Rule-based vs. Probabilistic Surface Realisation of Definite Descriptions

Francis Marques V. dos Santos
University of São Paulo (EACH)
Av. Arlindo Bettio, 1000 -
São Paulo, Brazil
55-11-30911004
fmvsantos@usp.br

Daniel Bastos Pereira
University of São Paulo (EACH)
Av. Arlindo Bettio, 1000 -
São Paulo, Brazil
55-11-30911004
daniel.bastos@usp.br

Ivandré Paraboni
University of São Paulo (EACH)
Av. Arlindo Bettio, 1000 -
São Paulo, Brazil
55-11-30911004
ivandre@usp.br

ABSTRACT

We describe the evaluation work of two standard approaches to the surface realisation of definite descriptions as Portuguese text. Taking as an input a non-linguistic representation of the description to be generated, a rule-based approach makes use of grammar constraints to compute the appropriate surface string, whereas a competing probabilistic model applies n-gram statistics to the same task. The results of both systems are compared against a corpus of human-produced descriptions and their advantages are discussed.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: *language generation*.

General Terms

Algorithms, Languages.

Keywords

Natural Language Generation, Definite Descriptions, Surface Realisation.

1. INTRODUCTION

Natural Language Generation (NLG) systems – which produce textual descriptions from (usually) non-linguistic input data - come to play when simple, 'canned' text is not sufficient, and hence greater (i.e., closer to human performance) linguistic variation is required [4]. Starting from a high-level communicative goal of providing a textual description of a domain entity (e.g., 'john') the NLG system builds up a plan to represent the input data (known as the *Document Planning* task.) The plan is successively refined by performing sentence aggregation, lexicalization and referring expressions generation (comprising the *Microplanning* task) up to the point in which a mapping from semantic representation to text is drawn. This final conversion step, the *Surface Realisation* task, is the focus of the present work.

Surface realisation takes as its input an abstract, language-independent specification of the text to be generated, and produces output text in the target language. The task has been traditionally performed using many of the techniques that are familiar to other NLP fields. Our present work proposes - and subsequently compares - a knowledge-intensive approach to surface realisation based on pre-defined *grammar rules*, and a *probabilistic model* in which no linguistic knowledge is hand-coded. To this end, we shall limit the discussion to the task of providing a surface realisation for *definite descriptions* as in "*the red chair facing backwards*", whose impact on text coherence and cohesion has long been argued for [2] and, accordingly, correspond to one of the essential components of NLG systems.

Both our rule-based and probabilistic models are tested against a corpus of definite descriptions produced by human speakers, from which we expect to gain a better understanding of their individual strengths and weaknesses.

2. INPUT DATA

As input data, we follow [6] and use instances of descriptions taken from the TUNA corpus [1,5], a database of situations of reference collected primarily for the study of reference phenomena and referring expressions generation algorithms. TUNA descriptions were produced by 45 native or fluent speakers of English, participants of a controlled experiment with the sole purpose of identifying each intended referent.

We focus on a subset of TUNA descriptions distributed as training data for the REG-2008 challenge[1], consisting of a collection of 319 descriptions in the Furniture domain. The referents in this domain are pieces of furniture (sofas, desks etc.) of different sizes and colours, presented in a 3 x 5 grid on a computer screen so that their position within the grid could be used as a referable attribute as well.

TUNA descriptions are uniquely identifying sets of semantic properties, each of them represented as an attribute-value pair as in `<NAME="type" VALUE="chair">` in XML format. The following is an example of one such description, which could be realised as "*the large red chair, in the second column on the top*".

```
<DESCRIPTION>
   <ATTRIBUTE ID="a295" NAME="size" VALUE="large" />
   <ATTRIBUTE ID="a296" NAME="colour" VALUE="red" />
   <ATTRIBUTE ID="a297" NAME="type" VALUE="chair" />
   <ATTRIBUTE ID="a299" NAME="y-dimension" VALUE="1" />
   <ATTRIBUTE ID="a301" NAME="x-dimension" VALUE="2" />
</DESCRIPTION>
```

Our goal is to produce a textual description (in Portuguese) from the above semantic representation that should ideally be as close as possible to human performance. To this end, we take advantage of the Portuguese reference set described in [6], hereby called *Reference* set. The data consist of the above 319 instances of TUNA attribute sets, each of them accompanied by a manually-produced realisation as a Portuguese definite description. The Portuguese descriptions were translated from the original

[1] http://www.nltg.brighton.ac.uk/research/reg08

(English) TUNA word strings and subsequently normalized to remove noisy data. For details we report to [6].

3. SURFACE REALISATION MODELS

For the purposes of this work we will consider two approaches to the task of surface realisation of definite descriptions: the probabilistic model described in [6] and a simple rule-based system to be used as a (strong) baseline for the former. In what follows each of them is briefly discussed in turn.

The probabilistic approach consists of a simple application of n-gram statistics to produce the most likely surface string for a given input description (i.e., a set of attribute-value pairs as described in the previous section.) The general approach can be viewed as a simplification of standard statistical NLG techniques such as [7].

The generation task is performed in three steps: first, we compute all possible (unordered) sets of phrases in which the description could be realised. For example, a description comprising two semantic properties `<NAME="colour" VALUE="red">` and `<NAME="type" VALUE="chair">` would be associated with two phrase sets: s1={*vermelho, cadeira*} and s2={*vermelha, cadeira*} to allow for the gender variation of the 'red' colour value in Portuguese (non-Portuguese speakers should note that, unlike s2, s1 does not present the required gender agreement between the two phrases and therefore is ungrammatical.)

Second, we compute (or "overgenerate") all possible permutations of each phrase set (s1 and s2 in the above example) that matched a pre-defined definite description template suitable to Portuguese phrase order, once again with gender variation. In the case of the TUNA Furniture domain, this template is simply a pre-defined ordering of *determiner* + *noun* + *adjectives*. For instance, the above semantic input would have the following four possible realisations: (1) "*a cadeira vermelho*", (2) "*o cadeira vermelho*", (3) "*a cadeira vermelha*" and (4) "*o cadeira vermelha*", in which only (3) provides the necessary gender agreement between the three constituents.

Finally, we decide which of these (e.g., 1-4) alternatives is the most likely output string with the aid of a bigram language model trained on the 40-million words NILC Portuguese corpus [3]. The language model was built using the tool described in [8].

The NILC corpus consists of a large collection of Brazilian newspapers articles which are highly unlikely to contain (unless by chance) descriptions of the kind that we intend to generate (e.g., "the green fan facing left"), as they would be extremely unusual in this domain. Note however that whether a description actually occurs in the corpus or not should have little or no effect in the expected results.

Descriptions produced in this way currently have a number of limitations, chief among them the lack of long-distance gender agreement, as in "*a mesa grande vermelho*" (the large red table) in which "*mesa*" (table, feminine) should agree with (red, feminine) "*vermelha*" in a well-formed Portuguese description.

As an alternative to this approach, we have also developed a set of standard grammar rules to generate Portuguese definite descriptions, taking gender, number and structural constraints into account. The grammar (actually a DCG) covered the entire set of definite descriptions that we intend to generate. The following is an example of a simple grammar rule for generating descriptions such as "*the red chair in the first column*", in which G and N stand for the gender and number agreement that constrains all constituents but the location phrase. Note that, unlike English, the word order in Portuguese requires the adjective to follow the noun (e.g., "*a cadeira vermelha na primeira coluna*".)

```
sn(G,N) -->       determiner(G,N),
                  noun(G,N),
                  quality_adjective(G,N),
                  location_phrase.
```

The grammar rules were embedded in a hybrid Java/Prolog application that takes as its input a set of attributes in XML format and returns a well-formed Portuguese surface string. In this case, agreement is forced by the pre-defined rules and hence no inconsistencies are possible. In cases in which there are more than one possible output string (e.g., due to synonymy) the grammar will return only the first solution found. As we shall discuss later, this arbitrary behaviour will have a number of consequences for our evaluation work.

In both systems, a certain amount of errors is to be expected since the TUNA data include a number of non-standard attribute usages (i.e., expressions that the task participants were not supposed to have used) which our systems could not be expected to handle. Moreover, neither the rule-based or the probabilistic systems were designed to generate descriptions that combine several attributes into a single phrase, e.g., realising both the *x-dimension* and the *y-dimension* attributes as a single reference to a corner, as in "the upper right corner". Although in these cases we notice that referring to a corner was far more frequent in the data, both system will simply produce 1-2-1 mappings as in "the 5th column in the top row". To fully appreciate the extent of this weakness, consider for instance the need to refer to a person using the attributes 'human', 'male' and 'young', in which both systems would simply attempt to describe the intended referent as "the person that is young and male", as opposed to a much more adequate description as "the boy".

4. EVALUATION

Our evaluation work consisted of comparing the set of *Reference* descriptions to those produced by each of the two surface realisation approaches described in the previous section, hereby called *Rule-based System* and *Probabilistic System* sets. More specifically, we computed Accuracy and String-edit scores for each *Reference-System* pair using the *teval* tool provided by the REG-2008 team[1].

String-edit is defined as the number of insert, delete and replace operations required to take place on the *Reference* description to make it identical to a (*Rule-based* or *Probabilistic) System* description, whereas Accuracy shows whether the two word strings are exactly the same. Thus, higher accuracies and lower edit distances are best.

Table 1 summarizes our findings, in which the results for the *Probabilistic* model are taken from [6].

Table 1. Rule-based vs. Probabilistic surface realisation

Evaluation Criteria	Rule-based	Probabilistic
Accuracy	0.69	0.24
String-edit distance	1.26	2.69

The rule-based approach clearly outperforms the probabilistic model according to both criteria, generally producing descriptions that are much closer to those in the *Reference* set. This was, in our view, to be fully expected for a number of reasons.

Firstly, except for the limitations described in the previous section, the rules are nearly ideal in the sense that they capture the precise grammar constraints expressed in the *Reference* set, and were indeed *built from* that data. The accuracy rate below 70% in this case can be almost entirely explained by the overly strict definition of this measure, which is suitable for a referring expressions generation competition (and it was indeed adopted in the REG-2008 Challenge) but it is much less appropriate for our current purposes. Given that Accuracy simply counts the number of matches between *System* and *Reference* sets, any *System* description that is not completely identical (word by word) to its *Reference* counterpart is penalized even if they are synonymous. For example, the comparison between "*the chair in the left column*" and "*the chair in the first column*" reduces the accuracy score even though both descriptions are perfectly acceptable (and found in large numbers in the corpus.)

Secondly, we are aware that our probabilistic approach is still in its infancy, making use of an overly simple bigram model and a relatively small training corpus for current standards in the field. In particular, it is clear that a bigram model cannot handle long-distance gender and number dependencies

On the other hand, one of the greatest benefits commonly associated with statistical NLP is also evident, that is, language-independency. Despite its current limitations, the probabilistic approach is in principle capable of generating text in any arbitrary language as long as a sufficiently large training corpus is provided. By comparison, the rule-based approach would require a language specialist to write new rules from scratch, which may be a costly or labour-intensive work.

5. FINAL REMARKS

We have evaluated two approaches to the surface realisation task of Portuguese definite descriptions: a rule-based and a probabilistic model. Both systems have been tested against a corpus of human-produced definite descriptions and, as expected, the use of rules achieved much higher accuracy than the probabilistic model, although limited to the language that it was originally designed for.

As future work we intend to improve our probabilistic approach by making use of larger corpora and a more expressive (e.g., 3- or 4-gram based) language model, and use the rule-based system as a strong baseline for further analysis.

6. ACKNOWLEDGMENTS

The authors acknowledge support by FAPESP (2006/03941-7), CNPq (484015/2007-9) and the University of São Paulo "Ensinar com Pesquisa" grant.

7. REFERENCES

[1] Gatt, A., I. van der Sluis, and K. van Deemter. 2007. Evaluating algorithms for the generation of referring expressions using a balanced corpus. 11th European Workshop on Natural Language Generation, pp. 49–56.

[2] Halliday, M and R. Hassan. 1976. Cohesion in English. Longman English Language Series 9. Longman, London.

[3] Nunes, Maria das Graças Volpe, Fabiano M. Costa Vieira, Cláudia Zavaglia, Cássia R. C. Sossolote and Josélia Hernandez. 1996. A construção de um léxico para o português do Brasil: lições aprendidas e perspectivas. 2º. Propor. Curitiba, pp. 61-70.

[4] Reiter, E. and Robert Dale. 2000. Building natural language generation systems. Cambridge University Press.

[5] van Deemter, K., I. van der Sluis and A. Gatt. 2006. Building a semantically transparent corpus for the generation of referring expressions. 4th International Conference on Natural Language Generation, INLG-2006 Special session on Data Sharing and Evaluation.

[6] Pereira, D. B. and Ivandré Paraboni. 2008. Statistical Surface Realisation of Portuguese Referring Expressions. 6th International Conference on Natural Language Processing (GoTAL-2008) Gothenburg - Sweden. Springer LNAI vol. 5221. pp. 383-392.

[7] Langkilde, I. and Kevin Knight: 1998. Generation that Exploits Corpus-Based Statistical Knowledge. COLING-ACL 1998. pp. 704-710.

[8] Pereira, Daniel Bastos and Ivandré Paraboni. 2007. A Language Modelling Tool for Statistical NLP. 5th Workshop on Information and Human Language Technology (TIL-2007). Anais do XXVII Congresso da SBC. Rio de Janeiro. pp.1679-1688.