Machine Translation Pre-training for Data-to-Text Generation - A Case Study in Czech

Mihir Kale and Scott Roy
Google Research
{mihirkale,hsr}@google.com

Abstract

While there is a large body of research studying deep learning methods for text generation from structured data, almost all of it focuses purely on English. In this paper, we study the effectiveness of machine translation based pre-training for data-to-text generation in non-English languages. Since the structured data is generally expressed in English, text generation into other languages involves elements of translation, transliteration and copying - elements already encoded in neural machine translation systems. Moreover, since data-to-text corpora are typically small, this task can benefit greatly from pre-training. We conduct experiments on Czech, a morphologically complex language. Results show that machine translation pre-training lets us train end-to-end models that significantly improve upon unsupervised pre-training and linguistically informed pipelined neural systems, as judged by automatic metrics and human evaluation. We also show that this approach enjoys several desirable properties, including improved performance in low data scenarios and applicability to low resource languages.

1 Introduction

Data-to-Text refers to the process of generating accurate and fluent natural language text from structured data such as tables, lists, graphs etc.(Gatt and Krahmer, 2018) For example, consider Figure 1, in the context of a restaurant booking system. The system must take a meaning representation (MR) as input - in this case represented in the form of a dialogue act (*inform*) and a list of key value pairs related to the restaurant - and generate fluent text that is firmly grounded in the MR.

In this work, we focus on generating text in non-English languages and show that it is possible to significantly reduce this accuracy gap by pre-training fully lexicalized models on an NMT task. For an example motivating the use of NMT, consider Figure 1 once again. In order to generate semantically correct and natural sounding text in Czech (Marathi), a data-to-text model would need to learn the following skills:

Figure 1: Generating text from structured data. Aligned segments from the structured data and natural language have the same color.

- *Translate* the slot value "dinner" to the target language
- *Copy* the phone number correctly
- *Inflect* the restaurant name

In the case of Marathi, which has a different script, there is the additional challenge of *Transliterating* the restaurant name as well.

It is unreasonable to expect neural data-to-text models to learn all these skills, especially since the size of most NLG [1] datasets is quite small. However, modern neural machine translation systems are already fairly adept at translating, transliterating, copying, inflecting etc. Consequently, we hypothesise that the parameters of an NMT model will act as a very strong prior for an NLG model.

[1] While NLG is a broad term, in this paper, we use NLG and data-to-text interchangeably.

Proceedings of The 13th International Conference on Natural Language Generation, pages 91–96,
Dublin, Ireland, 15-18 December, 2020.

2 Related Work

Earlier work on NLG was mainly studied rule-based pipelined methods (Reiter and Dale, 2000; Siddharthan, 2001; Stent et al., 2004), but recent works favor end-to-end neural approaches. Wen et al. (2015) proposed the Semantically Controlled LSTM and were one of the first to show the success of neural networks for this problem, with applications to task-oriented dialogue. Since then, some works have focused on alternative architectures - Liu et al. (2018) generate text by conditioning language models on tables, while Puduppully et al. (2019) propose to explictly model entities present in the structured data. With the advent of BERT (Devlin et al., 2018), the unsupervised pre-training + fine-tuning paradigm has shown to be remarkably effective, leading to improvements in many NLP tasks. While the above works focus on unsupervised pre-training, Siddhant et al. (2019) and Schuster et al. (2018) examine transfer learning via neural machine translation for NLU tasks. Recently, Chi et al. (2019) found multilingual unsupervised pre-training techniques to be effective for cross-lingual language generation tasks like summarization and question generation. Similar to our work, Saleh et al. (2019) used machine translation pre-training in their winning entry to the WNGT 2019 shared task (Hayashi et al., 2019). In this work, we also offer further insights on the usefulness of machine translation by conducting controlled experiments in various settings - limited labeled data, low resource languages, comparison with unsupervised pre-training etc. We also support our findings with human evaluations.

3 Model Architecture

We use the transformer (Vaswani et al., 2017) based encoder-decoder architecture by casting data-to-text as a seq2seq problem, where the structured data is flattened into a plain string consisting of a series of intents and slot key-value pairs. More exotic architectures have been suggested in prior work, but the findings of Dušek et al. (2018) show that simple seq2seq models are competitive alternatives, while being simpler to implement. Secondly, the transformer architecture is state-of-the art for NMT. Thirdly, keeping the pre-train and fine-tune architectures the same allows us to easily transfer knowledge between the two steps by parameter initialization.

4 Pre-train + Fine-tune

Our modeling approach is simple. We first use a parallel corpus to train a sequence-to-sequence transformer based neural machine translation model. Next, we fine-tune this NMT model using a data-to-text corpus for a small number of steps. All the model parameters are updated in the fine-tuning process. In practice, we found that a bidirectional model, which can translate from English to the target language and vice-versa, performed slightly better.

5 Baselines

We compare with the following baselines:
Scratch A baseline where all the parameters are learned from scratch, without any kind of transfer learning. This is a 1-layer Transformer model. Larger models trained from scratch did not improve performance.
Unsupervised pre-training baseline Monolingual data is generally far easier to obtain than bilingual data, which makes unsupervised pre-training techniques more attractive. Interestingly, Wu and Dredze (2019) and Pires et al. (2019) find that pre-training BERT models on a combination of languages can lead to surprisingly effective cross-lingual performance on NLU tasks, without using any parallel data. Of the myriad unsupervised techniques, we choose the span masking objective employed by T5 (Raffel et al., 2019), MASS (Song et al., 2019) etc. for our baseline since it has been shown to outperform other alternatives like BERT. During pre-training, spans of text are masked in the input sentence and fed to the encoder. The decoder must learn to output the masked spans.
TGen is a freely available open-source NLG system based on seq2seq + attention. Dušek and Jurčíček (2019) create a pipelined system consisting of : a TGen based model that outputs delexicalized text, a classifier that ranks the beam search hypotheses and a language model which does the lexicalization by picking the exact surface form. We denote this combined system, consisting of all 3 components as *tgen-sota*. It is also currently the state-of-the-art for the data-to-text corpus that we use for downstream evaluation. Note that the lexicalization step requires access to lexicon data containing all the morphological forms of words and entities. Unlike *tgen-sota*, our proposed model is trained end-to-end to directly generate lexicalized outputs, which is a much harder task. We also

part	Train	Dev	Test
Unique MRs	144	51	53
Corpus size	3,569	781	842

Table 1: Czech NLG dataset statistics. The unique MRs are counted after delexicalizing the slots.

do not rely on any external lexical data.

Its not realistic to assume that every NLG system is first developed for English. As such, our setting does not assume the existence of a similar dataset in English. Therefore, translation based baselines (eg: first running the English model and then translating the output) are not applicable here.

6 Experimental Setup

6.1 Datasets

Pre-training We use the Czech-English parallel corpus provided by the WMT 2019 shared task. The dataset comprises of 57 million translation pairs, automatically mined from the web. In order to facilitate a fair comparison, we use this corpus for our unsupervised pre-training baselines as well. This effectively results in 114 million monolingual sentences, equally split between English and Czech.

NLG We use the recently released Czech Restaurant dataset (Dušek and Jurčíček, 2019). Data related statistics can be found in Table 1. The delexicalized MRs in the test set never appear in the training set. As a result, models must learn to generalize to MRs with unseen slot and intent combinations.

6.2 Training details

For NMT and MASS, we train transformer models with 93M parameters (6 layers, 8 heads, 512 hidden dimensions). They are trained for 1 million steps with Adam optimizer and a batch size of 1024. For NLG, all our models are fine-tuned for 10K steps with a batch size of 32. We do not perform any hyperparameter tuning. Decoding is performed using beam search, with a beam width of 8. All the transformer based models are implemented in the Lingvo framework (Shen et al., 2019) based on Tensorflow (Abadi et al., 2016). The *tgen-lex* baseline is trained using the open-source repository with the exact hyperparameters as used by Dušek and Jurčíček (2019). The best checkpoints are selected based on validation set BLEU score.

6.3 Data pre-processing

Our vocabulary consists of a sentencepiece model with 32,000 tokens (Kudo and Richardson, 2018) shared between English and Czech. It is computed on English and Czech sentences from the pre-training corpus. In order to facilitate a fair comparison, we maintain the same vocabulary across all the transformer based models and baselines. No special rules or pre-processing is done to tokenize the structured data - we simply feed it as a plain string. The input sequence is pre-pended with a task specific token - `[TRANSLATE]` for translation, `[GENERATE]` for NLG. Following Aharoni et al. (2019), we pre-pend a second token to specify the desired output language - `<2en>` for English and `<2cs>` for Czech.

6.4 Metrics

Following prior work (Dušek and Jurčíček, 2019), we use the suite of word-overlap-based automatic metrics from the E2E NLG Challenge [2], supporting BLEU (Papineni et al., 2002), NIST (Doddington, 2002), ROUGE-L (Lin, 2004), METEOR (Lavie and Agarwal, 2007), CIDEr (Vedantam et al., 2015). We also compute a Slot Error Rate (SER) metric to gauge how well the generated text reflects the structured data. We calculate how many of the slot values in the structured data have been mentioned in the generated text. An example is marked as correct only if all the slot-values in the structured data are present in the output [3].

7 Results and Discussion

7.1 Main Results

We report results in Table 2. The *scratch* baseline performs quite poorly, as expected. While unsupervised transfer learning (*mass*) performs better, pre-training via machine translation (*nmt*) gives the best results by large margin. *nmt* brings down the SER to just 1.9, a 20 point gain over *mass*, while improving the BLEU score by 8 points. Similar trends are observed in the other metrics as well. These results give credence to our hypothesis that machine translation can be a strong pre-training objective for data-to-text generation in non-English languages.

[2] https://github.com/tuetschek/e2e-metrics

[3] Note that SER can be reliably computed only for delexicalizable slots. As a result, the binary *kids_allowed* slot is ignored.

model	BLEU ↑	SER ↓	NIST ↑	METEOR ↑	ROUGE-L ↑	CIDEr ↑
tgen-sota †	21.96	2.75	4.77	23.32	42.95	2.18
scratch	11.66	63.18	3.06	15.79	28.27	0.84
mass	17.72	24.82	4.22	21.16	38.94	1.75
nmt	**26.35**	**1.9**	**5.24**	**25.81**	**47.07**	**2.60**

Table 2: Results. ↑ implies higher is better, while ↓ arrow implies lower is better. † We compute SER metrics on outputs provided to us by the authors. The other metrics are taken from the paper (Dušek and Jurčíček, 2019)

Compared to the state-of-the-art pipelined *tgen-sota* system, *nmt* compares favorably, showing improvements on all metrics, including a 4 point improvement in BLEU. Recall that *tgen-sota* involves training 3 separate models (seq2seq for generation, classifier for ranking and language model to pick the correct surface form). In contrast, our approach is simple and end-to-end.

7.2 Human Evaluation

Since automatic metrics have been shown to be inadequate for generation tasks, we also conduct human evaluations on a set of 200 examples randomly sampled from the test set. Concretely, we measure two metrics - accuracy and fluency

Accuracy: Human raters are shown the gold text and the predicted text and are instructed to mark the generated text as accurate if it correctly conveys the meaning of the gold text. This effectively catches errors due to hallucinations, incorrect grounding etc. Each example is rated by 3 raters, and we consider an example to be correct if at least two raters say so.

Fluency: We show the predicted text to raters and ask them how natural and fluent the text sounds on a 1-5 scale, with 5 being the highest score. Again, each example is rated by 3 raters. We average the scores across all the ratings to get the fluency score.

We conduct accuracy and fluency evaluations for our best model (*nmt*), *mass* and *tgen-sota*. Results are reported in table 3. *tgen-sota* produces accurate output, but lags behind *nmt* and *mass* in terms of fluency. *mass* produces fluent output on account of its strong language model but scores low on accuracy. *nmt* on the other hand, gets the highest scores on both metrics - 97.5% for accuracy and 4.83 for fluency.

Overall, automatic and human evaluation results strongly point to the applicability of this approach to real-world NLG systems.

model	accuracy ↑	fluency ↑
nmt	**97.5**	**4.83**
tgen-sota	94.0	4.48
mass	90	4.77

Table 3: Human evaluations for accuracy and fluency

Training Size	Model	BLEU ↑	SER ↓
100	scratch	3.03	78.5
	mass	4.42	78.74
	nmt	**15.45**	**31.82**
1000	scratch	7.37	70.19
	mass	9.80	66.15
	nmt	**21.17**	**4.51**
Full	scratch	11.66	63.18
	mass	17.72	24.82
	nmt	**26.35**	**1.9**

Table 4: Experiments with low-resource NLG

7.3 Low resource NLG

In this section we study the effects of transfer learning when the size of the fine-tuning corpus is small. We create two random subsets from the NLG training data of size 100 and 1000. Results are reported in Table 4. We find that once again, *nmt* offers substantial gains over *mass*. When fine-tuning on 1000 examples, pre-training with NMT is substantially better than fine-tuning *mass* on the *entire* dataset (3.5k examples). Remarkably, with just 100 examples, our model outperforms training from scratch on the entire training set. These results lead us to believe that machine translation based pre-training can lead to substantial cost savings with respect to training data annotation.

7.4 Low-resource machine translation

Our previous experiments use NMT models trained on a fairly large corpus. However, for many languages, the amount of available parallel data can be small. Therefore, to study the impact of the size of bitext corpus, we run experiments in a simulated

Pre-train	Model	BLEU ↑	SER ↓
1.6B	nmt-50m	26.35	1.9
160M	nmt-5m	23.70	1.43
16M	nmt-500k	22.52	12.47
1.6B	mass	17.72	24.82

Table 5: NLG fine-tuning with low-resource NMT. The first column indicates the number of tokens used for pre-training.

low-resource setting. We train machine translation models on 10% (5.7 million examples, medium resource, denoted as *nmt-5m*) and 1% (570K examples, low resource, denoted as *nmt-500k*) of the data and use them for fine-tuning the NLG task.

Next, we fine-tune each of these models on the data-to-text task. From the results in Table 4, we see that while the high resource model performs the best, the medium resource models is not far behind in terms of BLEU. Both the high and medium resource models have a comparable SER. Even the low resource model, pre-trained on just 1% of the translation corpora is significantly better than *mass*, which has been pre-trained on almost 1.6 billion tokens. The results indicate that machine translation based transfer learning can be successfully applied even when the size of parallel corpus is small, and thus holds promise for low-resource languages.

8 Conclusion

In this work we investigated neural machine translation based transfer learning for data-to-text generation in non-English languages. Using Czech as a target language, we showed that such an approach enables us to learn simple, fully lexicalized end-to-end models that outperform competitive baselines. Experimental results suggest several desirable properties including improved sample efficiency, robustness to unseen values and potential applications to low resource languages. At the same time, the approach can also be leveraged to improve performance of delexicalized models.

Studying pre-training on a wide variety of languages, especially those with different scripts, is a direct line of future work. Combining unsupervised and translation based pre-training is also a promising avenue and has already shown good results for NLU tasks (Lample and Conneau, 2019).

Acknowledgments

We would like to thank Markus Freitag for insightful discussions and Ondřej Dušek for providing the *tgen-sota* model outputs.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. *arXiv preprint arXiv:1903.00089*.

Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. 2019. Cross-lingual natural language generation via pre-training. *arXiv preprint arXiv:1909.10481*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc.

Ondřej Dušek and Filip Jurčíček. 2019. Neural generation for czech: Data and baselines. *arXiv preprint arXiv:1910.05298*.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the e2e nlg challenge. *arXiv preprint arXiv:1810.01170*.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Hiroaki Hayashi, Yusuke Oda, Alexandra Birch, Ioannis Konstas, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Katsuhito Sudoh. 2019. Findings of the third workshop on neural generation and translation. *arXiv preprint arXiv:1910.13299*.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

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. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. 2018. Table-to-text generation by structure-aware seq2seq learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Ratish Puduppully, Li Dong, and Mirella Lapata. 2019. Data-to-text generation with content selection and planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6908–6915.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Ehud Reiter and Robert Dale. 2000. *Building natural language generation systems*. Cambridge university press.

Fahimeh Saleh, Alexandre Berard, Ioan Calapodescu, and Laurent Besacier. 2019. Naver labs Europe's systems for the document-level generation and translation task at WNGT 2019. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 273–279, Hong Kong. Association for Computational Linguistics.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2018. Cross-lingual transfer learning for multilingual task oriented dialog. *arXiv preprint arXiv:1810.13327*.

Jonathan Shen, Patrick Nguyen, Yonghui Wu, Zhifeng Chen, Mia X Chen, Ye Jia, Anjuli Kannan, Tara Sainath, Yuan Cao, Chung-Cheng Chiu, et al. 2019. Lingvo: a modular and scalable framework for sequence-to-sequence modeling. *arXiv preprint arXiv:1902.08295*.

Aditya Siddhant, Melvin Johnson, Henry Tsai, Naveen Arivazhagan, Jason Riesa, Ankur Bapna, Orhan Firat, and Karthik Raman. 2019. Evaluating the cross-lingual effectiveness of massively multilingual neural machine translation. *arXiv preprint arXiv:1909.00437*.

Advaith Siddharthan. 2001. Ehud reiter and robert dale. building natural language generation systems. cambridge university press, 2000. *Natural Language Engineering*, 7(3):271–274.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentations in spoken dialog systems. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 79–86, Barcelona, Spain.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of bert. *arXiv preprint arXiv:1904.09077*.