Enriching the WebNLG corpus

Thiago Castro Ferreira[1] **Diego Moussallem**[2,3] **Sander Wubben**[1] **Emiel Krahmer**[1]
[1]Tilburg center for Cognition and Communication (TiCC), Tilburg University, The Netherlands
[2]AKSW Research Group, University of Leipzig, Germany
[3]Data Science Group, University of Paderborn, Germany
{tcastrof,s.wubben,e.j.krahmer}@tilburguniversity.edu
moussallem@informatik.uni-leipzig.de

Abstract

This paper describes the enrichment of WebNLG corpus (Gardent et al., 2017a,b), with the aim to further extend its usefulness as a resource for evaluating common NLG tasks, including Discourse Ordering, Lexicalization and Referring Expression Generation. We also produce a silver-standard German translation of the corpus to enable the exploitation of NLG approaches to other languages than English. The enriched corpus is publicly available[1].

1 Introduction

Natural Language Generation (NLG) is the process of automatically converting non-linguistic data into a linguistic output format (Reiter and Dale, 2000; Gatt and Krahmer, 2018). Recently, the field has seen an increase in the number of available focused data resources as E2E (Novikova et al., 2017), ROTOWIRE (Wiseman et al., 2017) and WebNLG (Gardent et al., 2017a,b) corpora.

Although theses recent releases are highly valuable resources for the NLG community in general, all of them were designed to work with end-to-end NLG models. Hence, they consist of a collection of parallel raw representations and their corresponding textual realizations. No intermediate representations are available so researchers can straight-forwardly use them to develop or evaluate popular tasks in NLG pipelines (Reiter and Dale, 2000), such as Discourse Ordering, Lexicalization, Aggregation, Referring Expression Generation, among others. Moreover, these new corpora, like many other resources in Computational Linguistics more in general, are only available in English, limiting the development of NLG-applications to other languages, which is currently an emerging theme in NLG research community – see, for instance, the increased availability of SimpleNLG tools to languages other than English (Mazzei et al., 2016; Bollmann, 2011; Vaudry and Lapalme, 2013; Ramos-Soto et al., 2017) and the recent Multilingual Surface Realization task (Mille et al., 2018).

This paper describes how we addressed the aforementioned issues by enriching the WebNLG corpus with intermediate representations and exploiting the possibilities of automatically translating the corpus to a second language (German). To this end, we first manually replaced all referring expressions in the WebNLG texts with general tags in a process called *Delexicalization*. The original texts and the delexicalized templates were then translated to the German language using an existing state-of-art English-German Neural Machine Translation (NMT) system (Sennrich et al., 2017), providing a silver-standard version of the corpus in another language. Finally, by automatically processing the original texts and the delexicalized templates for both English and (translated) German versions of the corpus, we obtained a collection of gold-standard referring expressions and discourse orderings. In sum, the main contributions of this study (all publicly available) are:

- A silver-standard version of the WebNLG corpus in German.
- A collection of 16,661 delexicalized templates in English and 16,292 in (silver-standard) German.
- A collection of 86,345 referring expressions to 1,771 Wikipedia entities and constants in English and (silver-standard) German.

[1] https://github.com/ThiagoCF05/webnlg

Proceedings of The 11th International Natural Language Generation Conference, pages 171–176,
Tilburg, The Netherlands, November 5-8, 2018.

- Discourse ordering information of 20,370 instances of the WebNLG corpus.

The paper is organized as follows: Section 2 briefly describes the WebNLG corpus, Section 3 depicts our delexicalization procedure in detail (briefly introduced in Castro Ferreira et al. 2018), Section 4 explains the process of translating the corpus texts to German, Section 5 introduces the intermediate representations automatically extracted from the corpus in its original and delexicalized forms, and, finally, Section 6 discusses the contributions and prospects of future work.

2 WebNLG corpus

The WebNLG corpus (Gardent et al., 2017a) is a parallel dataset initially released for the eponymous NLG challenge, where participants had to automatically convert non-linguistic data from the Semantic Web into a textual format (Gardent et al., 2017b). The source side of the corpus are sets of *Resource Description Framework* (RDF) triples, in which, each one of them is formed by a Subject, Predicate and Object. The Subject and Object are constants or Wikipedia entities, whereas predicates represent a binary relation between these two elements in the triple. The target side contains English texts, obtained by *crowdsourcing*, which describe the source triples.

The corpus consists of 25,298 texts describing 9,674 sets of up to 7 RDF triples (an average of 2.62 texts per set) in 15 domains: Astronaut, University, Monument, Building, Comics Character, Food, Airport, Sports Team, Written Work, City, Athlete, Artist, Means of Transportation, Celestial Body and Politician (last 5 were available only in the test set). Figure 1 shows an example of a set of 5 RDF triples and its corresponding English text.

3 Delexicalization

To account for data sparsity and unseen entities, many "end-to-end" NLG models work by first generating a *delexicalized* template, where references are represented by general tags (Konstas et al., 2017; Castro Ferreira et al., 2017). The referring expressions are only surface realized once the template is generated. Motivated by these studies, we manually delexicalized the training and development parts of the WebNLG corpus, generating gold-standard templates. The test part of the corpus was not included in this study, since only its source RDF triples were publicly available at the time of writing.

We started the delexicalization process by automatically mapping each entity in the source representation to a general tag. All entities that appear on the left and right side of the triples were mapped to AGENTs and PATIENTs, respectively. Entities which appear on both sides in the relations of a set were represented as BRIDGEs. To distinguish different AGENTs, PATIENTs and BRIDGEs in a set, an ID was assigned to each entity of each kind (PATIENT-1, PATIENT-2, etc.).

Once all entities were mapped to different general tags in the text, the first two authors of this study manually replaced the referring expressions in the original target texts by their respective tags. Each annotator delexicalized half of the texts, and the few difficult cases were resolved in discussions with the co-authors. Figure 2 shows the entity mapping and the delexicalized template for the example in Figure 1.

In total, we delexicalized 20,370 different texts which describe 7,812 distinct sets of RDF triples, resulting in 16,661 distinct templates in English. Together with the original texts, we translated the delexicalized templates to German, and extracted a collection of referring expressions and discourse ordering information of the corpus as explained in the following sections.

4 Translation

We translated the original texts and the delexicalized templates by relying on the University of Edinburgh's Neural MT System for WMT17 (Sennrich et al., 2017; Bojar et al., 2017a). Not only are the training models publicly available[2], but this system is state-of-the-art in translating English-to-German at the time of writing[3], which guarantees we obtain arguably the best silver-standard resource currently feasible.

The University of Edinburgh system was modeled in a deep encoder attention-decoder architecture. Its translation model was trained on back-translated monolingual data (Sennrich et al., 2016a) in order to augment the training data. To have an open vocabulary, the rare words, which pose a well-known problem in NMT systems,

[2] http://data.statmt.org/wmt17_systems

[3] http://matrix.statmt.org/matrix/systems_list/1869

Subject	Predicate	Object
Appleton_International_Airport	location	Greenville,_Wisconsin
Greenville,_Wisconsin	isPartOf	Ellington,_Wisconsin
Greenville,_Wisconsin	isPartOf	Menasha_(town),_Wisconsin
Greenville,_Wisconsin	country	United_States
Appleton_International_Airport	cityServed	Appleton,_Wisconsin

↓

The Appleton International Airport is located in Greenville, Wisconsin, United States and serves the city of Appleton, Wisconsin. Greenville is part of the town of Menasha and Ellington, Wisconsin.

Figure 1: Example of a set of triples (top) and corresponding text (bottom).

Tag	Entity
AGENT-1	Appleton_International_Airport
BRIDGE-1	Greenville,_Wisconsin
PATIENT-1	United_States
PATIENT-2	Appleton,_Wisconsin
PATIENT-3	Menasha_(town),_Wisconsin
PATIENT-4	Ellington,_Wisconsin

↓

AGENT-1 is located in **BRIDGE-1** , **PATIENT-1** and serves the city of **PATIENT-2** . **BRIDGE-1** is part of **PATIENT-3** and **PATIENT-4** .

Figure 2: Mapping between tags and entities for the related delexicalized/wikified templates.

were segmented into sub-word units using Byte Pair Encoding (BPE) (Sennrich et al., 2016b).

To translate a sentence, the University of Edinburgh submission trained 4 left-to-right and 4 right-to-left models. The left-to-right models were ensembled to produce the 50 most likely translation hypotheses while the right-to-left models were then used to re-rank the outcomes from the left models. The process resulted in 20,370 texts and 16,292 delexicalized templates in German.

5 Automatic extraction process

In both English and (translated) German versions of the corpus, we used the original texts and the delexicalized templates to automatically extract a collection of referring expressions and discourse ordering information.

5.1 Referring expression collection

For both English and (translated) German versions, we automatically extracted a collection of referring expressions by tokenizing the original texts and delexicalized templates, and then finding the non-overlapping items. For instance, by processing the text in Figure 1 and its delexicalized template in Figure 2, we extracted referring expressions, for instance, "The Appleton International Airport" to ⟨ *AGENT-1, Appleton_International_Airport* ⟩, "Greenville, Wisconsin" and "Greenville" to ⟨ *BRIDGE-1, Greenville,_Wisconsin* ⟩, "the town of Menasha" to ⟨ *PATIENT-3, Menasha_(town),_Wisconsin* ⟩. In total, we obtained 86,345 referring expressions to 1,771 Wikipedia entities and constants, in which 72.6% (62,689) are proper names, 4.9% (4,230) pronouns, 22.13% (19,108) descriptions and 0.4% (318) demonstrative referring expressions.

5.2 Discourse Ordering

As depicted in Figure 1, each instance of the original WebNLG corpus consists of a set of triples and its respective text. However, in many cases, the order in which the triples are introduced in the set is not the same in which they are realized in the text. For instance, in Figure 1, the triple ⟨ *Appleton_International_Airport*, **cityServed**, *Appleton,_Wisconsin* ⟩ is the third argument expressed on the text, while it is represented as the 5th (last) one in the input set of triples. This is just a singular example of others that even exist in this instance. In order to solve the problem, we noticed that we could extract the order of the arguments in the text by looking into the order of the general tags in the delexicalized template, as Algorithm 1 shows.

The algorithm iterates over the words in a template (lines 4-16). If a word is a general tag (line

Algorithm 1 Discourse ordering pseudo-code

```
1: function ORDER(tripleset, template)
2:     orderedset ← ∅
3:     prevTags ← ∅
4:     for all word ∈ template do
5:         if ISTAG(word) then
6:             for all prevTag ∈ prevTags do
7:                 triples ← tripleset[prevTag, word]
8:                 if |triples| > 0 then
9:                     triple ← triples[0]
10:                    orderedset ← orderedset ∪ triple
11:                    tripleset ← tripleset \ orderedset
12:                end if
13:            end for
14:            prevTags ← prevTags ∪ word
15:        end if
16:    end for
17:    return orderedset
18: end function
```

5), the algorithm looks for a remaining instance on the triple set which relates the visited tag with a previous visited one (line 7). If a triple is found (line 8), this one is added on the ordering list (line 10) and removed from the input set (line 11).

6 Discussion

This study introduced an enriched version of the WebNLG corpus, easily usable on the evaluation of popular tasks of pipeline NLG models as Discourse Ordering, Lexicalization, Aggregation and Referring Expression Generation. Moreover, a silver-standard version of the corpus in German is provided, hopefully making it more useful for the exploration of NLG in other languages or for the study of Multilingual Surface Realization (Mille et al., 2018). We discuss below the main aspects of our results.

Delexicalization This process was applied to obtain 16,661 English templates and, after the translation process, 16,292 German templates. These representations can be used in the development of template-based NLG systems or Lexicalization models.

Automatic extraction process Using original texts and templates, we extracted important intermediate resources from the corpus, as a collection of referring expressions and discourse ordering information for English and (silver-standard) German. The former resource can be used to evaluate referring expression or wikification models, whereas the second may be a good resource for discourse ordering, content planning, and also for the aggregation task when combined with sentence tokenization information.

Translation While analyzing the translations from English to German, we could perceive that the NMT system did not face any big problem for translating the delexicalized templates. The main challenge was faced with transliterations and coreferences in the texts. The genitive case is an example, as in "Elliot See 's Besatzung war ein Testpilot.", where the apostrophe ('s) is placed wrongly. The same happens to the sentence "Bill Oddie Tochter ist Kate Hardie", where the name "Oddie" should have had the "s" in the end of this German sentence. In terms of transliterations, the preposition "von" played a key role in the challenge, as in the case of the reference "University of Texas", wrongly transliterated to "Universität von Texas" instead of the correct form "Universität Texas". These problems are well-known in WMT English-German tasks and still take a place even using the best NMT model (Koehn, 2009; Bojar et al., 2017b).

Conclusion This study aimed to enrich the WebNLG corpus, facilitating its use in popular tasks of pipeline NLG models as well as in other languages than English. In future work, we envision translating the corpus for other morphologically rich languages, as Brazilian Portuguese (Moussallem et al., 2018). Furthermore, we intend to experiment and come up with good automatic methods to improve the aforementioned challenges and generate useful silver-standard resources for NLG.

Acknowledgments

This work is part of the research program "Discussion Thread Summarization for Mobile Devices" which is financed by the Netherlands Organization for Scientific Research (NWO). It has also been supported by the National Council of Scientific and Technological Development from Brazil (CNPq) under the grants 203065/2014-0 and 206971/2014-1.

References

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017a. Findings of the 2017 conference on

machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214. Association for Computational Linguistics.

Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, et al. 2017b. Findings of the 2017 conference on machine translation (wmt17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214.

Marcel Bollmann. 2011. Adapting simplenlg to german. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 133–138. Association for Computational Linguistics.

Thiago Castro Ferreira, Iacer Calixto, Sander Wubben, and Emiel Krahmer. 2017. Linguistic realisation as machine translation: Comparing different MT models for AMR-to-text generation. In *Proceedings of the 10th International Conference on Natural Language Generation*, INLG'17, pages 1–10, Santiago de Compostela, Spain. Association for Computational Linguistics.

Thiago Castro Ferreira, Diego Moussallem, Ákos Kádár, Sander Wubben, and Emiel Krahmer. 2018. Neuralreg: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1959–1969. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017a. Creating training corpora for NLG micro-planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL'17, pages 179–188, Vancouver, Canada. Association for Computational Linguistics.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017b. The WebNLG challenge: Generating text from RDF data. In *Proceedings of the 10th International Conference on Natural Language Generation*, INLG'17, pages 124–133, Santiago de Compostela, Spain. Association for Computational Linguistics.

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.

Philipp Koehn. 2009. *Statistical machine translation*. Cambridge University Press.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL'17, pages 146–157, Vancouver, Canada. Association for Computational Linguistics.

Alessandro Mazzei, Cristina Battaglino, and Cristina Bosco. 2016. Simplenlg-it: adapting simplenlg to italian. In *INLG*, pages 184–192.

Simon Mille, Anja Belz, Bernd Bohnet, Yvette Graham, Emily Pitler, and Leo Wanner. 2018. The First Multilingual Surface Realisation Shared Task (SR'18): Overview and Evaluation Results. In *Proceedings of the 1st Workshop on Multilingual Surface Realisation (MSR), 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–10, Melbourne, Australia.

Diego Moussallem, Thiago Ferreira, Marcos Zampieri, Maria Cláudia Cavalcanti, Geraldo Xexéo, Mariana Neves, and Axel-Cyrille Ngonga Ngomo. 2018. Rdf2pt: Generating brazilian portuguese texts from rdf data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The e2e dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206, Saarbrucken, Germany. Association for Computational Linguistics.

A. Ramos-Soto, J. Janeiro-Gallardo, and Alberto Bugarín. 2017. Adapting SimpleNLG to spanish. In *10th International Conference on Natural Language Generation*.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge University Press, New York, NY, USA.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The university of edinburgh's neural mt systems for wmt17. In *Proceedings of the Second Conference on Machine Translation*, pages 389–399. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725. Association for Computational Linguistics.

Pierre-Luc Vaudry and Guy Lapalme. 2013. Adapting simplenlg for bilingual english-french realisation. In *ENLG*, pages 183–187.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.