Exploring Data Augmentation in Neural DRS-to-Text Generation

Anonymous ACL submission

Abstract

Neural networks are notoriously data-hungry. This represents an issue in cases where data are scarce such as in low-resource languages. Data augmentation is a technique that is commonly used in computer vision to provide neural networks with more data and for increasing their generalization power. When dealing with data augmentation for natural language, however, simple data augmentation techniques similar to the ones used in computer vision such as rotation and cropping cannot be employed because they would generate ungrammatical texts. Thus, data augmentation needs a specific design in the case of neural data-to-text systems, especially for a structurally rich input format such as the ones used for meaning representation. This is the case of the neural natural language generation for Discourse Representation Structures (DRS-to-Text), where the logical nature of DRS needs a specific design of data augmentation. In this paper, we adopt a novel approach in DRS-to-Text to selectively augment a training set with new data by adding and varying two specific lexical categories, i.e. proper and common nouns. In particular, we propose to use WordNet supersenses for producing new training sentences using both *in*-and-*out* context nouns. We present a number of experiments for evaluating the role played by augmented lexical information. The experimental results prove the effectiveness of our approach for data augmentation in DRS-to-Text generation.

1 Introduction

Data augmentation is a systematic way of increasing data examples by altering the original data with controlled variations (Feng et al., 2021). It is a prevalent technique in computer vision (CV) for increasing dataset size by introducing slightly different and contextually similar examples (Yang et al., 2022).

Augmentation approaches are also becoming popular in many Natural Language Processing (NLP) applications as well. The most commonly used approaches to augment textual data are based on random swapping, random insertion, random deletions, synonyms replacement, back translation, and using generative models to get new context-aware data (Feng et al., 2021; Shorten and Khoshgoftaar, 2019). Notice that data augmentation in NLP is a very challenging task due to the constraint of producing a grammatical augmented text (Hou et al., 2018). Moreover, given the continuous nature of images, in CV the augmented version of an image rarely is *pragmatically* incorrect. In contrast, in NLP, preserving the contextual meaning of the sentence is, usually, a hard constraint. Indeed, bad model performance can be the consequence of augmented textual data that is grammatically incorrect or out-of-scope (Dong et al., 2017).

Recently, researchers working on text generation from meaning representations, i.e., graph-based Abstract Meaning Representation (AMR) (Banarescu et al., 2013; Flanigan et al., 2016) or Discourse Representation Structure (DRS) (Noord, 2019; van Noord et al., 2018; Basile and Bos, 2011; Wang et al., 2021; Amin et al., 2022), have put their efforts into generating text from logical representations, and vice-versa, using transformers and encoder-decoder-based neural models (Noord, 2019; van Noord et al., 2018; Wang et al., 2021; Amin et al., 2022). In this paper, we consider the specific problem of augmenting data in the context of neural DRS-to-Text generation task. DRS represents textual information in the form of events, concepts, and entities, i.e., names as discourse referents usually represented as variables in DRS, and logical relations between these entities i.e., quantifiers, conjunctions, negations, disjunctions, etc. (Bos, 2021; Kamp and Reyle, 1993; Jaszczolt, 2023). In Fig. 1 a graphical representation of DRS in box format (on the left), its flattened version i.e., clausal format (on the right), and its corresponding textual representation (on the bottom) is displayed.

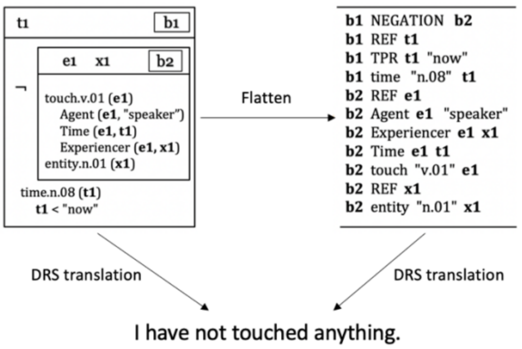

Figure 1: Box format and Clausal format of DRS along with their textual representation.

Neural DRS-to-Text generation is a type of data-to-text generation task that takes the logical representation of a sentence as input and generates text as output (Wang et al., 2021; Amin et al., 2022). This is an application of text generation from structured input data similar to knowledge graphs (Flanigan et al., 2016), RDF triplets data (Gardent et al., 2017), and tables (Parikh et al., 2020). Note that, in contrast to tables and graphs, the ability to represent the structured logical nature of the input as a DRS generation allows for a more fine-grained investigation of the relation between input and output in DRS-to-Text. In other words *"changing the meaning of a DRS in a controlled way, the robustness of systems can be monitored in detail and assessed accordingly"* (Wang et al., 2021). However, this robustness property discourages the application of large language models (LLMs) for augmenting data because LLMs would generate noise in the augmented data (Feng et al., 2021; Hou et al., 2018; Dong et al., 2017) – see also Section 4.

In this paper, we exploit the robustness property of neural DRS-to-Text generation by designing and evaluating data augmentation for the specific categories of (i) proper nouns and (ii) common nouns. In particular, we have designed and evaluated a procedure for augmenting a DRS training dataset by adding *context-aware* new sentences that are produced by varying the proper and common nouns in the original sentences. We consider different strategies and propose to use Supersenses Tagging (SST) for creating new training sentences using both *in-and-out* context nouns. In this way, we want to analyze the role played by lexical information in the performance of a neural DRS-to-Text system.

The research questions and contributions addressed in this paper are:

- Is it possible to augment a logical data representation such as DRS?
- How to generate new data that is contextually similar to the original one?
- What is the role played by the *in-and-out* contextual vocabulary for char-level and word-level decoder models? And what is the role of grammatical-semantic-pragmatic-world knowledge in learning?
- Does augmentation result in an increase or decrease in model performance?
- What is the behavior of the state-of-the-art large language models i.e., ChatGPT, while analyzing DRS structures?

To the best of our knowledge, apart from the preliminary work on augmentation of verbs presented in (Amin et al., 2022), this is the first on data augmentation in DRS-to-Text generation analyzing its impact on model performance.

Notice that our augmentation techniques could generate factually incorrect texts (e.g., starting from "at dawn, the sun rises", "at night, the sun rises" could be generated. However, since humans can generate texts that are not factually correct (consider, for example, a sci-fi story), preventing this situation would actually be not beneficial, but detrimental for the system.

The statistical nature of the neural networks does not allow for an easy analysis of the kind of knowledge really learned by the system. When we provide a specific example as *Brad Pitt is an actor*, the network is learning that the verb follows the subject (e.g. grammatical competence), and/or that a man can be an actor (semantic and pragmatic knowledge), and/or that a specific man is an actor (world knowledge)? How can we exploit this multi-level nature of neural learning? A side effect of our study on data augmentation is to investigate on these theoretical questions as well.

The paper is structured as follows: in Section 2, we describe the procedure adopted for noun augmentation; in Section 3, we give architectural insights on the neural DRS-to-Text pipeline; in Section 4, we describe the experimental results of DRS-to-text generation that uses (1) automatic metrics on a standard test set, (2) a reduced test set comparing the neural system with two general LLMs,

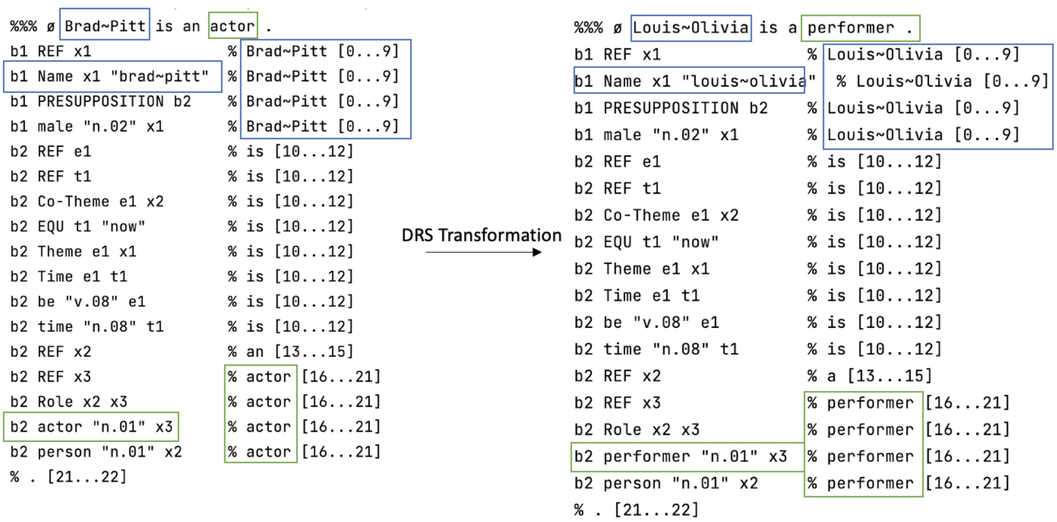

Figure 2: Graphical representation of the DRS transformation as a proper noun (in blue) and common noun (in green). The DRS on the left generates the sentence *Brad Pitt is an actor.*, while the DRS on the right generates *Louis Olivia is a performer.*

and (3) applying both automatic and human evaluation metrics. Finally, in Section 5, we conclude the paper.

2 Logical Data Augmentation with Nouns

Data augmentation is a relatively complex task in the case of neural DRS-to-Text: each augmented example in the training set consists of a pair of a new DRS structure together with a new corresponding sentence. While applying systematic transformations on training data, it is essential to keep track of both types of data representations as they are treated as *input value pairs* in the neural model. So, data transformations should be identical and symmetrical on both elements by considering the order of meaning representations and textual translations.

In the DRS-to-Text generation task, we applied different augmentation techniques for augmenting proper nouns and common nouns. We have used the gold version of the Parallel Meaning Bank[1] (PMB) dataset, which is organized in the usual train-dev-test split.

A graphical representation of transformation for proper (highlighted in blue) and common (highlighted in green) nouns in DRS is shown in Fig. 2: the DRS on the left generates the sentence *Brad Pitt is an actor*, while the DRS on the right generates *Louis Olivia is a performer* (see Table 1).

[1]The PMB is developed at the University of Groningen as part of the NWO-VICI project "Lost in Translation – Found in Meaning" (Project number 277-89-003), led by Johan Bos.

2.1 Proper Noun Augmentation

For proper nouns, we considered two specific name entity (NE) categories, which are the proper name of a person (PER) and of a place i.e., city, state, or country (GPE). We have used spaCy NE recognizer (https://spacy.io) to extract proper nouns from the text. There are a total amount of 3773 proper noun instances for PER and GPE. The proper nouns are divided as follows: person names 57%, city names 30%, state names 6%, country names 6%, and 1% other types as shown in Fig. 3.

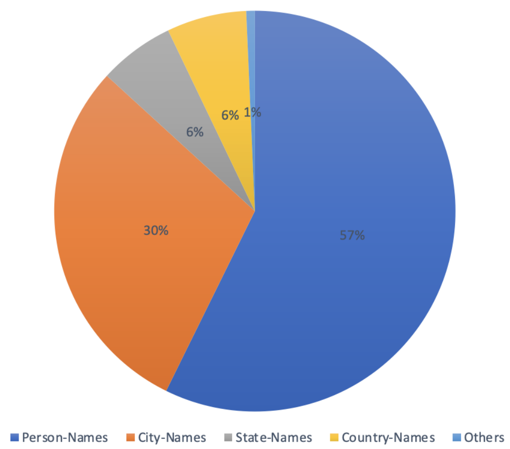

Figure 3: Distribution of proper noun entities in Gold-PMB dataset.

We have used two procedures for replacing proper nouns to analyze the impact of adding external linguistic information to the dataset vocabu-

Transf Type	Original Text	Transformed Augmented Text
Proper Noun	Brad Pitt is an actor. Alice and Bob work for this company. Turin is a beautiful city. Indiana is a very famous state. China is one of the top 5 populous countries in the world.	Louis Olivia is an actor. Maria and Tom work for this company. Venice is a beautiful city. Georgia is a very famous state. India is one of the top 5 populous countries in the world.
Common Noun	Brad Pitt is an actor. Alice and Bob work for this company. Turin is a beautiful city. We painted the house green. The book rested on the table.	Brad Pitt is a performer. Alice and Bob work for this corporation. Turin is a beautiful metropolis. We painted the building green. The novel rested on the desk.
Proper and Common Noun	Brad Pitt is an actor. The Mona Lisa hung above the antique table. Alice and Bob work for this company. Noah and Sophia watched a movie at the local theater. Oliver and Isabella enjoyed the view of the mountains from the cabin.	Louis Olivia is a performer. The Leonardo da Vinci hung above the antique desk. Maria and Tom work for this corporation. Liam and Emma watched a film at the local cinema. Daniel and Lily enjoyed the view of the peaks from the lodge.

Table 1: Different flavors of augmentation applied to the dataset as single and blended data transformations.

lary[2]. (1) Replacing them with other proper nouns inside the same dataset, i.e., *inside context*. (2) Replacing them with proper nouns outside the dataset, i.e., *outside context*.

For replacing proper nouns via *outside context* approach, we choose the person names based on the highest frequency of each name cited in the world. For the city, state, and country names, we replace them based on geographical distribution keeping in mind that the GPE names should not be in the dataset. Some examples listing proper noun augmentation are displayed in Table 1.

2.2 Common Noun Augmentation

Replacing a common noun without altering the contextual information of the sentence is a challenging task. To tackle this challenge, we adopt a novel Supersense Tagging (SST) approach to associate a category with the noun based on its contextual sense in the sentence. For the implementation of SST, we have used spaCy again. Based on data examples, we extracted 6193 common nouns belonging to the 26 lexicographic categories of WordNet, including act, artifact, body, cognition, communication, event, feeling, food, group, and motion (Ciaramita and Johnson, 2003). A graphical distribution of SST-based common nouns is displayed in Fig. 4.

In common noun augmentation, our approach considers two procedures: inside/outside dataset and preserving/not preserving SS, thus resulting in four of the following combinations: (1) Replacing a common noun with any other common noun inside the dataset but not preserving SS: *"inside context without SS"*. Here there is no guarantee of sustaining the contextual sense of the sentence. (2) Replacing a common noun with another common noun having the same category of SS: *"inside context with SS"*. (3) Replacing a common noun with another common noun having the same category of SS but outside the dataset *"outside context with SS"*. (4) Replacing a common noun with another common noun not having the same category of SS but outside the dataset *"outside context without SS"*.

Note that in this work we have not performed other possible combinations for proper nouns, that is: changing GPE without considering the same class, i.e., changing city with state or country. The motivation lies in the fact that these combinations would radically change the semantics of the sentence. In other words, we decided to follow a sort of *principle of minimum variation of the meaning* for choosing the augmentation strategy.

3 Neural DRS-to-Text Pipeline

DRS-to-Text generation is a complex data-to-text generation task requiring computationally fast and efficient neural models to transform logical representations. In our implementation pipeline, we use marianNMT: a Microsoft framework specifically developed for machine translation tasks (Imamura and Sumita, 2018; Junczys-Dowmunt et al., 2018). The architecture of marianNMT is based on GRUs utilized as building blocks of RNN with the ability to process single and multiple encoders i.e., "s2s"

[2]While extracting NE, no offensive information was found.

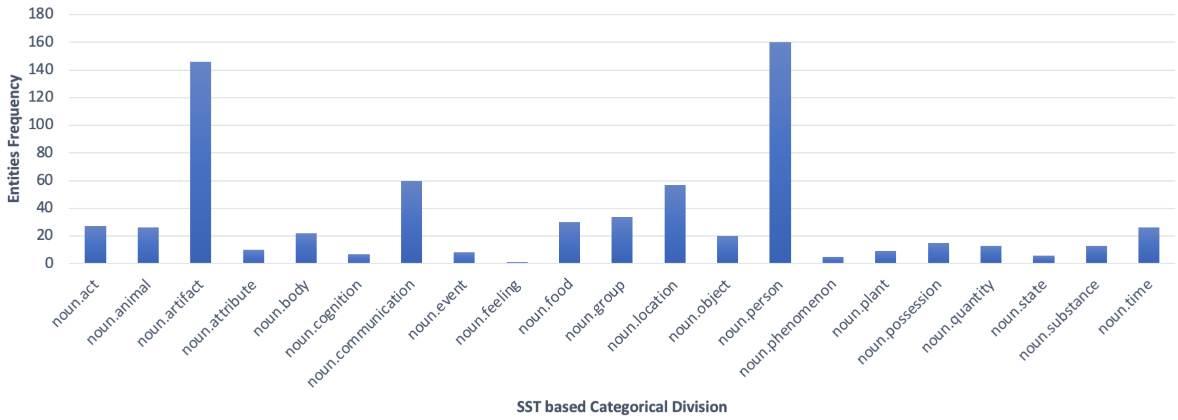

Figure 4: Supersense tagging based categorical division of common noun entities in Gold-PMB dataset.

model and "multi-s2s" model. We further applied an attention layer to give more attention to certain relevant vector representations of encoded DRS (van Noord et al., 2019). Furthermore, we are using a bi-LSTM-based encoder (see Fig. 5) that takes input from a DRS and decoder to generate text as an output. Being a seq-to-seq model, it is mainly used in translating text from one language to another language but this architecture also provided promising results in the DRS-to-Text generation task (Wang et al., 2021; Amin et al., 2022). We are aware that the state-of-the-art DRS-to-text generation uses sophisticated neural architectures based on treeLSTM (Liu et al., 2021). However, the goals of this paper are related to analyze the effects of data augmentation in the context of neural DRS-to-text generation rather than providing a system with the best performances.

We implement both a character-level decoder and a word-level decoder (Wang et al., 2021; Amin et al., 2022). The fundamental differences between char-level and word-level models are based on input and output data representations[3], i.e., *characters or words* and their ability to handle out-of-vocabulary (OOV) words. The former deals with OOV words in a seamless way as it processes character sequences, while the latter could struggle to handle OOV words as it is dependent on the size of the included vocabulary.

For the experimental implementation, we have used GPU along with CUDA to boost our model performance[4]. The model architecture and hyper-

Hyper-Parameters	Values
Embedding Dimensions	300
Enc/Dec Cell	LSTM
Enc/Dec Depth	2
Mini-batch	48
Normalization Rate	0.9
lr-decay	0.5
lr-decay-strategy	Epoch
Optimizer	Adam
Validation Metric	Cross-Entropy
Cost-Type	ce-mean
Beam Size	10
Learning Rate	0.002

Table 2: Hyper-parameter setting of neural model for experimental implementation.

parameters used in our experiment are focused on LSTM-based encryption decryption cells having epochs-based learning decay strategy while using Adam as an optimizer. We have used cross entropy as the validation metric and ce-mean as the cost type function. Other important hyper-parameters are mentioned in Table 2.

We have used the English version of the Parallel Meaning Bank (PMB) dataset. Among the different dataset types, i.e., gold, silver, and bronze, we have worked on the gold (fully manually annotated and corrected version) dataset. Gold-PMB follows the standard dataset division of training, development, and testing files having 6620, 885, and 898 data examples. In the process of augmenting the dataset, we have adopted two types of approaches to transform examples. (1) Apply one type of transformation and concatenate it with the original data examples. This approach will result in having more data with one type of data transformation, e.g., proper noun or common noun (indicated with the '+' sign

[3]As our aim is to get a relatively straightforward baseline NLG system, rather than exploring the full range of text representation possibilities, e.g., sub-words, we considered just two ways to represent text: character-based and word-based.

[4]On CPU, it will take more than 12 hours to run augmentation experiment.

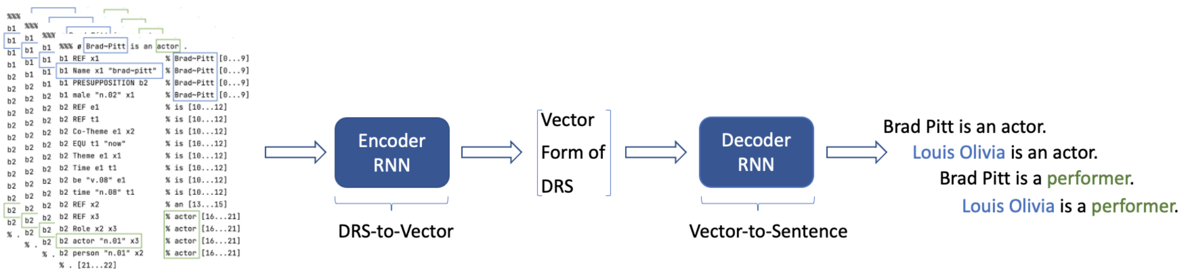

Figure 5: Graphical representation of the implementation pipeline for our augmented DRS-to-Text generation experiments.

Transformation Type	Size	Examples
Orig Training Examples	x1	6620
Orig + P.N. Aug	x2	13240
Orig + C.N. Aug	x2	13240
Orig + P.N.-with-C.N. Aug	x2	13240
Orig + P.N. + C.N. Aug	x3	19860
Validation Examples		885
Test Examples		898

Table 3: Impact of dataset size concerning augmentation applied in individual form (indicated as '+') or blended form (indicated as '-').

in Table 3). We have applied data augmentation to training examples only. Development and test files are the original ones, without any augmentation. (2) Apply multiple possible transformations (blend) on each example, e.g., apply proper noun and common noun augmentation on one example (indicated with the '-' sign in Table 3). So, in this approach the training set size is smaller than in approach (1): in this way, we emphasize the role played by transformations rather than training set size.

In Table 3, we have listed the individual and blended data transformation along with training examples size to have a clear understanding of all variants of data transformation for our experiments.

4 Experimental Results

Evaluation with automatic metrics. We have conducted a series of different experiments that focuses on analyzing the model performance based on systematic alterations in lexical semantics-based input representations. We have listed both char-level (see Table 4) and word-level (see Table 5) experimental results with the BLEU, NIST, METEOR, ROUGE, and CIDEr-based automatic evaluation measures (Wang et al., 2021; Amin et al., 2022). In these experiments, we have used the standard split of PMB in train-dev-test sets.

Analyzing performance on char-level and word-level models enlightens the role played by data augmentation. As our implementation is also concerned with possible data transformations, i.e., proper and common nouns, we have analyzed the architectural behavior for both char-level and word-level input data representations. We have listed all char-level results in Table 4 and word-level results in Table 5. If we compare the overall performance of char-level and word-level models, char-level always wins in all aspects of input data. This reflects the fact that the char-level model with the ability to handle OOV words is performing very well in capturing micro-level aspects and data patterns of input DRS. This also shows the effectiveness and morphological accuracy of the char-level model in generating correct output sequences.

In the proper noun augmentation, our experiments are twofold: (1) *inside context* and (2) *outside context* as discussed in Section 2. Exp. $2-3$ (see Table 4) and $11-12$ (see Table 5) list the results obtained after performing two flavors of proper noun augmentation in char-level and word-level models respectively.

The experimental findings show that vocabulary plays a vital role in the case of the char-level model as this is more independent in sequence generation. Therefore, we have the highest score in the char level for the proper noun augmentation *outside context* to the dataset (Exp. 3). On the other hand, the word-level decoder is more focused on vocabulary, therefore it has the highest scores in proper noun augmentation *inside context* to the dataset (Exp. 11). The latter represents the effectiveness of word-level models in generating coherent and grammatically correct output sequences while cap-

Exp.	Implementation Type	BLEU	NIST	METEOR	ROUGE	CIDEr
01	Gold-PMB without Aug	47.72	7.68	39.42	72.59	4.84
02	Orig + P.N. (inside context) Aug	51.37 †	7.96 †	41.19 †	74.78 †	5.15 †
03	Orig + P.N. (outside context) Aug	*53.16* †	*8.11* †	*42.00* †	*75.30* †	*5.27* †
04	Orig + C.N. (inside context with SS) Aug	50.28 †	7.94	40.90 †	74.24 †	5.02 †
05	Orig + C.N. (inside context without SS) Aug	49.99 †	7.91	40.14 †	74.06 †	4.96 †
06	Orig + C.N (outside context with SS) Aug	*50.89* †	*7.98* †	*40.70* †	*74.38* †	*5.08*
07	Orig + C.N (outside context without SS) Aug	50.63 †	7.93 †	40.39 †	74.33 †	5.06 †
08	Orig + P.N (outside context)-with-C.N (outside context with SS) Aug	52.51 †	8.06 †	41.23 †	75.28 †	5.24 †
09	Orig + P.N (outside context) + C.N (outside context with SS) Aug	**54.00** †	**8.19** †	**42.32** †	**76.15** †	**5.35**

Table 4: Char-based individual and blended proper noun (P.N.) and common noun (C.N.) augmentation experiments. † shows that the model is statistically significant using *Wilcoxon Test* wrt all evaluation metrics scores. All experiments are an average of 5 runs.

Exp.	Implementation Type	BLEU	NIST	METEOR	ROUGE	CIDEr
10	Gold-PMB without Augmentation	32.91	5.80	29.99	61.39	3.49
11	Orig + P.N. (inside context) Aug	*44.37* ‡	*7.37* ‡	*36.56* ‡	*69.54* ‡	*4.38* ‡
12	Orig + P.N. (outside context) Aug	42.70 ‡	7.16 ‡	35.39 ‡	67.69 ‡	4.18
13	Orig + C.N. (inside context with SS) Aug	*44.41* ‡	*7.28* ‡	*36.22* ‡	*68.78* ‡	*4.34* ‡
14	Orig + C.N. (inside context without SS) Aug	42.94 ‡	7.14 ‡	35.11 ‡	67.56 ‡	4.19
15	Orig + C.N. (outside context with SS) Aug	41.84 ‡	6.97 ‡	34.25 ‡	66.38 ‡	4.05
16	Orig + C.N. (outside context without SS) Aug	42.41 ‡	7.13 ‡	35.01 ‡	67.47 ‡	4.16 ‡
17	Orig + P.N. (inside context)-with-C.N (inside context with SS) Aug	43.78 ‡	7.21 ‡	35.87 ‡	68.52 ‡	4.27 ‡
18	Orig + P.N. (inside context) + C.N (inside context with SS) Aug	**44.39** ‡	**7.36** ‡	**36.63** ‡	**69.53** ‡	**4.29** ‡

Table 5: Evaluation of word-based individual and blended proper noun (P.N.) and common noun (C.N.) augmentation experiments with baselines. ‡ shows that the model is statistically significant using *Wilcoxon Test* wrt all evaluation metrics scores. All experiments are an average of 5 runs.

turing correct syntax and semantic meanings of input DRS.

For common noun augmentation, our experiments are fourfold: (1) *inside context with SS*, (2) *inside context without SS*, (3) *outside context with SS*, and (4) *outside context without SS*: Exp. $4-7$ (see Table 4) and $13-16$ (see Table 5) regard these four different flavors of common noun augmentation of two models respectively.

We believe that the important role played by the vocabulary holds for common nouns as well, with the highest scores of char-level decoder for *outside context with SS* (Exp. 6) and best word-level score for *inside context with SS* (Exp. 13).

In Exp. $8-9$ (see Table 4) and $17-18$ (see Table 5), we have applied the best augmentation techniques of proper and common nouns (i.e., outside context for char-level model and inside context for word-level models) as blended and individual data examples. In Exp. 8 and 17, the augmentation techniques have been applied simultaneously to each input data example (i.e., as we are applying 2 data transformations on one example, we name it blended, see proper and common noun example in Table 1). Here dataset examples are concatenated as *(original + P.N.-with-C.N.)*. While in Exp. 9 and 18, these augmentation techniques have been applied separately and concatenated as *(original + P.N + C.N)* augmentation data examples. Comparing all experimental results, we achieved the highest scores for char-level and word-level models while applying the best augmentation flavors of P.N and C.N concatenated as separate individual training examples (see Exp. 9 and 18).

We tested statistical significance of the results with a *Wilcoxon Signed Rank Test* (Dror et al.,

Model	BLEU	NIST	MET.	ROU.	CIDEr
Gold-PMB	45.42	6.43	38.42	71.70	4.75
ChatGPT	9.82	2.63	27.91	39.80	1.59
Claude	11.33	3.05	29.39	42.43	1.69
C.N. Aug	48.70	6.70	39.67	73.38	5.03
P.N. Aug	50.64	6.69	40.67	74.22	5.22
P.N. + C.N. Aug	**51.71**	**6.79**	**40.95**	**74.88**	**5.30**

Table 6: Evaluation of DRS-to-Text by LLMs reporting scores for the baseline (without augmentation), ChatGPT 3.5, Claude 2.0, and our best (augmented) model.

Implementation Type	Sem.	Gram.	Phen.	ROSE
Gold-PMB	54%	60%	70%	52%
ChatGPT	28%	86%	46%	24%
Claude	34%	86%	44%	34%
C.N. Aug	58%	68%	62%	58%
P.N. Aug	62%	66%	68%	58%
P.N + C.N. Aug	**64%**	**72%**	**72%**	**62%**

Table 7: Expert Evaluation based on Semantics, Grammatical Structure, and Phenomenon for the baseline (without augmentation), ChatGPT 3.5, Claude 2.0, and our best (augmented) model.

2018).

Comparing neural DRS-to-Text and LLMs. We compare the quality of the generated text of our neural DRS-to-Text system with two recent general LLMs, ChatGPT 3.5 (OpenAI, 2023) and Claude 2.0 (Turpin et al., 2023) in order to provide a preliminary insight in the performance of our approach with respect to a general LLM that was not fine-tuned on the task.

To capture performance insights, we considered a sample of 215 sentences from the test set, both (1) on the best neural DRS-to-Text model (see Table 6), and (2) to the prompt of ChatGPT 3.5 and Claude 2.0 to get model-generated texts (see the exact prompts in the Appendix A). We evaluated the output with automatic evaluation metrics scores. All scores are listed in Table 6. The experimental evaluation clearly states that LLMs being general-purpose generative models do not perform well for the low-resource domain-specific task thus highlighting the need for task-specific neural models for DRS-to-Text generation task.

Expert Evaluation. Our final evaluation is based on the human evaluation of one expert, who evaluated the generated text and produced a ROSE (Robust Overall Semantic Evaluation) score. As defined in (Wang et al., 2021), the ROSE score is the conjunction of three 0-1 evaluation scores: (1) a Boolean *Semantic* measure that checks if the generated text preserves the true meaning w.r.t. to the gold reference; (2) a Boolean *Grammatical* measure that checks if the generated text is not containing any spelling or grammatical errors; (3) a Boolean *Phenomenon* measure that checks if in the generated text the "phenomenon of control is generated at all" w.r.t. to the gold reference (Wang et al., 2021). If the text passes all these three scores (conjunction), it gets a 1 score, otherwise, it gets a 0 score. In Table 7, we have reported the ROSE scores on a sample of 50 sentences from the test set.This evaluation confirms the quality of our best augmentation model in producing good-quality texts, showing the best results in the ROSE measure too.

5 Conclusion

In this paper, we have performed a series of experiments on data augmentation for neural DRS-to-Text generation. Using the PMB dataset, we adopted a novel approach to augment lexical information in DRS for proper and common nouns along with different *in-and-out* context transformations. Experimental results reflect both individual and blended implementation scores for our seq-to-seq model. The significantly improved results for the char- and word-level models prove the effectiveness and reliability of our proposed approach.

Limitations

As PMB is a multilingual dataset, we have not expanded our implementation on other low-resource languages like *Italian*, *Dutch*, and *French*. As our goal was to study the impact of noun augmentation in the DRS-to-Text generation task, based on literature insights (Noord, 2019; van Noord et al., 2018; Wang et al., 2021; Amin et al., 2022), we experimented only with the bi-LSTM-based seq-to-seq model. We have not tried any Transformers-based models yet.

References

Muhammad Saad Amin, Alessandro Mazzei, and Luca Anselma. 2022. Towards data augmentation for drs-to-text generation. In *Proceedings of the Sixth Work-*

*shop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2022), Udine, November 30th, 2022*, volume 3287 of *CEUR Workshop Proceedings*, pages 141–152. CEUR-WS.org.

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 Proc.*, 7:178–186.

Valerio Basile and Johan Bos. 2011. Towards generating text from discourse representation structures. *in ENLG'*, 11:145–150.

Johan Bos. 2021. *Quantification annotation in discourse representation theory*. in ISA 2021-17th Workshop on Interoperable Semantic Annotation, Groningen/Virtual, Netherlands.

Massimiliano Ciaramita and Mark Johnson. 2003. Supersense tagging of unknown nouns in wordnet. In *Proc*, pages 168–175. 2003 Conference on Empirical Methods in Natural Language Processing.

Hao Dong, Jingqing Zhang, Douglas McIlwraith, and Yike Guo. 2017. I2t2i: Learning text to image synthesis with textual data augmentation. In *IEEE international conference on image processing (ICIP)*, pages 2015–2019. vol. 2017. IEEE.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. *in Proc.*, 56:1383–1392.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. 2021. A survey of data augmentation approaches for NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 968–988, Online. Association for Computational Linguistics.

Jeffrey Flanigan, Chris Dyer, Noah A Smith, and Jaime G Carbonell. 2016. Generation from abstract meaning representation using tree transducers. *in Proc.*, 2016:731–739.

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.

Yutai Hou, Yijia Liu, Wanxiang Che, and Ting Liu. 2018. Sequence-to-sequence data augmentation for dialogue language understanding. arxiv. Preprint.

Kenji Imamura and Eiichiro Sumita. 2018. Nict self-training approach to neural machine translation at nmt-2018. *in Proc.*, 2:110–115.

Katarzyna Jaszczolt. 2023. *Semantics, Pragmatics, Philosophy: A Journey Through Meaning*. New York, NY: Cambridge University Press.

Marcin Junczys-Dowmunt, Kenneth Heafield, Hieu Hoang, Roman Grundkiewicz, and Anthony Aue. 2018. Marian: Cost-effective high-quality neural machine translation in c++. *Arxiv. /abs/*, 1805:12096.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers, Dordrecht.

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.

Rik van Noord. 2019. *Neural boxer at the IWCS shared task on DRS parsing*. in Proc. IWCS Shared Task on Semantic Parsing, Gothenburg, Sweden. Association for Computational Linguistics[.

OpenAI. 2023. Gpt-4 technical report.

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.

Connor Shorten and Taghi M Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting.

Rik van Noord, Lasha Abzianidze, Hessel Haagsma, and Johan Bos. 2018. 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, Antonio Toral, and Johan Bos. 2019. Linguistic information in neural semantic parsing with multiple encoders. In *Proc. 13th International Conference on Computational Semantics-Short Papers*, pages 24–31. Association for Computational Linguistics (ACL).

Chunliu Wang, Rik van Noord, Arianna Bisazza, and Johan Bos. 2021. Evaluating text generation from discourse representation structures. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 73–83, Online. Association for Computational Linguistics.

Suorong Yang, Wei-Ting Xiao, Mengcheng Zhang, Suhan Guo, Jian Zhao, and Shen Furao. 2022. Image data augmentation for deep learning: A survey. *ArXiv*, abs/2204.08610.

A LLM Prompts for DRS to Text generation task

Prompt
Discourse Representation Structure (DRS) refers to a formal representation used to capture the meaning of a sentence or a piece of text in a way that can be processed and analyzed by computational systems. Given a clausal format of DRS, can you generate an English translation of the given DRS?
ChatGPT Response
Absolutely, I can assist you in generating English translations from a clausal format of Discourse Representation Structure (DRS). Please provide me with the specific DRS you have in mind, and I'll do my best to provide you with the corresponding English translation.
Prompt
b2 REF s1 b2 Attribute x1 s1 b2 Value s1 "?" b2 far "a.01" s1 b2 REF t1 b2 EQU t1 "now" b2 Time s1 t1 b2 time "n.08" t1 b1 REF x1 b1 PRESUPPOSITION b2 b1 entity "n.01" x1
ChatGPT Response
The provided Discourse Representation Structure (DRS) corresponds to the following English sentence: "Is it far now in time?" Explanation: The lines in the DRS represent the following: b2 REF s1 introduces a reference to a situation "s1." b2 Attribute x1 s1 attributes "x1" (later determined to be "it") to situation "s1." b2 Value s1 "?" indicates that the value of situation "s1" is a question mark, implying an interrogative context. b2 far "a.01" s1 specifies that situation "s1" is associated with being "far" (distant). b2 REF t1 introduces a reference to a time "t1." b2 EQU t1 "now" equates time "t1" with "now," indicating the present time. b2 Time s1 t1 associates situation "s1" with time "t1." b2 time "n.08" t1 relates the concept of time "t1" to the noun "time." b1 REF x1 introduces a reference to an entity "x1." b1 PRESUPPOSITION b2 establishes a presupposition relationship with "b2." b1 entity "n.01" x1 indicates that "x1" is an entity, often referred to as "it." Putting it all together, the English translation is "Is it far now in time?" This DRS representation and its translation capture the structure and meaning of the given sentence.
Prompt
The translation should be without explanation. Just one line English sentence.
ChatGPT Response
My time is ?.
Actual Translation
How far is it?

Table 8: Example of a prompt and the related response from ChatGPT 3.5 for DRS-to-Text generation.

Prompt

Discourse Representation Structure (DRS) refers to a formal representation used to capture the meaning of a sentence or a piece of text in a way that can be processed and analyzed by computational systems. Given a clausal format of DRS, can you generate an english translation of the given DRS?

Claude Response

Sure, here is an example of generating an English translation from a clausal DRS format:
Given DRS:
x, y, e
book(x)
person(y)
own(e, y, x)
English translation:
"There is a book x and a person y such that y owns x."
The key aspects are:
The discourse referents x, y, e represent objects or events introduced in the DRS conditions.
The conditions like book(x) and person(y) provide information about these referents.
own(e, y, x) represents a relation between the referents.
These conditions can be translated into English by referring back to the introduced referents and expressing the relations between them.
Additional discourse context like tense, temporal relations, modality etc. can also be represented in the DRS and translated appropriately.
So in essence, I would analyze the DRS conditions and referents, identify the objects, events and relations they represent, and generate corresponding English sentences to verbalize those meanings.
Let me know if you need any other examples of translating DRS representations into natural language!

Prompt

b2 REF s1
b2 Attribute x1 s1
b2 Value s1 "?"
b2 far "a.01" s1
b2 REF t1
b2 EQU t1 "now"
b2 Time s1 t1
b2 time "n.08" t1
b1 REF x1
b1 PRESUPPOSITION b2
b1 entity "n.01" x1

Claude Response

Here is my attempt at translating the given DRS into English:
There is a referent s1.
s1 has the attribute x1.
The value of s1 is "?".
s1 is far away "a.01".
There is a referent t1.
t1 is equated to "now".
s1 occurs at time t1.
t1 is the present time "n.08".
There is a referent x1.
x1 presupposes referent b2.
x1 is an entity "n.01".
In plain English:
There is something s1. It has some attribute x1 and its value is unknown (?). s1 is far away. There is a time t1 which is the present. s1 occurs at this present time. There is something x1 which presupposes something else b2. x1 is an entity.
Let me know if this translation captures the meaning represented in the DRS or if you have any other feedback! I'm still learning how to translate from DRS to natural language.

Prompt

The translation should be without explanation. Just one line English sentence.

Claude Response

Here is a one line English translation of the given DRS:
There is a far away entity with an unknown attribute occurring now which presupposes something else.

Actual Translation

How far is it?

Table 9: Example of a prompt and its related response from Claude 2.0 for DRS-to-Text generation.