A Testbed for Portuguese Natural Language Generation

Eder Miranda de Novais, Rafael Lage de Oliveira, Daniel Bastos Pereira, Thiago Dias Tadeu, Ivandré Paraboni

[1] EACH, Universidade de São Paulo
Av. Arlindo Bettio, 1000, São Paulo, Brazil

{ eder.novais, rafaellage, daniel.bastos, thiagoo, ivandre }@usp.br

Abstract. *We present a data-text aligned corpus for Brazilian Portuguese Natural Language Generation (NLG) called SINotas, which we believe to be the first of its kind. SINotas provides a testbed for research on various aspects of trainable, corpus-based NLG, and it is the basis of a simple NLG application under development in the education domain.*

1. Introduction

One possible way of building a Natural Language Generation (NLG) system is by following a labour-intensive knowledge engineering approach: after a detailed domain analysis (possibly making use of a corpus of target documents and expertise knowledge), generation rules are hand-crafted to model every decision in the process, from Document Planning to Sentence Realisation (cf. Reiter & Dale, 2000). As an alternative, a machine learning approach may use corpus knowledge also as a knowledge source for decision-making in its own right (and not only in the requirement analysis of the system's development), allowing the development of trainable approaches to NLG.

Despite its many benefits (e.g., robustness, reusability etc.), a major drawback in a corpus-based approach to NLG is precisely the need for annotated data-text corpora. Resources of this kind may be costly to develop or, in the case of Brazilian Portuguese, simply unavailable in a usable form, which may explain why Portuguese NLG is still an underdeveloped research topic. Thus, as an effort to boost work in the field, we present a first data-text aligned corpus of Brazilian Portuguese, called the *SINotas* corpus, which can be seen as a ready-to-use testbed for research on various aspects of corpus-based NLG ranging from content selection to surface realisation (Oliveira et. al., 2009).

2. The *SINotas* Corpus

SINotas is an aligned data-text corpus originally devised as training data for a simple NLG application in which University grades obtained by students are described as short reports generated automatically from raw data (i.e., the numeric grades themselves) taken from their academic records. reports of this kind are potentially useful to both students keen to learn how their professors interpret their efforts, and to the professors themselves who may have an at-a-glance view of the student's progress.

The corpus consists of 241 paired data-text records of students' academic performance data in five courses taught by a single professor (the domain expert[1]) in an academic term and corresponding reports, describing, e.g., figures about attendance records, examination grades at various stages throughout the course, and the average grades attained by the class as a whole.

Text meaning (corresponding to the 'data' portion of each record in the corpus) consists of a set of 14 content messages represented in flat semantics as attribute-value pairs. The following Table 1 describes the messages and their possible values, including the number of instances actually found in the corpus. Attributes assigned 'nulo' (null) values stand for missing or irrelevant information (e.g., if a student has not sat for her substitutive exams, then *sub_aval* will be set to 'nulo'.) Attributes marked as '*' convey particularly sparse and/or heavily imbalanced values.

Table 1. Content messages.

Attribute	Description	Possible values / number of instances
provas_aval	Regular exams grades	nao_realizou(50), muito_abaixo(30), razoável(40), bom_mas_baixo(6), bom(84), muito_bom(19), excelente(12)
provas_turma	Same, as compared to the entire class	nulo(50), abaixo(100), acima(91)
progresso	Overall progress throughout the term	nulo(50), declinio(50), menor_meio(48), maior_meio(65), aumento(28)
sub_aval*	Substitutive exams grades	nulo(223), muito_abaixo (16), abaixo(2), acima(0)
sub_turma*	Same, as compared to the entire class	nulo(214), abaixo(11), acima(16)
eps_aval	Practical exercises grades	nao_realizou(56), muito_abaixo(2), razoável(5), bom_mas_baixo(2), bom(22), muito_bom(33), excelente(121)
dev_ep1	Whether exercises were compulsory	nulo(207), sim(34)
freq_aval	Attendance to the lectures	nulo(188), nenhuma(44), insuficiente(9)
corel_nota_falta*	Lower grades / attendance relation	nulo(215), sim(26)
mf_aval	Final term exams	muito_abaixo(81), razoável(41), bom_mas_baixo(5), bom(70), muito_bom(27), excelente(17)
mf_turma	Same, as compared to the entire class	nulo (58), abaixo(48), acima(135)
rec_aval	Recuperation exams grades	nulo(200), muito_abaixo(17), razoável(8), bom_mas_baixo(0), bom(16), muito_bom(0), excelente(0)
aband_rec*	Abandoned recuperation exams	nulo(235), sim(6)
rec_turma	Same, as compared to the entire class	nulo (204), abaixo(16), media(2), acima(19)

The 'text' portion of each record in *SINotas* contains a short (about 5-sentences long, and to some extent normalised) report conveying a series of statements about the overall progress of the student, such as "You fared well in the regular exams and your grades on this subject were above the average of your class" etc. As in Williams & Reiter (2005), the reports are entirely purpose-made, i.e., written so as to provide training data for a machine-learned NLG application.

Each text was segmented into meaningful units as suggested in Geldof (2003), and infrequent instances were eliminated as required by our target application (which of course reduced the possible linguistic variation of the output.) Put together, the data and the corresponding reports make the *SINotas* corpus, a structured collection of 241 records in XML format. Each record consists of three aligned sections: <*DATA*>

[1] In order to establish mappings from raw data (e.g., students' grades) to semantics (i.e., the interpretation of the data according to a professor), we followed a traditional AI knowledge acquisition methodology (Russel & Norvig, 2003) and collected aligned data-text instances produced by *a single author only*.

conveys the document semantics as a set of 14 messages *m1..m14*; <*TEXT*> represents the target document, segmented and annotated with content messages and part-of-speech information, and <*RST*> conveys its rhetorical structure (Mann & Thompson, 1987). The following is a fragment of one such record.

Table 2. A sample record in SINotas.

```
<RECORD ID="2185480644" SEQ="1">
 <DATA TERM="2" CLASS="1" COURSE="44" GENDER="m">
  <MSG ID="m1" NAME="provas_aval" VALUE="razoavel"/>
  <MSG ID="m2" NAME="provas_turma" VALUE="abaixo"/>
  <MSG ID="m3" NAME="progresso" VALUE="aumento"/>
   {more messages here}
  <MSG ID="m14" NAME="rec_turma" VALUE="media"/>
 </DATA>
 <TEXT>
  <SENTENCE ID="s1" SENTENCE-STRING="seu desempenho nas avaliacoes regulares foi
            razoavel, embora tenha ficado abaixo da media da turma ">
   <SEGMENT ID="sg1" SEGMENT-STRING="seu desempenho nas avaliacoes regulares foi
            razoavel" MSG="m1">
    <SEGMENT-TREE>
     <SN STRING="seu desempenho nas avaliacoes regulares" SOURCE="provas_aval"
            TYPE="atributo" GENDER="masc" NUMBER="sing"/>
     <VP>
      <VERB STRING="foi" BASE="ser" MODE="indicativo" TENSE="preterito"
            GENDER="masc" NUMBER="sing" PERSON="3"/>
      <COMPLEMENT STRING="razoavel" SOURCE="razoavel" TYPE="valor" GENDER="masc"
            NUMBER="sing" POS="adjetivo"/>
     </VP>
    </SEGMENT-TREE>
   </SEGMENT>
   {more segments here}
  </SENTENCE>
  {more sentences here}
 </TEXT>
 <RST>
  <RELATION ID="r0" TYPE="concession" NUCLEUS="s1-sg1" SATELLITE="s1-sg3"
        CONNECTOR="embora"/>
   {more RST relations here}
 </RST>
</RECORD>
```

Since we have abstracted away from the application raw data by modelling the underlying semantics as content messages, *SINotas* is not be regarded as a low-level NLG resource such as the SUMTIME-METEO corpus (Sripada et. al., 2003), which aligns text directly with domain data. In our case, the choice for a higher-level representation aims to enable the quick deployment of NLG applications, and it was also motivated by the need to render the data records anonymous.

3. Discussion

SINotas is intended to be a ready-to-use testbed for (Portuguese) NLG research. Despite the lack of sophistication of the underlying semantics, we expect *SINotas* to be useful for many other corpus-based NLG studies besides our target application, allowing the investigation of various aspects of language generation, from content selection to surface realisation, and based on different levels of knowledge representation. For

example, we may use machine-learning techniques to build NLG models for document planning, sentence planning etc.

Regarding our own research, we have so far used *SINotas* to develop the Document Planning module of an academic reports generator as a series of classifiers. using corpus-based knowledge from *SINotas* we have developed a Content Determination module in which both data interpretation and content selection subtasks (Reiter, 2007) are obtained from decision-tree induction, and the same principle has been applied to Document Structuring to perform both within-sentence and between-sentences structuring. Details are described in Oliveira et. al. (2009).

Acknowledgements

The authors acknowledge support by FAPESP and CNPq.

References

Geldof, S. (2003) “Corpus analysis for NLG”. In: Reiter, E., Horacek, H. and van Deemter, K. (Eds.) 9th European Workshop on NLG, pages 31-38.

Mann, W. C. and S. A. Thompson (1987) “Rhetorical Structure Theory: A Theory of Text Organisation”. L. Polanyi (ed.) The Structure of Discourse. Ablex, Norwood.

Oliveira, Rafael Lage de, Eder Miranda de Novais, Roberto Paulo Andrioli de Araujo and Ivandré Paraboni (2009) “A Classification-driven Approach to Document Planning”. Recent Advances in Natural Language Processing (RANLP-2009), Borovets, Bulgaria.

Reiter, Ehud and Robert Dale (2000) “Building Applied Natural Language Generation Systems”. Cambridge University Press.

Reiter, Ehud (2007) An Architecture for Data-to-Text Systems. Proc. of ENLG-2007.

Russell, S. and Norvig, P. (2003) “Artificial Intelligence: A Modern Approach”. Prentice-Hall.

Sripada, S., E. Reiter, J. Hunter, and J. Yu (2003) “Exploiting a parallel text-data corpus”. Corpus Linguistics 2003, pages 734–743.

Williams, S. and Ehud Reiter (2005) “Deriving content selection rules from a corpus of non-naturally occurring documents for a novel NLG application”. Corpus Linguistics 2005 workshop on Using Corpora for Natural Language Generation.