Master Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis Author	Memduh Gokirmak		
Thesis Title	Converting Prose into Poetry with Neural Networks		
Submission Year	2021		
Study Program	Computer Science	**Branch of Study**	Computational Linguistics

Review Author	Ondřej Dušek	**Role**	Opponent
Department	Institute of Formal and Applied Linguistics		

Review Text:

Contents Summary The thesis of Memduh Gokirmak focuses on the topic of building neural models for converting prose text into poetry. The author collected poetry data for three languages and then experimented with two modelling approaches – machine translation (MT), trained on automatically "deversified" texts as the source "language" and original poetry as the target, and generic pretrained language models (LMs) finetuned for prose-to-poetry conversion.

The author shows that MT works relatively well on versifying previously deversified poetry, but fails on generic prose text, probably due to a domain mismatch. The LMs are more successful in producing poetry even from an unrelated domain while treating the prose input more freely. Although the resulting texts do look more poetic than their input, especially regarding word choice, neither approach is able to make the texts rhyme.

The main text of the thesis includes an introduction, 4 numbered chapters, and a conclusion:

- The introduction explains the aim of the thesis and briefly sketches the methods used in in the experiments.
- Chapter 1 is a theory and literature overview, including both a description of the neural architectures used in the thesis and an overview of approaches to automatic poetry generation.
- Chapter 2 describes data collected to train the models (in English, Czech and Turkish) and automatic evaluation metrics used in the experiments.
- Chapter 3 is a detailed description of experiments with multiple variants of MT models. These are neural transformer models trained mainly on deversified poetry as source (prose texts resulting from backtranslating the target poetry – i.e., automatically translating to a pivot language and back to the source), with an alternative denoising autoencoder setting (translating from a few selected keywords into a full poetic text). The basic model is complemented with two additional features: gradient accummulation and rhyming loss. The chapter includes automatic evaluation and author's comments on the quality of the output poetry.
- Similarly, Chapter 4 describes experiments with LMs, where the focus is only on English. The author includes three settings – prompting the LM with a full deversified text or prompting with a few keywords extracted either from the deversified text, or from the target (i.e., similar to the autoencoding MT setting). The model variants are again evaluated and compared to the English MT models.
- The conclusion simply summarizes the results and gives a few ideas on possible extensions.

The appendix of the thesis includes links to the source code, including online demonstration Google Colab notebooks.

Overall Evaluation My overall impression of the thesis is mostly positive, with some caveats. I believe that the topic of the thesis is very interesting and the aims are very challenging. The text gives a reasonable amount of background, I liked the inclusion of pre-neural approaches to poetry generation. The methods chosen in the thesis – be it for data collection, model training, or evaluation – are reasonable and appropriate in the vast majority of

cases. The extent of the experiments performed is adequate for a master's thesis. I really like that the author used three different languages to evaluate his models. The main aims of the thesis were mostly reached – it demonstrates plausible ways to make prose texts more poetic, documented by reasonable evaluation. While the output poetic quality is probably not very useful in practice (especially regarding rhyming patterns), this is not unlike many state-of-the-art works and is expected given the fully data-driven methods used.

The only questionable part of the modelling approach appears to be the rhyming loss described on pg. 32. Here I am not certain this would work, and the results do not give much clue (see Questions). All the models used in the thesis also have an underlying inherent limitation in that they they have no explicit notion of syllables or rhymes. I believe that tokenization in the MT models could be optimized for this (e.g. using syllables as tokens, and using pronunciation tokens instead of orthography). There are, however, other complications (e.g. with homonyms) and it is hard to do with pretrained LMs where the vocabulary is pre-set at pretraining time.

The evaluation focuses mainly on model training and development, and it is heavily based on automatic metrics. This is, however, critically reflected by the author and understandable given that the results are not of a sufficient quality to use them for human evaluation. The author also shows that he inspected the outputs manually; however, a more formal error analysis would have been appropriate.

The text is written in a very comprehensible English and mostly adequate academic style, with a few rather informally phrased spots. The main problem of the text is its ease of reading, especially for a reader unfamiliar with the topic. Some details of the methods and evaluation are not explained with a sufficient level of detail or clarity, and there are some mix-ups and confusions in the related work section (see Detailed Comments and Questions). The usual guide through chapters, present in the introduction of most theses, is omitted in this one, which again makes orientation harder. The text is severely lacking in cross-references – many tables and figures are not referenced at all from the main body of text, and references to other parts of the text typically are not specific (i.e. by section number). The appendix is also not referenced from the main text, although it contains important links to the source codes and online demonstration of the models. While most relevant literature is properly cited and the author's own work can be clearly distinguished from previous works, the links to relevant literature are applied too sparingly – mostly just in one place (mainly in the first chapter). Repeating the reference at all relevant spots where the respective model/tool/phenomenon is mentioned in the text would be more appropriate and more helpful to the reader. Most notable is the almost complete omission of any citations in the introduction. While the relevant works are cited later, it is not very friendly to the reader.

Overall, I believe that the author has clearly showed his ability to peform independent scientific work, but the evaluation looks a little rushed and the resulting text would require significantly more polishing to be easily readable.

Detailed Comments These comments are mainly aimed at the author, to document problems with the readability of the text. I do not expect to address them at the defense (except at the author's express wish). First, comments on specific points in the text:

- Even for the basic concepts referred to in the literature review in Chapter 1, citations (e.g. of textbooks) would have been appropriate.
- The basic description of MT systems and LMs in Sections 1.2–1.4 is confusing and hard to understand without prior knowledge. There are also a few inaccuracies: embeddings have been used for RNNs already, they are not a new concept for transformers; BERT is an encoder-only model – it is only capable of classification, not generation.
- Some concepts in the poetry generation description in Section 1.5 are only referred to by abbreviations – these need to be explained and cited (e.g. FST, EM). The reference to autoencoders on p. 12 is confusing – is it a standard autoencoder, or some different variant? There is an important reference missing – Zhang & Lapata, EMNLP 2014 is probably the first use of RNNs for generating poetry.
- You should be descriptive and objective in the related works review – phrasing such as "this does not make sense to judge very harshly" on p. 10, "are commendable" on p. 12 is not appropriate.
- The use of the CMU pronouncing dictionary would be better described in more detail on pg. 19.
- You should describe how you obtained your hyperparameter values in both Chapter 3 and 4.

- The evaluation on news texts is not described clearly and in enough detail. You should clearly mark in Tables 3.5 and 4.3 that the last column is measured on different data than the remaining columns.
- The problem of training on a single GPU (pg. 22, 25) should explicitly mention that the aim is obtaining larger effective batch sizes, not necessarily just overall speedup.
- The remark on paraphrasing on pg. 27–31 is confusing and unclear, a rephrasing would be needed.
- It is not completely clear which exact GPT-2 variant you used – was it GPT-2 small (dubbed "gpt2" in the Huggingface Transformers library)? In addition, GPT-2 is English only, which could be stated as one more reason to focus on English in Chapter 4.
- Note that the freer handling of the input on the part of LM vs. MT is the probable cause in the drop of BLEU/CHRF. I am not saying that this drop is a bad thing, it is just to be expected.
- Regarding your comment on top of pg. 43, I would say the rhyming metric is actually very accurate. The problem is that the models are mostly not trained to rhyme, so the metric does not help much (apart from showing that the outputs do not rhyme).
- Regarding your comment on assessing meaning preservation on pg. 44, I think that just having human evaluators rate given input-output pairs would be sufficient. There is no need to give them full access to your system.

General comments regarding typesetting:

- All the figures and tables should be linked from the text; the words "Figure", "Table", "Section" etc. should be capitalized in cross-references (when used with a number).
- There are some strange tokens in the text, probably oversights ("mem" as reference on p. 11, "/todomeasure" on p. 18), there is an overflow line (p. 18) and an overflowing footnote (p. 19).
- All numeric values in tables should be properly aligned at the decimal point and use the same number of decimal digits (at the very least in the same column).
- You should be careful about capitalization in the references list (e.g. "rnn" on pg. 45).
- You should cite peer-reviewed versions in preference to arXiv papers (e.g. Bahdanau et al., Cho et al., Devlin et al. have published versions). You can use DBLP or Google Scholar to find these.
- You should always include enough details in citations – Nikolov et al. (2020a) does not include the arXiv ID. In addition, it is probably a duplicate of Nikolov et al. (2020b).

Questions I would like some of the questions to be addressed during the defense, if time permits. First, there are several points which need clarification:

- The "reconstructed rhyme" metric (p. 19) is unclear from your description – can you explain?
- What do you mean by the "verse ending character" on p. 20?
- What scoring do you use on pg. 31 to select the 15 tokens for the source sequence in the autoencoding MT-based setup?
- Why did you use the rhyming loss only for Czech? Why did you multiply it with the training loss – is there a mathematical reasoning behind this? I do not see why it should be implemented this way.
- What is the purpose of Table 3.6? I do not understand why it is included and what it is measuring.
- How do you obtain test set keywords for generating from prose/news (p. 39)?

Further, there are a few general questions on potential extensions of the work:

- What do you think about combining both of your training data sources for MT – autoencoding and deversification?
- Did you consider tracking or using poetic meters in any way? How could you include this feature in your models?
- Would you say that current MT approaches are too high-quality to deversify texts during backtranslation (i.e., the texts remain too poetic)? Could that have been a factor in the generalization of your MT-based poetry generation models?

I recommend that the thesis be defended.

I do not nominate the thesis for a special award.

Prague, 27 August 2021

Signature: