DaMata: A Robot-Journalist Covering the Brazilian Amazon Deforestation

André Luiz Rosa Teixeira[1], **João Gabriel Moura Campos**[2], **Rossana Cunha**[1], **Thiago Castro Ferreira**[1], **Adriana Silvina Pagano**[1] and **Fabio Gagliardi Cozman**[2]

[1]Laboratory for Experimentation in Translation, Federal University of Minas Gerais
{andrelrt,rossanacunha,thiagocf05,apagano}@ufmg.br

[2]Escola Politécnica, University of São Paulo
{joaogcampos,fgcozman}@usp.br

Abstract

This demo paper introduces *DaMata*, a robot-journalist covering deforestation in the Brazilian Amazon. The robot-journalist is based on a pipeline architecture of Natural Language Generation, which yields multilingual daily and monthly reports based on the public data provided by *DETER*, a real-time deforestation satellite monitor developed and maintained by the Brazilian National Institute for Space Research (INPE). *DaMata* automatically generates reports in Brazilian Portuguese and English and publishes them on the Twitter platform. Corpus and code are publicly available.[1]

1 Introduction

Robot-Journalism is one of the most promising Natural Language Generation (NLG) applications thanks to the high volume of structured data-streams available nowadays, which enables automated systems to report recurrent information with high-fidelity and low latency. Such automation can unburden journalists from tedious data reporting tasks, thus enabling human agents to devote efforts to more investigative coverage (Graefe, 2016).

The health of the Amazon rainforest has drawn global news coverage attention lately, when soaring deforestation raised public awareness worldwide (Escobar, 2020). Accurate information about deforested land in the Brazilian part of this territory is yielded by the country's National Institute for Space Research (INPE)[2] through several satellites and other monitoring systems. One of the most effective is DETER, a real-time deforestation warning surveillance system (Diniz et al., 2015).

Despite being publicly available[3], DETER's deforestation data output is in graphical and numerical style, which renders it less accessible to general audiences and demands coverage by human journalists specialized in the domain. To address this issue, we introduce *DaMata*, a robot-journalist based on an NLG pipeline architecture (Reiter and Dale, 2000; Gatt and Krahmer, 2018), which generates daily and monthly reports about deforestation in the Brazilian Amazon, both in Brazilian Portuguese[4] and English[5], and publishes them on Twitter.

2 System Overview

Unlike novel end-to-end systems, which may hallucinate content (Moryossef et al., 2019) and may be problematic for sensitive domains, *DaMata* follows the pipeline architecture proposed by Ferreira et al. (2019), which converts non-linguistic data into text in 6 steps: Content Selection, Discourse Ordering, Text Structuring, Lexicalization, Referring Expression Generation and Textual Realization. We have added a seventh step to *DaMata* – publication – responsible for sharing the generated news on Twitter. This kind of architecture, depicted in Figure 1, allows for trustworthy output as well as easier access and maintenance of the sub-modules and easier multi-domain and multi-language applications.

The grammar used by the model was built by first running the content selection step in previous data, generating 14 non-linguistic monthly reports and 25 daily ones. These non-linguistic reports were then manually verbalized and the input and output representations for each pipeline module were manually annotated. This process resulted in a list of possible discourse orders, text structures, lexicalizations, and referring expressions. When deployed, each module draws on the selected combination of templates using rule-based approaches.

The next sections describe each module.

[1] https://github.com/BotsDoBem/DEMO_INPE_COVID

[2] http://www.inpe.br/

[3] http://terrabrasilis.dpi.inpe.br

[4] https://twitter.com/DaMataReporter

[5] https://twitter.com/DaMataNews

Proceedings of The 13th International Conference on Natural Language Generation, pages 103–106,
Dublin, Ireland, 15-18 December, 2020.

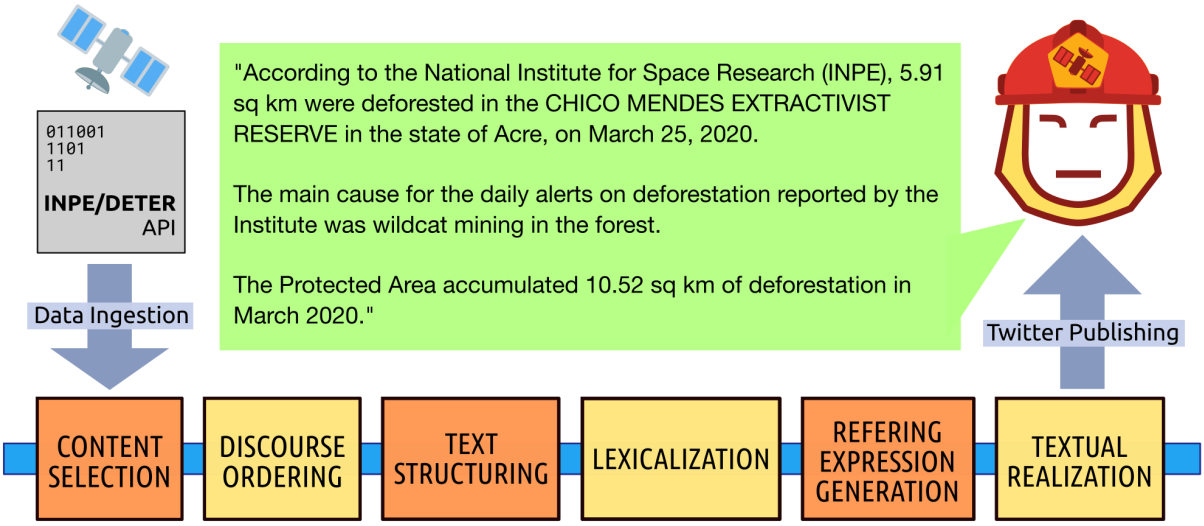

Figure 1: Robot-Journalist Pipeline Architecture and a human-readable output sample.

Content Selection This module is responsible for selecting relevant messages to be verbalized in the reports. Satellite data are gathered, processed, and interpreted by INPE, which provides this information through DETER API[6]. These data come in a structured format, a JSON file comprising pre-defined features, e.g., deforestation area, cause, date, city, state, etc., which are mapped to *DaMata's* database. The information is structured as *intent-attribute-value* messages:

```
DAILY_ALERT(area="5.91",city="None",uc=
"CHICO_MENDES_EXTRACTIVIST_RESERVE",day=
"25",month="3",year="2019",state="Acre",
location="amazon",daily_accumulation="1")
```

We adopt a rule-based approach to content selection. The system checks for monthly and daily updates on deforestation. For a monthly report, *DaMata* always verbalizes news on the amount of total deforestation in the Brazilian Amazon and the main cause for deforestation in the corresponding time period. Next, the system verbalizes month and annual variation in case of an increase in these values compared to the previous period. Finally, it reports news on the state, city, and nature protected area with the highest deforestation rates in the respective month, if the corresponding deforested area was greater than the established threshold (e.g., the sum of mean value and standard deviation of the deforestation areas time series).

For a daily report, our robot-journalist selects data on the amount of deforested area and the main cause of deforestation in a given city or nature protected area as well as the total area hitherto deforested in the respective month for the given region. The following is an example of the content selection module output:

```
TOTAL_DEFORESTATION(area,location,month,
year);
DAILY_ALERT(area,day,month,year,
daily_accumulation,location,state,city);
CAUSE(area,cause,location,month, year);
```

Discourse Ordering According to Jurafsky and Martin (2019), ordering events in discourse to construe a logical timeline enhances clarity and reader comprehension. Based on the list of possible intent orderings collected from our corpus (for instance 8 ordering templates for daily reports and 10 for monthly reports), we arrange our discourse messages by a rule-based approach. By using discourse ordering templates based on human generated texts, we ensure conveying the relative importance of each piece of information, as it is usually done in journalistic texts. For example, for the messages selected for a daily report – daily deforested area, main cause of deforestation and total deforestation in the month, a likely outcome order would be:

```
DAILY_ALERT → CAUSE →
TOTAL_DEFORESTATION
```

Text Structuring This step is responsible for structuring the information in sentences and paragraphs (Gatt and Krahmer, 2018; Ferreira et al., 2019), bearing in mind the character limitation on Twitter (280 characters per tweet) and how intents should be ordered. As in the previous phase,

[6] http://terrabrasilis.dpi.inpe.br/homologation/file-delivery/download/deter-amz/daily

our robot-journalist draws on the annotated corpus structures and also decides which structuring template to use by a rule-based approach. For our daily report example, this step would return:
`<P> <S>DAILY_ALERT</S> <S>CAUSE</S> </P>`
`<P> <S>TOTAL_DEFORESTATION</S> </P>.`

Lexicalization Once the order and structure of the intents are set, *DaMata* chooses a lexicalization template for each structured sentence. These templates are chosen from 55 sentence templates for monthly reports and 46 templates for daily reports assuring that the lexicalization step has plenty of options to choose from, thus, rendering more variety in text outputs.

The templates provide for gender and number inflection. Verbs and nouns reflect entity gender and number attribute values (especially in Brazilian Portuguese), e.g., *"Altamira accumulates 1 day with alerts"* vs. *"Altamira accumulates 8 days with alerts"*. A fill-template would not cater for number and gender inflection, and would also decrease variety in the final output.

To further increase variety and to ensure audience engagement, this module randomly picks one template from this pool of options. For example:
`DAILY_ALERT ⇒ According to INSTITUTE, AREA sq km VP[...] deforest in UC in STATE, on MONTH DAY, YEAR.`

Referring Expression Generation At this step in the pipeline, the system tracks entity tags in the lexicalized template and maps the appropriate referring expressions, replacing entities in context accordingly. In this module, *DaMata* uses a list of possible expressions for each entity, obeying some choice constraints. For a first reference to an entity in the text, a full description is used (e.g., `INSTITUTE ⇒ the National Institute for Space Research (INPE)`), whereas for subsequent references a random referring expression to the entity is chosen (e.g., `INSTITUTE ⇒ the National Institute for Space Research (INPE)`; `the Institute`; `INPE`; `it`; etc.).

Textual Realization The final step in our pipeline sequence performs the remaining adjustments to transform the abstract machine intermediate representations into human-readable text. Nominal and verbal inflections, contractions and detokenization are performed according to each grammar. The following is an example of the intermediate representation and the output for a verbal inflection output during this process:
`VP [aspect=simple, tense=past, voice=passive, person=3rd, number=plural] deforest → were deforested`

The final product is shown in the robot dialog box in Figure 1.

3 Conclusions

This study introduces *DaMata*, a robot-journalist based on a pipeline architecture of NLG, which generates and publishes multilingual daily and monthly news reports about deforestation in Brazilian Amazon on Twitter. By leveraging structured datastreams available for the domain, *DaMata* can provide low latency human-readable information to wider audiences. Moreover, due to its rule-base nature, the pipeline framework ensures high-fidelity data publishing and the random choice at template level enhances lexical variety and audience engagement.

For future work, we plan to publish satellite images indicating the deforested areas alongside the generated texts. A further interesting feature to be implemented is a chat-bot to the Twitter account, as we have noticed that Twitter recipients tend to react to the posts and seek to engage in conversations with the robot-journalists for further information. We also intend to add more languages for our robot to output, with a view to raising broader audience awareness about deforestation. Our robot-journalism proposal can be used to report on additional domains, particularly pertaining to sensitive issues that could benefit from ampler dissemination. For further improvement of our system's performance, we plan to add data-driven techniques to the existing pipeline modules.

Acknowledgments

Research funded by the Coordination for the Improvement of Higher Education Personnel (CAPES) under grants 88887.488096/2020-00, 88882.349188/2019-01 and 88887.508597/2020-00; the National Council for Scientific and Technological Development (CNPq), grant 312180/2018-7; the São Paulo Research Foundation (FAPESP), grant 2019/07665-4; and the State Funding Agency of Minas Gerais (FAPEMIG), grant APQ-01129-17.

References

Cesar Guerreiro Diniz, Arleson Antonio de Almeida Souza, Diogo Corrêa Santos, Mirian Correa Dias, Nelton Cavalcante da Luz, Douglas Rafael Vidal de Moraes, Janaina Sant'Ana Maia, Alessandra Rodrigues Gomes, Igor da Silva Narvaes, Dalton M Valeriano, et al. 2015. Deter-b: The new amazon near real-time deforestation detection system. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(7):3619–3628.

Herton Escobar. 2020. Deforestation in the brazilian amazon is still rising sharply. *Science*, 369(6504):613–613.

Thiago Castro Ferreira, Chris van der Lee, Emiel van Miltenburg, and Emiel Krahmer. 2019. Neural data-to-text generation: A comparison between pipeline and end-to-end architectures. In *EMNLP/IJCNLP*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Andreas Graefe. 2016. Guide to Automated Journalism. Technical report, Tow Center for Digital Journalism, Columbia University, New York.

Dan Jurafsky and James H Martin. 2019. *Speech and language processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3rd edition*. Pearson London.

Amit Moryossef, Yoav Goldberg, and Ido Dagan. 2019. Step-by-step: Separating planning from realization in neural data-to-text generation. In *Proceedings of NAACL*, Minneapolis, Minnesota.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge, UK: Cambridge University Press.