Generating Text from Anonymised Structures

Émilie Colin
Université de Lorraine / LORIA
Nancy, France
emilie.colin@loria.fr

Claire Gardent
CNRS / LORIA
Nancy, France
claire.gardent@loria.fr

Abstract

Surface realisation maps a meaning representation (MR) to a text, usually a single sentence. In this paper, we introduce a new parallel dataset of deep meaning representations and French sentences and we present a novel method for MR-to-text generation which seeks to generalise by abstracting away from lexical content. Most current work on natural language generation focuses on generating text that matches a reference using BLEU as evaluation criteria. In this paper, we additionally consider the model's ability to reintroduce the function words that are absent from the deep input meaning representations. We show that our approach increases both BLEU score and the scores used to assess function words generation.

1 Introduction

Surface realisation (SR), the ability to generate text from meaning representations (MR), is a key component of data-to-text generation. In this paper, we focus on surface realisation for French. We make two contributions.

First, we present a method for automatically creating a parallel dataset of sentences and their meaning representations[1].

Second, we propose a novel surface realisation approach which differs from previous work in that it relies on an extensive anonymization of the data. The underlying intuition behind our approach is that abstracting away from lexical content reduces data sparsity which in turn, should facilitate the learning of linguistic structure. We show that extensive anonymization indeed improves performance. To further assess the degree to which our model learns linguistic structure, we provide an analysis of the extent to which it handles the reintroduction in the generated sentence of the function words that are absent from the input.

2 Related Work

SR Corpora Various datasets have been introduced to support the learning of surface realisers.

The 2017 AMR SemEval generation shared task (May and Priyadarshi, 2017) provides a parallel corpus where the input semantic representations are AMRs (Abstract Meaning Representation, (Banarescu et al., 2013)) and the task is to generate a sentence verbalising that AMR.

Mille et al. (2018) derived multilingual MR-to-Text datasets from the UD (Universal Dependencies) treebanks[2] creating two types of input, shallow and deep. In the shallow input, the nodes of the UD dependency tree are scrambled to remove word order information and words are replaced by their lemmas. The generation task consists in ordering and inflecting the lemmas decorating the input tree. The deep input is closer to an applicative context. It abstracts away from the surface form by removing additional information from the UD tree and replacing syntactic edge labels with PropBank/NomBank labels.

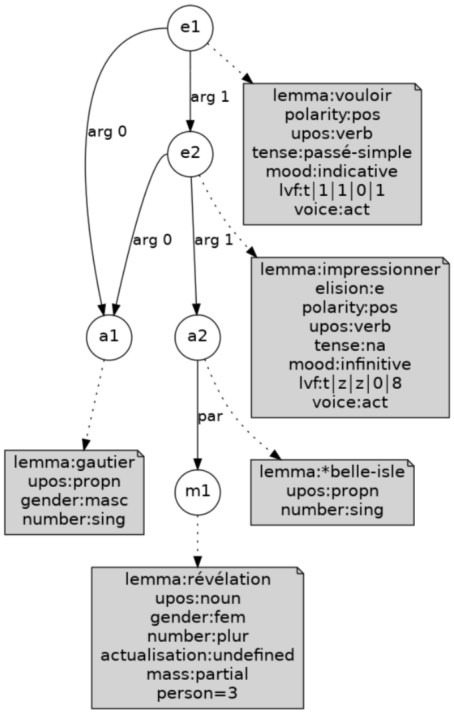
Finally, (Novikova et al., 2017) introduce a dataset where the input MRs are dialog moves.

All these datasets were expensive to build as they require extensive human intervention. The AMR datasets were built by manually annotating sentences with AMRs, the SR datasets were derived from hand-annotated treebanks and (Novikova et al., 2017)'s dialog moves were associated with text using crowdsourcing. Moreover, except for the SR task, these datasets all focus on English. In short, we depart from previous work in that we introduce a new dataset for French which is automatically derived from text using parsers.

[1] The corpus is available by simple request to the authors.

[2] http://universaldependencies.org

Proceedings of The 12th International Conference on Natural Language Generation, pages 112–117,
Tokyo, Japan, 28 Oct - 1 Nov, 2019.



(b_1) Lemma linearization:

```
( vouloir|pos|verb|ps|... arg_0
      gautier|propn|masc|sg )
( vouloir|pos|verb|ps|... arg_1
      impressionner|e|pos|verb|... )
( impressionner|e|pos|verb|... arg_0
      gautier|propn|masc|sg )
( impressionner|e|pos|verb|... arg_1
      *belle-isle|propn|sg )
par ( *belle-isle|propn|sg <
      révélation|noun|fem|plur|... )
```

(b_2) Anonymized linearization:

```
( e1|pos|verb|ps|... arg_0
      a1|propn|masc|sg )
( e1 arg_1|pos|verb|ps|[...]
      e2|e|pos|verb|... )
( e2|e|pos|verb|... arg_0
      a1|propn|masc|sg )
( e2|e|pos|verb|... arg_1
      a2|propn|sg )
par ( a2|propn|sg <
      m1|noun|fem|plur|... )
```

Figure 1: Example Input Meaning Representation and Linearizations for the sentence *Gautier voulut impressionner Belles-Isle par des révélations (Gautier wanted to impress Belles-Isle by some revelations)*

SR Models. Using the parallel MR-to-text corpora just described, various SR models have been proposed. We focus here on SR from deep meaning representations (AMRs and SR'18 deep track MRs) as this is closest to our proposal.

Early work on MR-to-text generation linearise the input graph and use various statistical methods to generate text (Flanigan et al., 2016; Song et al., 2017; Pourdamghani et al., 2016; Bohnet et al., 2010). Similarly, early neural approaches linearise the input graph and use a sequence-to-sequence (S2S) model. Konstas et al. (2017) achieve strong results on the AMR-to-text task by using data expansion and anonymising data entities while Cao and Clark (2019) additionally leverages syntactic information to improve performance. On the deep SR data, (Elder and Hokamp, 2018) uses data expansion and a factored S2S model.

Graph-to-sequence models have also been proposed using various graph encoders and testing on different datasets Marcheggiani and Perez-Beltrachini (2018); Song et al. (2018); Beck et al. (2018); Koncel-Kedziorski et al. (2019); Veličković et al. (2018).

Our approach is closest to the S2S model used by Elder and Hokamp (2018) in that it uses a factored S2S model to create rich node embeddings capturing the structure of the graph. We differ from that work in that we use full anonymization on input and output data.

3 Data

Instead of relying on hand annotations, we automatically create a silver corpus using syntactic parsing and post-processing. Creating the dataset consists of two main steps. First, we compare the output of three syntactic dependency parsers and keep only those sentences for which there is a strong consensus. Second, we derive a meaning representation from the syntactic parse.

The three parsers used are Grew[3] (Guillaume et al., 2012), Talismane[4] and the Stanford Dependency Parser[5]. As the parsers have different tokenization strategies (for instance, *"entre autres"* is treated by Talismane and Stanford as two tokens but analysed as one token *entre_autres* by Grew), the largest tokens (e.g., *entre_autres*) are used as basis for the alignment. For POS tags and dependency relations, the alignment is neither one-to-one nor one-to-many. We only keep those mappings whose frequency is above 95% for POS tags and 94% for dependency relations. Grew is used as a reference and a POS tag/Dependency relation from Talismane and the Stanford parser is judged compatible if it matches one of these mappings. We only keep those sentences for which a mapping could be found for all three parsers (for both POS tags and dependency relations) and which contain less than 70 tokens[6].

We then derive deep meaning representations from the parse tree by mapping grammatical functions (subject, etc) to semantic ones (arg0, arg1, etc.), removing function words (determiners, auxiliaries, relative pronouns, complementizers, non subcategorised prepositions), lemmatizing word forms and keeping only those features which cannot be learned e.g., the number and gender of a noun and the tense of a verb. Since determiners and auxiliary are removed from the meaning representations their number must be learned by the model or more generally, agreement constraints (between verb and subject and between noun, determiners and adjectives) must be learned. The mapping from syntactic to semantic relations makes use of a verb lexicon[7] which permits identifying which phrases are arguments (rather than modifiers). The arg0 relation is assigned to the subject of verbs in the active voice, arg1 the object and arg2 is used for prepositional arguments.

Figure 1 shows an example meaning representation.

We download 2,835 french books (3,806,889 sentences) from the Gutenberg website[8]. We filter out sentences that do not contain at least one noun and a verb or which contain incorrect bracketing or foreign (non-French) material. This yields a total of 1,700,000 sentences. We then apply the alignment procedure described above and filter out sentences with more than 70 tokens reducing the dataset further to a total of roughly 500K sentences[9]

[3]Version 0.48.0 http://grew.fr/

[4]Version 5.1.2, http://redac.univ-tlse2.fr/applications/talismane.html

[5]Version 2018-02-27, https://nlp.stanford.edu/software/lex-parser.shtml

[6]An example alignment is shown in the supplementary material.

[7]We use the LVF ("Les verbes français", the French verbs) by Jean Dubois and Françoise Dubois-Charlier, version LVF+1, http://rali.iro.umontreal.ca/rali/?q=fr/lvf.

[8]http://www.gutenberg.org/

[9]More descriptive statistics for the created corpus are given in the supplementary material.

4 Model

Model	BLEU-4	C-R	FW-F1
BL	64.35	0.59	0.906
Dual (Test:Non Anon.)	66.55	0.62	0.910
Anonymised	66.87	0.87	0.933
Dual (Test:Anon.)	68.31	0.87	0.937

Table 1: Results

Our approach extends a standard S2S model with full anonymization and factored token embeddings which capture the linguistic and structural information associated with each node in the input graph.

Full Anonymization. Anonymisation (also often dubbed, delexicalisation) has frequently been used in neural NLG to help handle unknown or rare words (Wen et al., 2015; Dušek and Jurcicek, 2015; Chen et al., 2018). Rare items are replaced by placeholders both in the input data or MR and in the output sentence or text. Models are trained on the anonymized data. Finally, a post-processing step ensures that the generated text is relexicalised using the placeholders original value. In these approaches, anonymization is restricted to rare items (named entities) and replaces them with a simple identifier. In contrast, we apply anonymization to all lemmatized content words, adverbs excepted. As we derive the input MR from the sentence parse tree (cf. Section 3), we keep track of which word form in the target sentence matches which lemma in the input MR and use this at post-processing time to relexicalise the anonymous output structure.

Factored Sequence-to-Sequence model. We use OpenNMT factored S2S model with attention. Each input token is represented by the concatenation of 18 embeddings whereby each embedding represents a distinct feature type (part-of-speech, gender, number, tense etc) (Alexandrescu and Kirchhoff, 2006). We focus on word ordering and use a table lookup to match output lemmas to word forms.

5 Experimental Settings

Evaluation Metrics. Following common practice in NLG, we evaluate our model using BLEU-4[10]. To further assess the ability of the model to learn linguistic structure, we also evaluate recall on content words (C-R) and F1 on function words (FW-F1). We define content words to be the lemmas present in the input meaning representation. Function words are the words in the (lemmatized) output sentence that are not content words.

Models. The baseline is a factored S2S model with attention trained on the original data (no anonymization). We compare this baseline with the same model trained on anonymised data (Anonymised) and with two models trained on a corpus consisting of both the original and the anonymised data. The intuition behind used both anonymised and non anonymised data for training is to see whether the combination of both sources of information can help. The first model (Dual Lexicalised) is tested on lexicalised data (does the adjunction of anonymised training data help improve generation from non anonymised meaning representations?) while the second (Dual Anonymised) is tested on anonymised data (does the adjunction of non anonymised training data help improve generation from anonymised meaning representations?). We did never anonymised adverbs, having them in outputs, waiting them in outputs.

Results. The results are shown in Table 1. The BLEU scores show that training on anonymised structures yields better results (+2.52 BLEU). Using both lexicalised and anonymised data for training further improves results which is not surprising given that the size of the training data has doubled. Interestingly, the delta is slightly better when testing on anonymised data (+3.96 vs. +2.2) which suggests that adding lexical information to the training data is more beneficial to anonymised generation than vice versa.

A similar trend can be observed concerning the handling of function words although the increase is much less.

Qualitative Analysis. On the whole test corpus (49,026 sentences), an automatic analysis of the results shows that (i) 34.76% of the generated sentences are the same as the reference sentence, (ii) relexicalisation fails for some input token for 34.07 % of cases and (iii) in total, 94.8% of the input content words are present in the output.

We also manually examined 50 randomly selected sentences[11] generated by our model. 34%

[10] We use sacrebleu (Post, 2018)

[11] The selected sentences are given in the supplementary

of these are exactly the same as the reference. 58% are both grammatically and semantically correct. 78% are grammatically correct. All verbs were found to be in the correct form (agreement and tense). We detected only one agreement error between a noun and its determiner. A detailed analysis of the errors found can be found in the supplementary material.

6 Conclusion

We introduced a new MR-to-text dataset for French and showed that full anonymization helps improve surface realisation. The automatic construction of parallel data using parsing makes available a detailed linguistic description of the target sentences to be generated. We are currently exploring how to extend the data creation method described in section 3 to better evaluate the ability of neural models to generate under syntactic constraints.

Acknowledgments

The research presented in this paper was partially supported by the E-FRAN Program (Espaces de Formation, de Recherche et d'Animation Numérique) within the framework of the METAL Project (Modèles et Traces au Service de l'Apprentissage des Langues).

References

Andrei Alexandrescu and Katrin Kirchhoff. 2006. Factored neural language models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 1–4. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186. Association for Computational Linguistics.

Daniel Beck, Gholamreza Haffari, and Trevor Cohn. 2018. Graph-to-sequence learning using gated graph neural networks. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 273–283, Melbourne, Australia. Association for Computational Linguistics.

material together with some statistics about their length.

Bernd Bohnet, Leo Wanner, Simon Mille, and Alicia Burga. 2010. Broad coverage multilingual deep sentence generation with a stochastic multi-level realizer. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 98–106. Association for Computational Linguistics.

Kris Cao and Stephen Clark. 2019. Factorising amr generation through syntax. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Mingje Chen, Gerasimos Lampouras, and Andreas Vlachos. 2018. Sheffield at e2e: structured prediction approaches to end-to-end language generation. Technical report, E2E Challenge System Descriptions.

Ondřej Dušek and Filip Jurcicek. 2015. Training a natural language generator from unaligned data. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 451–461. Association for Computational Linguistics.

Henry Elder and Chris Hokamp. 2018. Generating high-quality surface realizations using data augmentation and factored sequence models. In *Proceedings of the First Workshop on Multilingual Surface Realisation*, pages 49–53. Association for Computational Linguistics.

Jeffrey Flanigan, Chris Dyer, Noah A Smith, and Jaime Carbonell. 2016. Generation from abstract meaning representation using tree transducers. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 731–739.

Bruno Guillaume, Guillame Bonfante, Paul Masson, Mathieu Morey, and Guy Perrier. 2012. Grew: un outil de réécriture de graphes for le tal (grew: a graph rewriting tool for nlp). In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 5: Software Demonstrations*, pages 1–2.

Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. 2019. Text generation from knowledge graphs with graph transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

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)*, pages 146–157. Association for Computational Linguistics.

Diego Marcheggiani and Laura Perez-Beltrachini. 2018. Deep graph convolutional encoders for structured data to text generation. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 1–9, Tilburg University, The Netherlands. Association for Computational Linguistics.

Jonathan May and Jay Priyadarshi. 2017. Semeval-2017 task 9: Abstract meaning representation parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 536–545. Association for Computational Linguistics.

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 First Workshop on Multilingual Surface Realisation*, pages 1–12. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Nima Pourdamghani, Kevin Knight, and Ulf Hermjakob. 2016. Generating english from abstract meaning representations. In *Proceedings of the 9th International Natural Language Generation conference*, pages 21–25.

Linfeng Song, Xiaochang Peng, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2017. Amr-to-text generation with synchronous node replacement grammar. *arXiv preprint arXiv:1702.00500*.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. A graph-to-sequence model for AMR-to-text generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1616–1626, Melbourne, Australia. Association for Computational Linguistics.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. *International Conference on Learning Representations*. Accepted as poster.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721. Association for Computational Linguistics.