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 **Study branch** Computational Linguistics

Review author	Mgr. Martin Popel, Ph.D. **Role** Supervisor
Department	ÚFAL MFF UK

Review text:

The topic of the thesis is automatic conversion of prose into poetry, i.e. to preserve a meaning of a given input text (to some extent), while using poetic language and rhymes in the output. This contrasts with many existing approaches to automatic generation of poetry with no input or by continuing a poem given its first few words or verses, which is a much easier task. The goal of the thesis is to explore ways how to do such conversion using neural networks with as little supervision as possible, i.e. without using rhyme databases and manually written rules for metres, rhyme schemes etc. Most of the knowledge should be learned automatically from a provided (training) corpus of poems.

The author explores two main approaches: MT-based (using machine translation techniques and synthetically deversified poems) and LM-based (using a pre-trained language model, GPT-2). The MT-based approach is implemented for English, Turkish and Czech; the LM-based approach only for English.

The text is clearly structured into 4 chapters, written in English, which is mostly easy to read, but sometimes too informal, with occasional typos and typographical errors.

Chapter 1 (Background) covers all necessary topics and shows the author is familiar with them, but the description is not well suited as an introduction for readers not familiar with the given topics. I appreciate the number of related works covered in Chapter 1.5.

Chapter 2 describes the data resources and their preprocessing. The "preprocessing" involves also a development of high-quality English-to-Turkish and Turkish-to-English MT systems trained with backtranslation, which are about 6 BLEU points better than the best systems reported at `http://matrix.statmt.org` for WMT17. Such task was difficult enough to constitute a separate master thesis topic, but it is described only very briefly in the present thesis, so it cannot be regarded as such advantage.

Chapter 3 describes the MT-based experiments. A baseline training is improved slightly with gradient accumulation, but the overall generated poetry quality (meaning and rhymes) is still unsatisfactory, as reflected by the author in an insightful analysis. The analysis revealed two possible sources of problems: the models overfit to the domain of "deversified poems" and they do not promote generating rhymes in any special way. Two experiments are described aiming to solve the two problems: denoising autoencoder and rhyme-augmented loss. I consider these experiments well motivated, but neither of the experiments succeeded and the error analysis is minimal in this case. Also, the experiments are not described with enough details. For example regarding the autoencoder: How exactly are the keywords selected (why not TF-IDF as in Section 4.2)? Is the problem already in the keyword selection or in further steps? What are the learning curves? Maybe it is just a problem of diverged training (so a longer warmup or curriculum learning with shorter sequences first could help). Similarly with the rhyme-augmented loss: How is the derivative of the loss (needed for backpropagation) computed? Why are the two losses multiplied (instead of adding a logarithm of the rhyming loss to the log-likelihood loss)?

Chapter 4 briefly describes three LM-based experiments: using the whole deversified poem (as an input primer/prompt for GPT-2), using five keywords extracted from the deversified poem and from the original poem. These experiments were more successful than the MT-based ones, when focusing on the amount of rhymes generated in the prose test set and also when focusing on the subjective poetic form. Based on the few examples in the thesis I think the meaning has departed further from the input text, especially in the keywords-based experiments, but this is not necessarily a disadvantage.

Overall, I am satisfied with the thesis: the goal was ambitious and it was partially achieved. The author has proven his ability to perform independent scientific work. That said, there were several topics which would benefit from more experiments and analyses (as described above) and also the text could be improved (as described below). Perhaps the main weakness of the thesis is that there were not enough attempts at generating rhymes: e.g. trying syllable-based tokens or an auxiliary task that would help to encode phonetic (and prosodic) properties in the subword embeddings. Also, the meter and rhyme scheme of the generated verses were ignored within both training and evaluation.

Detailed comments:

No answers on these comments are expected during the defense. The goal is to illustrate some of the factual errors, usually easy to fix, but making the thesis text more difficult to understand.

- page 5: Residual connections do not go from a higher layer *back into the network as input.*
- page 9: BERT is not an encoder-decoder model. It is an encoder-only model.
- page 17: *The proportion of the lines in a rhyme* → *The proportion of the lines in a poem*
- page 19: *If it is a consonant, then we accept the rhyme. If it is a vowel, there must be another vowel* → *If it is a vowel, then we accept the rhyme. If it is a consonant, there must be another vowel*
- page 25: *We train these models with 8 GPUs.* Really? Or with a single GPU and multiplying the effective batch size 8 times using gradient accumulation? Also, gradient accumulation is usually slower than the baseline training when examples per hour are considered, so the *constraints on time* should be clarified (less examples are seen in a given time, but a better model is obtained, corresponding possibly to a much longer baseline training, as could have been seen by comparing Figures 3.1 and 3.7, if they showed also the training time).
- page 33: Figure 3.14: *Rhyme scores* → *BLEU and CHRF scores*
- It would be nice to provide more detailed evaluation of the LM-based experiments. For example, focusing on the differences between keywords and deversified keywords: what is the percentage of keywords preserved in the output in these two experiments and in the keywords training data?
- Table 4.1 (and 4.2, 4.4, 4.5) should show also the keywords extracted from the input (and it could highlight the keywords if present in the output).

I recommend the thesis to be defended.

I do not nominate the thesis for a special award.

Zubří, 31 August 2021

Signature: