Evaluating Text Generation from Discourse Representation Structures

Chunliu Wang
CLCG
Univ. of Groningen
chunliu.wang@rug.nl

Rik van Noord
CLCG
Univ. of Groningen
r.i.k.van.noord@rug.nl

Arianna Bisazza
CLCG
Univ. of Groningen
a.bisazza@rug.nl

Johan Bos
CLCG
Univ. of Groningen
johan.bos@rug.nl

Abstract

We present an end-to-end neural approach to generate English sentences from formal meaning representations, Discourse Representation Structures (DRSs). We use a rather standard bi-LSTM sequence-to-sequence model, work with a linearized DRS input representation, and evaluate character-level and word-level decoders. We obtain very encouraging results in terms of reference-based automatic metrics such as BLEU. But because such metrics only evaluate the surface level of generated output, we develop a new metric, ROSE, that targets specific semantic phenomena. We do this with five DRS generation challenge sets focusing on tense, grammatical number, polarity, named entities and quantities. The aim of these challenge sets is to assess the neural generator's systematicity and generalization to unseen inputs.

1 Introduction

Faithfully generating text from structured representations is an important task in NLP. Common tasks include generations from tables (Parikh et al., 2020), knowledge graphs (Gardent et al., 2017) and meaning representations (Horvat et al., 2015; Flanigan et al., 2016; Dušek and Jurčíček, 2019). Recently, many research efforts have focused on the graph-based semantic formalism Abstract Meaning Representation (AMR, Banarescu et al., 2013), with approaches based on machine translation (Pourdamghani et al., 2016; Konstas et al., 2017), specialized graph encoders (Song et al., 2018; Zhu et al., 2019; Cai and Lam, 2020; Zhao et al., 2020; Jin and Gildea, 2020) and pre-trained language models (Mager et al., 2020; Ribeiro et al., 2020).

However, far less attention has been given to generating text from formal meaning representation, such as Discourse Representation Structures (DRSs). DRSs are proposed in Discourse Representation Theory (Kamp and Reyle, 1993; Kadmon, 2001; Geurts et al., 2020), a well-studied semantic formalism, covering a wide range of linguistic phenomena. Differently from AMR, DRSs explicitly model scope, tense and definiteness. The lack of this information makes AMR-to-text challenging (Wang et al., 2020), but their inclusion presents a challenge for the generation methods as well, as they, for example, have to deal with a lot more variables in the representation (van Noord et al., 2018a). Another difference with AMR is that DRSs are in principle language neutral (at least the version of DRS that we use in this paper), with gold standard annotations publicly available in four languages (Abzianidze et al., 2017). For these reasons, developing portable and high-quality generation systems for DRSs is a promising research direction.

While there has been some initial work on DRS-to-text generation (Basile and Bos, 2011; Narayan and Gardent, 2014; Basile, 2015), most DRS-based work has focused on semantic parsing, that is mapping text to DRS (Liu et al., 2018; van Noord et al., 2018b, 2019; Liu et al., 2019b; Evang, 2019; van Noord et al., 2020; Fancellu et al., 2020). Our work has two main contributions. The first is on the modelling side, as we take the first step in DRS-to-text generation with neural networks.[1] Specifically, we use a bi-LSTM sequence-to-sequence model that processes linearized DRSs representations and produces English texts using a character-level decoder (see pipeline in Figure 1).

Our second contribution regards the evaluation of the produced text. Given the known limitations of reference-based automatic metrics for natural language generation (Reiter and Belz, 2009; Novikova et al., 2017a) and in particular for AMR-to-text (May and Priyadarshi, 2017; Manning et al., 2020), we design five DRS-specific challenge sets (Popović and Castilho, 2019) and use them to per-

[1] Concurrently to this work, Liu et al. (2021) published a DRS-to-text model that is based on tree-LSTMs.

Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021), pages 73–83

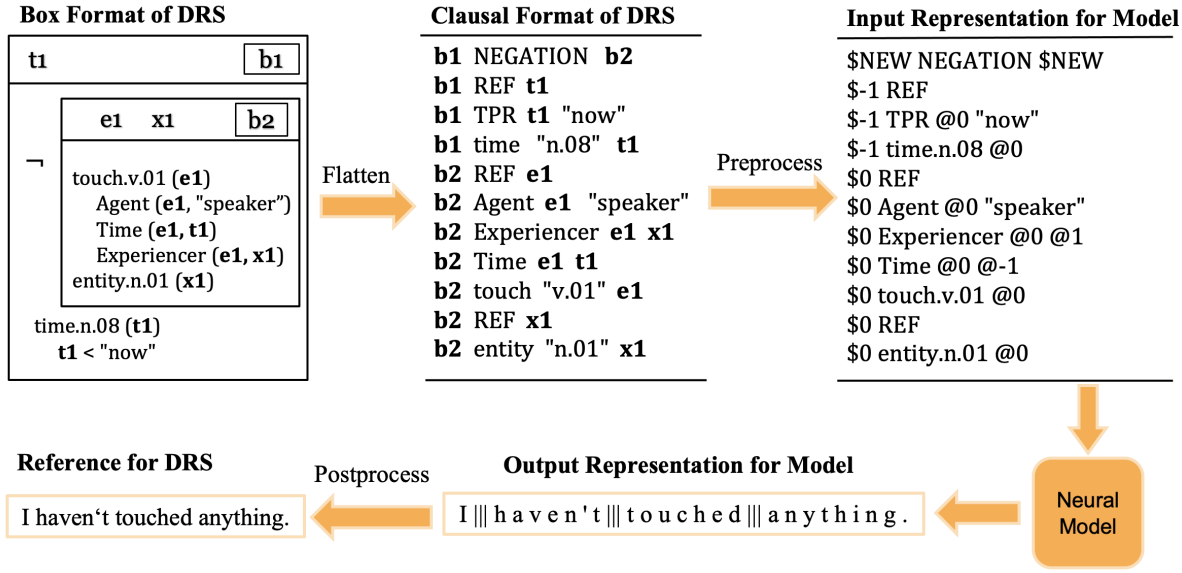

Figure 1: An example of the DRS data and a corresponding reference text with their processing procedures.

form a fine-grained manual evaluation. The general goal of these challenge sets is to assess the robustness of a DRS generator with respect to a number of linguistic phenomena. More specifically, we assess (i) generation systematicity with respect to three semantic phenomena (tense change, polarity change, singular↔plural switch), and (ii) generalization to unseen input literals (named entities and quantities). The idea is that by changing the meaning of a DRS in a controlled way, robustness of systems can be monitored in detail and assessed accordingly. Besides assessing the quality of a generator, these challenge sets also showcase the ease to which DRSs can be manipulated to express novel meaning combinations. All challenge sets are publicly available.[2]

2 Data and Methodology

In this section we describe the data and methodology we use for DRS generation. First we explain and motivate our representation of DRSs (input to the NLG system) and the generated text (see Figure 1 for a full overview of our source and target representations). Then we provide details of our NLG system, which is based on a recurrent neural network, and show how it is trained.

2.1 Input/Source Representation: DRSs

Discourse Representation Structures model the meaning of an entire text, ranging from isolated sentences to entire documents. A large repertoire of semantic phenomena is covered by DRSs, including quantification, negation, pronouns, comparatives, discourse relations, and presupposition. There are several variants of DRS; we use the fully interpretable version as employed in the Parallel Meaning Bank (Abzianidze et al., 2017), where concepts (triggered by nouns, verbs, adjectives and adverbs) are represented by WordNet synsets (Fellbaum, 1998), and semantic relations by Verbnet roles (Kipper et al., 2008).

DRS can be represented in box format or clause format (see Figure 1), where the letters *x*, *e*, *s*, and *t* are used for discourse referents denoting individuals, events, states, and time, respectively, and *b* is used for variables denoting DRSs. The clause format is a flat version of the standard box format, which represents DRS as a set of clauses. Due to its simple and flat structure, it has proven to be more suitable for machine learning tasks (van Noord et al., 2018a). The variables that occur in a DRS are rewritten using the relative naming method based on de Bruijn-indexing (Bruijn, de, 1972)).

We mostly follow van Noord et al. (2018b) in how to represent DRSs for neural processing, but make some important improvements. The idea is to represent meaningful units as atomic entites. These include the variable indices (`$0`, `@1`), the DRS operators (`REF`, `NOT`), the semantic relations (e.g., `Agent`, `Patient`, `Theme`), the deictic constants (`now`, `speaker`, `hearer`), and the concepts (e.g., `touch.v.01`).

The latter is a notable exception to van Noord et al. (2018b). By representing concepts, that correspond to WordNet-synsets, as single entities,

[2] https://github.com/wangchunliu/DRS-generation

we make sure that each concept is mapped to a language-independent embedding, even though its surface form may resemble the corresponding English word. This prevents the model from learning to predict target words (e.g., `touch`) by copying (part of) the characters that compose the Wordnet-synset (e.g., `touch.v.01`) in the input DRS.

The remaining parts of the DRSs are represented at the character-level. These include time/date expressions (e.g., `" 1 9 6 8 "`), value expressions such as scores (e.g., `" 2 - 0 "`), quantities (e.g., `" 2 6 0 0 "`) , and proper names (e.g., `" b r a d ~ p i t t "`). They are all enclosed in quotation marks in the DRS representation. It would not make sense to represent these entities as words because times, dates, and quantities are clearly of compositional nature. Names are literal expressions, and therefore also are best represented by separate characters. Moreover, this representation reduces the size of the vocabulary, which in turn could reduce the learning difficulty of the model.

2.2 Output/Target Representation: Text

The spectrum to represent text ranges from single characters on one end till (tokenised) words or multi-word expressions on the other end, and there are many possibilities in between too, for instance using byte-pair encodings to combine characters into sub-words. As our aim is to get a relatively straightforward baseline NLG system, rather than exploring the full range of text representation possibilities, we considered just two ways to represent text: character-based, where raw characters are separate entities and spaces are indicated by a special symbol (three vertical bars); or (tokenised) word-based, where tokenised words form the basic entities. The character-based approach has the advantage that post-processing is straightforward. The use of word-level representations is the classical approach in natural language processing, but requires a de-tokenisation step after generating. Tokenisation and de-tokenisation is carried out with the Moses tokenizer (Koehn et al., 2007).

2.3 Neural Generation Model

We use a standard recurrent encoder-decoder architecture with attention as implemented in the Marian toolkit (Junczys-Dowmunt et al., 2018), using two bi-directional LSTM layers (Hochreiter and Schmidhuber, 1997). In particular, we use an embedding size of 300 for both the encoder and decoder, a mini-batch size of 48 and the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.002. All hyper-parameters are shown in Table1. We use the English gold standard training, dev and test data of PMB 3.0.0[3], containing 6,620, 885 and 898 instances, respectively. During training, we merge the gold standard with the only partially manually annotated silver standard of 97,598 instances. Differently from van Noord et al. (2018b), we do not fine-tune on the gold standard data in a second step, as this did not lead to improved performance.

Parameter	Value	Parameter	Value
dim-emb	300	dim-rnn	300
dec-cell	lstm	enc-depth	2
enc-cell	lstm	dec-depth	2
mini-batch	48	lr-decay	0.5
lr-decay-strategy	epoch	normalize	0.9
beam-size	10	learn-rate	0.002
dropout-rnn	0.2	cost-type	ce-mean
label-smoothing	0.1	optim	adam
early-stop	3	valid-metric	cross-entropy

Table 1: Hyperparameter settings of our experiments.

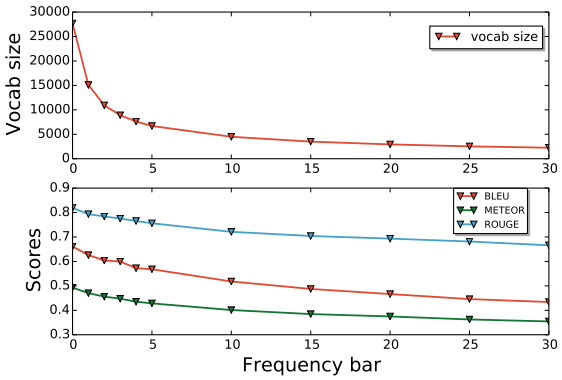

Figure 2: The correlation between the vocabulary size and the frequency threshold, along with the correlation between metric scores and the frequency threshold. Threshold set to 0 means using the full vocabulary.

Vocabulary For a word-level model, it can be beneficial to not include the full vocabulary. For example, it might learn to handle unknown words better if it was exposed to unknown word tokens during training. We experimented with the vocabulary size of the target representation on the development set, as is shown in Figure 2. We find that the we get best performance when including the full vocabulary, with decreasing performance as we decrease the vocabulary. We use this setting for our word-level experiments.

[3]https://pmb.let.rug.nl/data.php

3 Semantic Challenge Sets

Challenge sets are often used in Machine Translation to assess a model's ability to systematically deal with specific linguistic phenomena that may be infrequent in standard test sets (Popović and Castilho, 2019). Following this practice, we created five challenge sets for DRSs generation that focus on various semantic phenomena (see Table 2 and Figure 3). The variations are obtained by (manually) applying a minimal modification to a DRS and editing the corresponding text accordingly.

The resulting semantic challenge sets can be viewed as stress tests: if the generator performs well on these test suites it shows that it can deal with specific semantic phenomena adequately in unforeseen circumstances. We carry out these modifications on subsets of the PMB test data, and we group them into those that assess *systematic predictions* (tense, polarity, and grammatical number) and those that assess *generalisation to unseen input* (names and quantities). The specific challenge sets are described in detail below.

Original	Tom has three thousand books.
Tense	Tom had three thousand books.
Polarity	Tom does not have three thousand books.
Number	Tom has one book.
Names	Kirk has three thousand books.
Quantity	Tom has 3,200 books.

Table 2: Examples of how the challenge set DRSs are created. We show the reference texts of the modified DRSs here.

3.1 Tense Change

In English, tense is expressed by morphology and the use of auxiliary verbs. It is therefore a challenging phenomenon for NLG. There are three types of tense found in the DRSs of the Parallel Meaning Bank: past (t $<$ now), present (t $=$ now), and future tense (t $>$ now). Aspect is not covered in detail in the Parallel Meaning Bank, and therefore we won't address it in the paper and as a result it won't be part of the current semantic challenge sets.

For creating the challenge set, we used the following procedure. For the first 200 examples in the test set that contained information about tense in their corresponding DRSs, we changed the tense in the DRS: past to present or future, present to past or future, and future to past or present. The corresponding text was changed to reflect the change in tense. Example: She bought a vacuum cleaner at the supermarket. $\rightarrow$ She will buy a vacuum cleaner at the supermarket.

3.2 Polarity Change

As negation plays a crucial role to determine the truth conditions of a sentence, there has been ample interest in recognizing negation in text (Morante and Blanco, 2012; Basile et al., 2012) and translating accurately (Sennrich, 2017; Tang, 2020). Here we focus on generation, that is expressing negation appropriately in a sentence given a meaning representation. Negation is expressed in a DRS with a unary operator, introducing an embedded DRS. For the first 100 instances of the test set we removed negation if it was already present, or, more frequently, added it if it was not. Again, the corresponding reference text was changed to reflect this change in meaning. Example: I cooked dinner. $\rightarrow$ I didn't cook dinner.

3.3 Grammatical Number Change

Concrete quantities are expressed in DRSs with the relation `Quantity` and a number. For the first 100 examples that permitted this, we changed the quantity from a number greater than one to one, or vice versa. This set can be used to check whether the model can recognize the number and generate the correct plural form of nouns to get the correct noun phrase (Sennrich, 2017). Example: It currently employs 180 people. $\rightarrow$ It currently employs one person. As many languages (including English) have a different surface realisation for singular and plural, an NLG system needs to handle this correctly.

3.4 Names Change

The goal of this challenge set is to assess the behaviour of NLG systems that find unexpected (not seen in training data) proper names in the meaning representation input. We took the first 50 instances of the test set with named entities (persons, locations, organisations, artifacts) and modified the DRSs in such a way that the names entities are replaced by alternative, but realistic names of the same type of entity and gender (in case of persons), that do not occur in the training data. Consider a sentence with the name "Howard Caine", with `Name(x, howard~caine)` in its corresponding DRS. We change this into a real name outside the coverage of the training data, e.g., `Name(x, howard~carpendale)`. This should generate

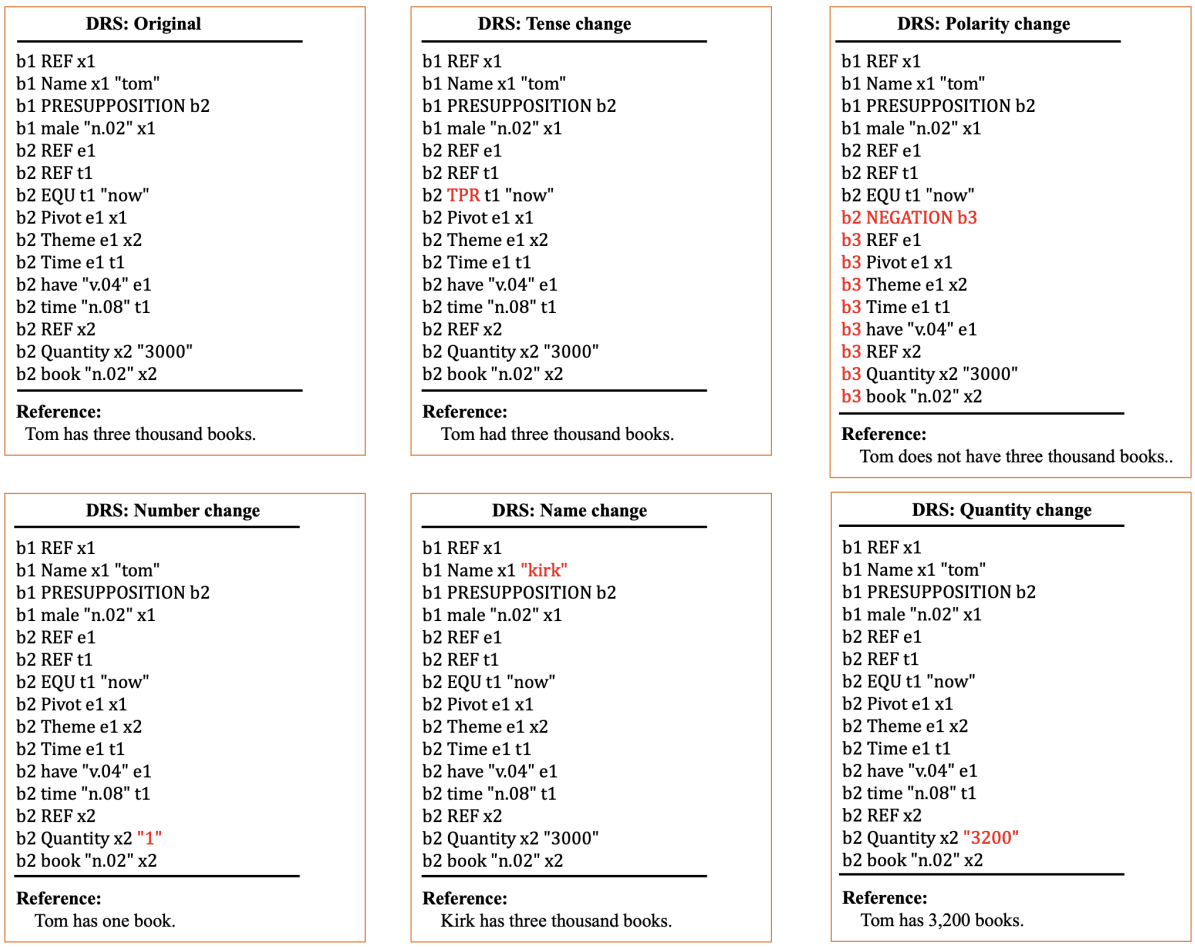

Figure 3: Examples of how the challenge set DRSs are created. Modified DRSs correspond to Table 2.

"Howard Carpendale", for which word-based systems would be expected to face more difficulties than character-based systems.

3.5 Quantities Change

In addition to named entities in meaning representation, the numeral expressions can also be changed to expressions that were never seen in the training data. We took the first 50 instances of the test set with numbers and then changed the numbers in the DRS representation to unknown quantity expressions, represented as a sequence of characters. For example, we changed `Quantity(x, 150)` to `Quantity(x, 152)`. This way, we test if the model can systematically generalize to generate the right numeral expression, even though it has not seen this particular sequence of characters before.

4 Assessment Methods

We consider two types of assessment for the generated English sentences. Our point of departure are the well-known automatic metrics based on word overlap. We complement these with manual metrics carried out by human experts.

4.1 Standard Automatic Metrics

We use three standard metrics measuring word-overlap between system output and references. They are BLEU (Papineni et al., 2002) used as standard in machine translation evaluation and very common in NLG, METEOR (Lavie and Agarwal, 2007), and ROUGE-L (Lin, 2004), which were applied in the COCO caption generation challenge as well as other NLG experiments (Novikova et al., 2017b; Dušek et al., 2020). As is well known, these standard metrics give a first, rough impression about the quality of the generated output, but often reveal only part of the story. This is why we also consider a further form of assessment.

4.2 Expert Assessment

Inspired by work of Jagfeld et al. (2018) and Belz et al. (2020), we believe that the manual evaluation method for our task should be simple in definition,

	BLEU	METEOR	ROUGE
Char-level (raw)	69.3	51.8	84.9
Word-level (tok)	64.7	47.8	81.8

Table 3: Performance of English DRS-to-text with two output representations, averaged over three runs.

easy to reproduce and high in generalization ability. The output of our NLG system was manually assessed by one expert. This was carried out by assigning three binary dimensions (either 0 or 1) to each generated text: (1) semantics; (2) grammaticality, and (3) phenomenon. As shown in Table 5: the first dimension, *semantics*, gets a score 1 if the meaning of the output reflects that of the underlying meaning representation, and 0 otherwise. The second dimension, *grammaticality*, receives a score 1 if the sentence is grammatical and free of spelling mistakes (but possibly gibberish), and 0 otherwise. The third dimension, *phenomenon*, gets a 1 if the phenomenon of control is generated at all, and 0 otherwise. We summarise these three dimensions into one score by taking the product of these numbers, and refer to this score as ROSE (Robust Overall Semantic Evaluation). Hence, a ROSE-score of 1 is given to output that is perfect (three ones); a ROSE-score of 0 is given if one of the three scores yields zero. Note that, usually, if the score for *phenomenon* is 0, then it follows that the score for *semantics* is 0, too.

5 Results and Analysis

Table 3 shows the performance of the models based on characters and words. The character-level model clearly outperforms the model based on word-tokenised text on all three automatic metric scores. This is in line with work on DRS parsing (van Noord et al., 2018b, 2019; Liu et al., 2019a) and other NLG tasks (Goyal et al., 2016; Agarwal and Dymetman, 2017; Jagfeld et al., 2018), where character-based models outperform word-based models. We will use the character-level model for the rest of the experiments in this paper.

5.1 Challenge Sets

Table 4 shows the overall results on the challenge sets for both the automatic evaluation results and manual evaluation. We can see that performance is hardly affected for the number, quantity and names challenge sets on the automatic evaluation metrics. It seems that our character-based model can indeed learn the shallow information contained in the input data and copy it to generate, even if these subsets (numbers, quantities and name entities) in the DRSs do not appear in the training set. However, for tense and polarity, all three automatic metrics are significantly lower in the challenge sentences than in the original sentences. Through the observation of the generated texts of the tense challenge set, we find that it is difficult for the model to generate future tense sentences, but past tense and present tense can be generated well. The original test set contained not so many DRSs in future tense, but in the challenge set we added relatively many of them, which likely caused the lower performance on the challenge set.

With regards to the polarity challenge set, inspection of the output shows that a common error is to confuse "never" with "not". This difference in meaning is reflected in a DRS by the relative order of the reference time and the DRS negation operator. Interestingly, recent work in machine translation (Tang, 2020) and language modelling (Ettinger, 2020) has also shown that state-of-the-art neural models still struggle with handling negation.

Although the results of the automatic evaluation metrics in the last three challenge sets have no obvious changes compared with the original data sets, our manual evaluation results show that the performance of the model in all challenge sets is lower than the original data sets. This further shows that there is not always a positive correlation between automatic evaluation and manual evaluation, and it is still necessary to rely on manual evaluation.

5.2 Error Analysis

Table 5 shows a number of interesting outputs of our DRS-to-text model. Sometimes, the model outputs a combination of characters that is clearly wrong, such as in (a), though it still captured the phenomenon that the challenge set checks for (tense). Sentence (b) is a common mistake for the polarity challenge set: the model generates a negation in a grammatical way, but it is not the correct one. In (c) we show a mistake that occurs for the tense challenge set, in which the model was not able to capture the correct tense. Sentence (d) shows that the model sometimes has trouble with longer character-level sequences of numbers. Perhaps the model learned that the sequence "1 5" is generated as "fifteen" as text, which in this case resulted in the wrong output. In (e), the model

		BLEU		METEOR		ROUGE		Sem.		Gram.		Phen.		ROSE	
	#	Orig	Chal	Orig	Chal	Orig	Chal	Orig	Chal	Orig	Chal	Orig	Chal	Orig	Chal
Tense	200	68.4	55.8	50.9	44.8	85.0	76.1	80.0	71.0	92.0	87.5	99.5	86.5	78.0	64.0
Polarity	100	68.1	37.4	50.8	37.9	85.0	66.1	80.0	52.0	96.0	81.0	100.0	99.0	78.0	49.0
Number	100	72.5	69.2	53.7	53.4	85.7	86.4	80.0	79.0	95.0	84.0	100.0	95.0	77.0	69.0
Names	50	69.1	71.9	53.0	53.5	87.2	87.8	82.0	76.0	94.0	84.0	100.0	98.0	82.0	74.0
Quantity	50	69.7	68.0	56.4	50.6	86.0	83.4	88.0	72.0	98.0	90.0	92.0	84.0	86.0	70.0

Table 4: Performance of the character-level model for five different challenge sets. We report scores on both the original input (Orig) of the challenge sets and the actual challenge sets (Chal). The first three scores are automatic metrics, while the last four scores are accuracies based on human evaluation (see Section 4.2). **Sem.**, **Gram.**, and **Phen.** stand for *Semantics*, *Grammaticality* and *Phenomenon*, respectively.

	Reference text	Generated text	Sem.	Gram.	Phen.	ROSE
(a)	She liked short skirts.	She liked short tomical.	0	0	1	0
(b)	Tom does not have three thousand books.	Tom never has three thousand books.	0	1	1	0
(c)	The small skirt will be pink.	The small skirt was pink.	0	1	0	0
(d)	He left 157 minutes ago.	He left fifteen minutes ago.	0	1	0	0
(e)	I checked it nine times.	I checked it nine.	0	0	1	0
(f)	We are painting the house green.	I paint the house green.	1	1	1	1
(g)	That hat cost around fifty dollars.	This hat cost about 50 dollars.	1	1	1	1
(h)	When I painted this picture, I was 23 years old.	I painted the picture when I was twenty-three years old.	1	1	1	1

Table 5: Examples of generated texts from the challenge set DRSs, compared with reference texts. Note that the input for the model is a linearized DRS, not the reference text.

managed to capture the phenomenon (quantity), but did this in an non-grammatical way not preserving the meaning. Sentence (f) is interesting, because the DRS representation does not differentiate between "I" and "We", meaning the model can not be expected to (always) output the correct version. Therefore, such differences are not counted as a mistake during human evaluation. Finally, the output of (g) and (h) shows the necessity of human evaluation: the model produced sentences that captured the meaning perfectly, but used a different surface realization than in the reference text.

6 Conclusion and Future Work

We presented an end-to-end neural approach to generate natural language from Discourse Representation Structures. Our model is based on a bi-LSTM sequence-to-sequence architecture taking linearized DRSs as input. Comparing character level with word level for producing text, it achieves higher BLEU, METEOR and ROUGE scores on the former.

For a better understanding of our generator's robustness and its reliability, we designed several challenge sets focusing on specific semantic phenomena (tense, polarity, grammatical number) and types of unseen input (quantity and named entities). Automatic and manual evaluations on these challenge sets point out to negation as the most challenging phenomenon for DRS generation, followed by tense. By contrast, changes in grammatical number and generalizations to unseen quantities or names are well handled by the model.

Altogether, our results suggest that neural generation from DRSs is a very promising research direction, but more work is needed to ensure reliability in real-world applications. We hope that our challenge sets will foster future research on this topic and eventually lead to truly robust DRS generators. The challenge sets, as we have presented them, can be further refined, and other linguistic phenomena can be added as well. Possibilities that spring to mind are challenge sets for pronouns, definite descriptions, comparatives, aspect, and discourse particles. And obviously, we need to generate challenge sets for languages other than English, which might trigger language-specific phenomena as well that could be suitable for challenge sets for DRS generation.

Acknowledgements

This work was funded by the NWO-VICI grant "Lost in Translation—Found in Meaning" (288-89-003). The first author is supported by the China Scholarship Council (CSC201904890008). Arianna Bisazza was partly funded by the Netherlands Organization for Scientific Research (NWO) under project number 639.021.646. The Tesla K40 GPU used in this work was kindly donated to us by the NVIDIA Corporation. We would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster. Finally, we thank the anonymous reviewers for their insightful comments.

References

Lasha Abzianidze, Johannes Bjerva, Kilian Evang, Hessel Haagsma, Rik van Noord, Pierre Ludmann, Duc-Duy Nguyen, and Johan Bos. 2017. The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.

Shubham Agarwal and Marc Dymetman. 2017. A surprisingly effective out-of-the-box char2char model on the e2e nlg challenge dataset. In *SIGDIAL Conference*, pages 158–163.

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, Sofia, Bulgaria. Association for Computational Linguistics.

Valerio Basile. 2015. *From logic to language: Natural language generation from logical forms*. Ph.D. thesis, University of Groningen.

Valerio Basile and Johan Bos. 2011. Towards generating text from discourse representation structures. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 145–150, Nancy, France. Association for Computational Linguistics.

Valerio Basile, Johan Bos, Kilian Evang, and Noortje Venhuizen. 2012. UGroningen: Negation detection with discourse representation structures. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 301–309, Montréal, Canada. Association for Computational Linguistics.

Anya Belz, Simon Mille, and David M. Howcroft. 2020. Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 183–194, Dublin, Ireland. Association for Computational Linguistics.

N.G. Bruijn, de. 1972. Lambda calculus notation with nameless dummies, a tool for automatic formula manipulation, with application to the church-rosser theorem. *Indagationes Mathematicae (Proceedings)*, 75(5):381–392.

Deng Cai and Wai Lam. 2020. Graph transformer for graph-to-sequence learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7464–7471.

Ondřej Dušek and Filip Jurčíček. 2019. Neural generation for Czech: Data and baselines. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 563–574, Tokyo, Japan. Association for Computational Linguistics.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2020. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech & Language*, 59:123 – 156.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Kilian Evang. 2019. Transition-based DRS parsing using stack-LSTMs. In *Proceedings of the IWCS Shared Task on Semantic Parsing*, Gothenburg, Sweden. Association for Computational Linguistics.

Federico Fancellu, Ákos Kádár, Ran Zhang, and Afsaneh Fazly. 2020. Accurate polyglot semantic parsing with DAG grammars. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3567–3580, Online. Association for Computational Linguistics.

Christiane Fellbaum. 1998. Wordnet: An electronic lexical database. *The MIT Press, Cambridge, Ma., USA*.

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, San Diego, California. Association for Computational Linguistics.

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

Bart Geurts, David I. Beaver, and Emar Maier. 2020. Discourse Representation Theory. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, spring 2020 edition. Metaphysics Research Lab, Stanford University.

Raghav Goyal, Marc Dymetman, and Eric Gaussier. 2016. Natural language generation through character-based RNNs with finite-state prior knowledge. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1083–1092, Osaka, Japan. The COLING 2016 Organizing Committee.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Matic Horvat, Ann Copestake, and Bill Byrne. 2015. Hierarchical statistical semantic realization for Minimal Recursion Semantics. In *Proceedings of the 11th International Conference on Computational Semantics*, pages 107–117, London, UK. Association for Computational Linguistics.

Glorianna Jagfeld, Sabrina Jenne, and Ngoc Thang Vu. 2018. Sequence-to-sequence models for data-to-text natural language generation: Word- vs. character-based processing and output diversity. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 221–232, Tilburg University, The Netherlands. Association for Computational Linguistics.

Lisa Jin and Daniel Gildea. 2020. Generalized shortest-paths encoders for AMR-to-text generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2004–2013, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marcin Junczys-Dowmunt, Kenneth Heafield, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. 2018. Marian: Cost-effective high-quality neural machine translation in C++. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 129–135, Melbourne, Australia. Association for Computational Linguistics.

Nirit Kadmon. 2001. *Formal Pragmatics*. Blackwell.

Hans Kamp and U. Reyle. 1993. From discourse to logic: Introduction to model theoretic semantics of natural language, formal logic and discourse representation theory. *Language*, 71(4).

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42:21–40.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

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, Vancouver, Canada. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2018. Discourse representation structure parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 429–439, Melbourne, Australia. Association for Computational Linguistics.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2019a. Discourse representation parsing for sentences and documents. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6248–6262, Florence, Italy. Association for Computational Linguistics.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2019b. Discourse representation structure parsing with recurrent neural networks and the transformer model. In *Proceedings of the IWCS Shared Task on Semantic Parsing*, Gothenburg, Sweden. Association for Computational Linguistics.

Jiangming Liu, Shay B. Cohen, and Mirella Lapata. 2021. Text generation from discourse representation structures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 397–415, Online. Association for Computational Linguistics.

Manuel Mager, Ramón Fernandez Astudillo, Tahira Naseem, Md Arafat Sultan, Young-Suk Lee, Radu Florian, and Salim Roukos. 2020. GPT-too: A language-model-first approach for AMR-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1846–1852, Online. Association for Computational Linguistics.

Emma Manning, Shira Wein, and Nathan Schneider. 2020. A human evaluation of AMR-to-English generation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4773–4786, Barcelona, Spain (Online). International Committee on 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, Vancouver, Canada. Association for Computational Linguistics.

Roser Morante and Eduardo Blanco. 2012. *SEM 2012 shared task: Resolving the scope and focus of negation. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 265–274, Montréal, Canada. Association for Computational Linguistics.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 435–445, Baltimore, Maryland. Association for Computational Linguistics.

Rik van Noord, Lasha Abzianidze, Hessel Haagsma, and Johan Bos. 2018a. Evaluating scoped meaning representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Rik van Noord, Lasha Abzianidze, Antonio Toral, and Johan Bos. 2018b. Exploring neural methods for parsing discourse representation structures. *Transactions of the Association for Computational Linguistics*, 6:619–633.

Rik van Noord, Antonio Toral, and Johan Bos. 2019. Linguistic information in neural semantic parsing with multiple encoders. In *Proceedings of the 13th International Conference on Computational Semantics - Short Papers*, Gothenburg, Sweden. Association for Computational Linguistics.

Rik van Noord, Antonio Toral, and Johan Bos. 2020. Character-level representations improve DRS-based semantic parsing Even in the age of BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4587–4603, Online. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017a. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017b. 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, Saarbrücken, Germany. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. 2020. ToTTo: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1173–1186, Online. Association for Computational Linguistics.

Maja Popović and Sheila Castilho. 2019. Challenge test sets for MT evaluation. In *Proceedings of Machine Translation Summit XVII Volume 3: Tutorial Abstracts*, Dublin, Ireland. European Association for Machine Translation.

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, Edinburgh, UK. Association for Computational Linguistics.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Comput. Linguist.*, 35(4):529–558.

Leonardo FR Ribeiro, Martin Schmitt, Hinrich Schütze, and Iryna Gurevych. 2020. Investigating pretrained language models for graph-to-text generation. *arXiv preprint arXiv:2007.08426*.

Rico Sennrich. 2017. How grammatical is character-level neural machine translation? assessing MT quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382, Valencia, Spain. Association for Computational Linguistics.

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.

Gongbo Tang. 2020. *Understanding Neural Machine Translation: An investigation into linguistic phenomena and attention mechanisms*. Ph.D. thesis, Uppsala University, Department of Linguistics and Philology.

Tianming Wang, Xiaojun Wan, and Hanqi Jin. 2020. AMR-to-text generation with graph transformer. *Transactions of the Association for Computational Linguistics*, 8:19–33.

Chao Zhao, Marilyn Walker, and Snigdha Chaturvedi. 2020. Bridging the structural gap between encoding and decoding for data-to-text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2481–2491, Online. Association for Computational Linguistics.

Jie Zhu, Junhui Li, Muhua Zhu, Longhua Qian, Min Zhang, and Guodong Zhou. 2019. Modeling graph structure in transformer for better AMR-to-text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5459–5468, Hong Kong, China. Association for Computational Linguistics.