images or as the things represented in the images, for instance.

Acknowledgments

This research is funded by the Spanish Ministry of Education and Science (TIN2006-14433-C02-01 project) and the UCM and the Dirección General de Universidades e Investigación of the CAM (CCG07-UCM/TIC-2803).

References

Cahill, Lyne. 1998. Lexicalisation in applied NLG systems. *Technical Report* ITRI-99-04.

Elhadad, Michael. 1993. Technical Report CUCS-038-91. Columbia University.

Elhadad, Michael and Robin, Jacques. 1996. Technical Report 96-03. Department of Computer Science, Ben Gurion University.

Gervás, Pablo. 2006. SurReal: a Surface Realization module. *Natural Interaction based on Language Group Technical Report*, Universidad Complutense de Madrid, Spain.

Gervás, Pablo. 2007. TAP: a Text Arranging Pipeline. *Natural Interaction based on Language Group Technical Report*, Universidad Complutense de Madrid, Spain.

Hervás, Raquel and Gervás, Pablo. 2007. NIL: Attribute Selection for Matching the Task Corpus Using Relative Attribute Groupings Obtained from the Test Data. *First NLG Challenge on Attribute Selection for Generating Referring Expressions (ASGRE)*, UCNLG+MT Workshop, Machine Translation Summit XI, Copenhagen.

Reiter, Ehud and Dale, Robert. 1992. A fast algorithm for the generation of referring expressions. *Proceedings of the 14th conference on Computational Linguistics*, pp. 232-238. Association for Computational Linguistics.