SINotas: the Evaluation of a NLG Application

Roberto P.A. de Araujo, Rafael L. de Oliveira, Eder M. de Novais,
Thiago D. Tadeu, Daniel B. Pereira, Ivandré Paraboni
School of Arts, Sciences and Humanities - University of São Paulo (USP / EACH)
Av. Arlindo Bettio, 1000 São Paulo Brazil
E-mail: {roberto.araujo, rafaellage, eder.novais, thiagoo, daniel.bastos, ivandre} @usp.br

Abstract

SINotas is a data-to-text NLG application intended to produce short textual reports on students' academic performance from a database conveying their grades, weekly attendance rates and related academic information. Although developed primarily as a testbed for Portuguese Natural Language Generation, SINotas generates reports of interest to both students keen to learn how their professors would describe their efforts, and to the professors themselves, who may benefit from an at-a-glance view of the student's performance. In a traditional machine learning approach, SINotas uses a data-text aligned corpus as training data for decision-tree induction. The current system comprises a series of classifiers that implement major Document Planning subtasks (namely, data interpretation, content selection, within- and between-sentence structuring), and a small surface realisation grammar of Brazilian Portuguese. In this paper we focus on the evaluation work of the system, applying a number of intrinsic and user-based evaluation metrics to a collection of text reports generated from real application data.

1. Introduction

Data-to-text Natural Language Generation (NLG) concerns the study and development of applications that produce text reports from non-linguistic (e.g., numeric) data (Reiter, 2007). In previous work (Novais et. al., 2009; Oliveira et. al., 2009) we have addressed the first stages of a data-to-text NLG architecture by implementing a Document Planning module based on a series of classifiers[1] trained on a data-text parallel corpus.

By attaching a surface realisation grammar to the existing system, we have presently completed a simple data-to-text NLG application that we call the SINotas system. In SINotas, grades obtained by undergraduate students and additional numeric data (such as weekly attendance rates and related information) are described as short text reports generated automatically from raw data taken from the students' academic records. Although developed primarily as a testbed for Portuguese Natural Language Generation research (cf. Novais et. al, 2009), SINotas reports turn out to be of interest to both students keen to learn how their professors would describe their efforts, and to the professors themselves, who may benefit from an at-a-glance view of the student's performance.

As a first assessment of the whole SINotas system, we carried out an evaluation work in collaboration with a group of potential users (namely, undergraduate students) who provided us with real data used as input to the system. The generated reports were subject to a number of intrinsic and user-based evaluation metrics, whose results are the main focus of this paper. For a more detailed description of the individual system components, we refer to Oliveira et. al. (2009). For issues related to the corpus used as training data for SINotas, see Novais et al. (2009).

2. Evaluation

We started the SINotas project by building a small parallel data-text NLG corpus comprising a collection of students' academic records (represented as attribute-value vectors), and corresponding text reports that describe each record individually. This data-text aligned structure- called the SINotas corpus - is described in Novais et. al., (2009).

Briefly, the SINotas corpus consists of 241 paired data-text records of students' academic performance data in five courses taught by a single professor (the domain expert) in an academic term, and their corresponding text reports, which were manually written by the same professor. The use of a single author as the main source for knowledge acquisition was required to establish meaningful mappings from raw data (e.g., students' grades) to semantics (i.e., the interpretation of the data according to a professor), as different experts will have different opinions on, e.g., what constitutes a 'good' grade in a given course.

We notice that in our work the entire data-text alignment task was performed manually, and represented explicitly in the corpus. In other words, the domain expert analysed each data record individually and wrote a suitable text report to describe it. For examples of automatic or semi-automatic data-text alignment techniques applied to NLG, see, e.g., Barzilay and Lapata (2003) and Kelly et. al. (2009).

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 accompanying report describes this data set in a simplified form as discussed below, conveying on average 5 sentences each. The following Figure 1 shows an example of one such text report (rendered in English.)

[1] For other uses of serialized classifiers in NLG (e.g., applied to the surface realisation task), see for example Smets et. al. (2003) and Marciniak and Strube (2005).

> *Your performance in the regular exam was good, and also above the average for your class.*
>
> *Your grades had experienced an increase in the middle of the term, but fell again towards the end.*
>
> *Your performance in the substitutive exam was slightly below the average for your class.*
>
> *On the other hand, the results of your practical assignment were excellent.*
>
> *Your final results were excellent and also above the average - congratulations!*

Figure 1: A sample text report.

The main simplification observed in the text reports is the limited use of discourse relations. In the SINotas corpus, there are only three kinds of within-sentence discourse markers, corresponding to *concession*, *joint* and *contrast* RST relations (Mann & Thompson, 1987).

Moreover, a sentence can be linked to another using only two possible RST relations, namely, *contrast* or *elaboration* (or none at all). While this level of simplification serves the purposes of our research project (namely, providing a ready-to-use testbed for Portuguese NLG research), we are aware that this may be insufficient from the point of view of a real-world application, as suggested by the rather sketchy example in previous Figure 1.

2.1 Document Planning Evaluation

The SINotas corpus was used as training data for a series of classifiers applied to four Document Planning subtask as described in Oliveira et. al., (2009). These are divided into (numeric) data interpretation, content selection, within-sentence structuring and between-sentences structuring.

Briefly, data interpretation computes all possible messages derivable from the application input data; next, content selection decides which of the generated messages should appear in the output text; third, within-sentence structuring aggregates sets of messages into sentences using appropriate RST relations; finally between-sentence structuring uses additional RST relations to tie the individual sentences together in a coherent discourse structure. Each of these procedures is modelled as a classification problem. For details, see Oliveira et. al., (2009).

The average results of Weka J48 decision-trees induction using 10-fold cross-validation (Witten & Frank, 2005) for the four Document Planning subtasks were deemed satisfactory as discussed in Oliveira et. al. (2009), and are summarized in the following Table 1 for illustration purposes only.

Criteria	P	R	F	Bs.
Data interpretation	0.883	0.885	0.886	0.575 [2]
Content selection	0.934	0.948	0.937	0.892 [3]
Within-sentence structuring	0.705	0.667	0.685	0.708 [4]
Between-sentences structuring	0.935	0.907	0.921	0.923 [5]

Table 1: Precision (P), Recall (R), F-measure (F) and Baseline (Bs) average results for the main Document Planning tasks (cf. Oliveira et. al., 2009).

Given that the output of a module is taken as the input to the next, individual results of each task may not provide an accurate picture of the actual error rates to be expected in the output text. Thus, we decided to extend the evaluation work by performing a simple analysis of the classification errors over the entire set of 241 documents.

We computed the number of reports conveying missing or wrong values in either content messages or discourse relations (i.e., counting both the errors stemming from data interpretation or content selection, and those that were produced in the discourse structuring stages), and classified each error either as being caused by the machine learning algorithm, or simply as an error inherited from the previous stage.

Data interpretation and content selection tasks jointly determine the content messages that appear in the output text, and were responsible for 124 classification errors distributed across 86 reports, being 53 instances of missing values and 54 misclassified content messages, plus 17 instances of messages that were not supposed to appear in the output. The errors in these two initial stages percolated to within-sentence and between-sentences structuring tasks, which are to a lesser extent responsible for additional errors (20 and 5 instances, respectively.) The following Table 2 summarizes these results, showing the number of output reports only.

Criteria	Local errors	Cumulative errors
Data interpretation / selection	86	86
Within-sentence structuring	20	106
Between-sentences structuring	5	111

Table 2: Local and cumulative classification errors (in number of output reports).

The above results suggest that despite the higher scores for each individual classification task in previous Table 1, the data interpretation and content selection classifiers are indeed responsible for the vast majority of errors made by

[2] Selecting the most frequent value for each class.

[3] Selecting all non-null input messages for realisation.

[4] Selecting the most frequent meaningful class (i.e., disregarding negative instances.)

[5] Selecting the NULL relation for all cases (i.e., not performing any document structuring.)

the system, whereas the discourse structuring stages need far less improvement. This was in our view to be expected given that converting numeric data to content messages is far more complex (and considerably more prone to ambiguity) than selecting RST relations in our relatively simple discourse structure (e.g., decisions such as linking a positive and a negative result using a 'contrast' RST relation are fairly straightforward in our simplified corpus.)

2.2 Overall System's Corpus-based Evaluation

The Document Planning module presented in Oliveira et. al. (2009) was attached to a surface realisation grammar, making a complete (although still rather simple) data-to-text NLG application that we have called the SINotas system. As a first assessment of the application as a whole, we have applied a range on intrinsic and user-based metrics as follows.

First, using the same (numeric) data that produced the SINotas corpus described in Novais et. al., (2009) as an input to the system, we re-generated all existing reports (241 instances in total) and compared them to the original (i.e., hand-written) text using a number of intrinsic evaluation metrics. In doing so, the quality of the system was evaluated as a whole, that is, all NLG subtasks from Document Planning to Surface Realisation were evaluated as a single unit.

Three metrics were used in this evaluation: BLEU (Papineni et. al., 2002), NIST (NIST, 2002) and Edit-distance scores (as in, e.g., Bangalore et. al. (2006)), in all cases using the corpus text as a gold standard. The following results were obtained.

Criteria	Score
BLEU	0.70
NIST	6.54
Edit-distance	2.46

Table 3: Surface realisation evaluation (corpus-based.)

We notice that some of the results (and most evidently, BLEU scores) are still relatively low due to the fact that our system currently does not handle certain special characters such as Portuguese accent marks and upper casing, a point that we will come back to later.

As these intrinsic evaluation metrics are not sensitive to differences in word usage (e.g., the use of synonymy for the sake of variation counts as an 'error'), and also because they do not allow us to single out the individual aspects of the system that may need improvement, we decided to complement this by performing a simple user-based evaluation of the application in which individual aspects of the system performance could be taken into account.

2.3 Overall System's User-based Evaluation

The user-based evaluation was carried out as follows. Taken previously unseen data as an input, we generated 26 reports and asked the users (i.e., students) whose results they describe to assign scores (from 0 to 5) according to five criteria.

We notice that this rather informal evaluation method is not intended to provide an accurate picture of the system functionality, but simply to gather first-hand evidence of required improvements to the system. For a thorough discussion on the intrinsic, extrinsic and user-based evaluation metrics in the evaluation of NLG systems, see for example Belz & Reiter (2006).

The following criteria were considered: humanlikeness (i.e., whether the text seemed to be written by humans), grammaticality, clarity, accuracy and content selection (i.e., whether each individual piece of text was consistent with the input data.) Additionally, we have also collected written opinions on various aspects of the text report to guide future improvements. The results are summarized in the following Table 4.

Criteria	Average	SD
Humanlikeness	1.9	1.5
Grammaticality	4.0	0.6
Clarity	4.7	0.5
Accuracy	4.3	1.0
Content Selection	4.3	0.9

Table 4: Overall system evaluation (user-based, scores ranging from 0 to 5.)

The system fared generally well according to all but one criterion (humanlikeness), in which case it was heavily penalized. The reasons for this are twofold: first, since most students knew in advance that they were participating in the evaluation of a NLG system, they may have correctly guessed that their documents were among those that were generated by the system, and may have been biased in their answers to the question 'Do you think that your report was generated by a computer system?".

Second, we are aware that our generated reports do seem 'artificial' in the sense that they comprise a series of rather unrelated statements. This was a decision made when preparing the training data (i.e., the reports in the SINotas corpus have been intentionally normalized to make the classification tasks feasible given data sparseness and other challenges) and the system is actually highly faithful to the original examples, to the point that it is virtually impossible to distinguish human from machine-generated documents. In other words, what is reflected in this evaluation is actually the lack of humanlikeness of the manually-written reports in Novais et. al. (2009).

Regarding the issue of grammaticality, we notice that the system does not generate ungrammatical sentences. Once again, grammaticality was penalized by the lack of Portuguese accent marks and upper case, which are not handled by our current surface realisation grammar. With hindsight, we should not have overlooked these pos-editing details in the system evaluation. The required improvements are now underway and we do not expect further issues regarding grammaticality.

3. Conclusion

In this paper we have described the preliminary evaluation work of an ongoing data-to-text NLG project called SINotas, a system primarily developed as a testbed for NLG research in the Portuguese language, and which is able to generate short text reports of students' academic performance from raw numeric data. SINotas was implemented as a series of classifiers trained on a small aligned data-text corpus, and it was evaluated using a number of intrinsic (i.e., corpus-based) and user-based metrics.

The overall results of the evaluation suggest that SINotas performed considerably well, with two important exceptions: data interpretation and surface realisation. Data interpretation errors were mainly due to the imprecise nature of data-to-concept mapping task, which is evident in the SINotas corpus and, as a result, reflected in the generated text reports. More work is still required to refine the behaviour of the data interpretation subtask in our system.

With respect to the surface realisation task, we notice that the text reports obtained low scores for the 'humanlikeness' criterion in the user-based evaluation. Besides our project decision to keep the training data simple (and which necessarily leads to the generation of rather unsophisticated output text), this was mainly due to the fact that the participants knew in advance that the text was computer-generated (or at least many of them thought so.) With hindsight, we should have applied a more robust user-based evaluation technique such as, e.g., the Two-panel Evaluation Methodology described in Lester and Porter, (1997).

A new round of user-based evaluation is currently underway, and as future work we intend to take these insights into account and expand the coverage of the current system, besides applying more formal extrinsic evaluation techniques when appropriate.

4. Acknowledgements

The authors acknowledge support by FAPESP, CNPq and the University of São Paulo. We are also grateful to the subjects who took part in the evaluation work for their valuable feedback, and to the anonymous reviewers of the original submission.

5. References

Bangalore, S., O. Rambow and S. Whittaker. (2000) Evaluation metrics for generation. *Proceedings of the 1st International Conference on Natural Language Generation (INLG '00)*, pp.1-8.

Barzilay, R. and M. Lapata (2003) Collective Content Selection for Concept-To-Text Generation. Proceedings of the HLT'05 conference, pp. 331–338.

Belz, A. and E. Reiter (2006) Comparing Automatic and Human Evaluation of NLG Systems. *Proceedings of EACL'06*, pp. 313-320.

Kelly, C., A. Copestake and N. Karamanis (2009) Investigating Content Selection for Language Generation using Machine Learning. *Proceedings of the 12th European Workshop on Natural Language Generation (ENLG-2009)*. Athens, Greece.

Lester, C. J. and B.W. Porter (1997) Developing and Empirically Evaluating Robust Explanation Generators: The KNIGHT Experiments. *Computational Linguistics* 23(1), pp. 65-101.

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.

Marciniak, T. and M. Strube (2005) Using an Annotated Corpus As a Knowledge Source For Language Generation". *Proceedings of the Corpus Linguistics'05 Workshop Using Corpora for NLG (UNNLG-2005)*, pp. 19-24.

NIST (2002) Automatic Evaluation of Machine Translation Quality using n-gram Co-occurrence Statistics".*www.nist.gov/speech/tests/mt/doc/ngram-study.pdf*

Novais, Eder Miranda de, Rafael Lage de Oliveira, Daniel Bastos Pereira, Thiago Dias Tadeu and Ivandré Paraboni (2009) A Testbed for Portuguese Natural Language Generation. *Proceedings of the 7th Brazilian Symposium in Information and Human Language Technology (STIL-2009)*.

Oliveira, Rafael Lage de, Eder Miranda de Novais, Roberto Paulo Andrioli de Araujo and Ivandré Paraboni (2009) A Classification-driven Approach to Document Planning. *Proceedings of the Recent Advances in Natural Language Processing Conference (RANLP-2009)* pp. 324-329.

Papineni, S., T. Roukos, W. Ward, and W. Zhu (2002) Bleu: a method for automatic evaluation of machine translation. *Proceedings of ACL-02*, pp. 311–318.

Reiter, E. (2007) An Architecture for Data-to-Text Systems. *Proceedings of ENLG-2007*, pp. 97-104.

Smets, M., M. Gamon, S. Corston-Oliver and E. Ringger (2003) French Amalgam: A machine-learned sentence realization system. Proceedings of the TALN-2003 conference.

Witten, I. H. and E. Frank (2005) *Data Mining: Practical machine learning tools and techniques*. 2nd edition, Morgan Kaufmann, San Francisco.