Character-Level Poetry Generation for Macedonian Language

***Abstract*—In this work I will present a method to perform a character-level poetry text generation in Macedonian language. The method uses a multi-layer recurrent neural network with LSTM. The approach that we are going to use is a character-level approach, where the model will do a next character prediction instead of a next word prediction, which is a more conventional approach.**

I. Introduction

Text Generation is a type of Language Modelling problem, which is the core problem for a number of natural language processing tasks such as speech to text, conversational system, and text summarization. A trained language model learns the likelihood of occurrence of a word based on the previous sequence of words used in the text. Language models can be operated at character level, n-gram level, sentence level or even paragraph level. In this project, I will be presenting a character level text generation method that uses a recurrent neural network that can generate poetry in Macedonian language.

Poetry generation is an interesting problem to tackle because it is a field in which humans are considered to be far more superior than the computers. We have seen in numerous examples how computers are as good as human or even better sometimes in mechanical tasks but they usually struggle when creativity comes to play. In this project we will tackle that problem and see how far we can get in achieving a human level creativity.

The reason for choosing it to be in Macedonian language is because something like this hasn't been done with this language in particular..

II. Data

In order for the model to generate poetry, it needs a poetry type of data.

Considering the fact that it is a small country and the number of resources is fairly limited, gathering enough data was a challenge.

I have gathered data from many online available poems from various Macedonian authors. The amount of data was way smaller than the desired one so I got some additional data from different forums where people posted their poems.

Since the amount of data was still not enough, I have decided to also add song lyrics from various artists to the dataset.

At the end, I ended up with **104584** lines of text, which consisted of **466189** words, from which, **67962** were unique ones.

For the training I have used 90% of the data and the rest 10% were used in the validation process.

III. Methods

For the purpose of this task, I have decided to use a multi-layer recurrent neural network with LSTM cell.

The architecture of the network is as follows:

- number of layers: 3
- hidden size: 768
- dropout: 0.5

The architecture has an LSTM and a fully connected linear layer. The network consists of 3 layers and a hidden size of 768. For the regularization part, I have decided to use a dropout with 0.5 probability.

For the training part, I have used the following parameters:

- optimizer: Adam
- loss: cross entropy
- learning rate: 0.001
- batch size: 64
- epochs: 40
- sequence length: 150

The training was done on a GPU and it took about an hour to complete, which was reasonable, considering that the size of the data is not that big. I have used an Adam optimizer with 0.001 learning rate and AMSGrad [9] enabled, which helps with the exploding gradient issue. For the objective function, I have used a cross entropy loss. The length of the sequences was limited to 150 characters. In order to speed up the training, I have processed the data in batches of size 64. The training took 40 epochs and 9600 iterations.

The text generation task can be approached in different ways. It can be done on a word level, n-gram level, character level, etc. I have decided to use a character level approach because that way the vocabulary size is way smaller. It just contains the characters in the alphabet and the punctuation. Also the character level approach does not require word2vec embeddings and one-hot encoding is sufficient for the task. The total size of the vocabulary is 186, which we can see is far smaller than what we would have had if we used a word-level approach. The downside of this approach though is that in requires more data and the training time takes longer.

IV. Experiments and results

As mentioned in the previous section, the model was trained for 40 epochs.

During the training, the data was split into 90% training data and 10% validation data.

On the Figure 1 below we can see the values of the training and validation loss.

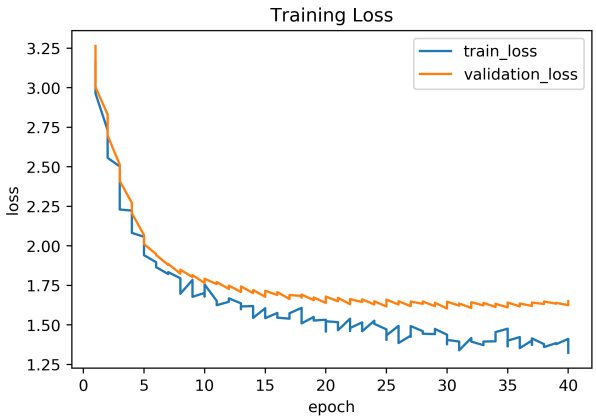

Figure 1. Loss for training and validation data

From the Figure 1 we can see that after 15 epochs, the validation loss starts becoming flat, while the training loss keep decreasing slowly.

The final training loss is: **1.3240** and the final validation loss is: **1.6501**.

For the evaluation purposes, I have generated 1000 characters and counted the number of correctly generated words and divided it by the total number of generated words.

Since the model is doing a character level generation, it is expected to have some misspelled words.

The evaluation results are as follows:

Number of correct words: 187, number of incorrect words 6, accuracy: 0.9689

On the Figure 2 we can see a generated sample.

Since the reader probably isn't familiar with the language, I have also used Google Translate to translate the text to English language. We can see the translation on the Figure 3.

From the translation, we can see that the generated text looks like a poetry text but still there is room for improvement, at least on the grammar side, which I assume that can be fixed with a bigger dataset.

V. Conclusion

In this report, I have presented a character-level approach for poetry generation for Macedonian language, using a multi-layer recurrent neural network with LSTM cell.

The results are relatively good as we saw that most of the words were correctly spelled.

Unfortunately there is no way to evaluate how good the poetry is since that is a subjective matter and there is no universal metric for that kind of task.

полетај го просторот.
И те зема со мене на тебе,
каде е мојата средба.
Стоевме за тебе не сме по него
и носат секилности,
се невини, некои да ми се споредуваат.

Како проклети сончеви погледи непривезени
оваа коска на нив да се однесеме,
а јас не се сеќавам на себе

Толку пријатно небо,
по небото светлина,
пред свилена светлина,
вечна наша верби и треперливи изгрејсонца
сега веќе не е важно
кога сум само толку далечно
и да се потплетам во темнината.

Не ми спие, туку сум и не знам
и ти ми продолжи во ноќта.

Празнина со насмевка во мојата средба
и столб одмината
и пак да ги скриеш.
Ти ме сакаше,?
со твојот сјај негов
и секоја вечер пееше
со некоја друга.
Имам сила да те повредам
исчезнуваш и сега не се раѓам.
Немам приказна.
Но не плачи.
Но немам последен пат,
не знам како е само,
ти, ти се, таа е твојот свет.
Тебе во себе нема небо,
сите сме имање, сонце да не продолжиме
и не е плач,
но сега сум јас со себе.

На мисла ни солзи им ги виде,
низ своите прсти во некако меѓу мене
и солз

Figure 2. Generated text sample

We also saw that the lack of data was an issue for this task and with more data, the results could be significantly improved.

Although we saw that computers are getting pretty good in this kind of tasks, there is still a long way to for the computers to achieve a human-level grammar and creativity.

```
Flee the space.
And he takes you with me on you,
where is my meeting.
We were not after you
and carry axioms,
are innocent, some to compare to me.

Like cursed sunny views untouched
let us take this bone to them,
and I do not remember myself

Such a pleasant sky,
the sky is light,
before silken light,
eternal our faith and trembling sunrise
now it does not matter anymore
when I'm just as distant
and to roll into the darkness.

She does not sleep, but I do not know
and you continued in the night.

Crying with a smile at my meeting
and pillar bye
and hide them again.
Did you love me ,?
with your splendor
and every night he sang
with another.
I have the strength to hurt you
You are disappearing and now I'm not born.
I have no story.
But do not cry.
But I do not have the last time,
I do not know how it is,
you are, it is your world.
There is no heaven in you,
we are all having, let's not keep the sun
and not crying,
but now I am with myself.

Thought to them he saw tears,
through his fingers in some way between me
and tears
```

Figure 3. English translation of the generated text

VI. Future work

As we saw in the section *IV*, the model still needs improvement. One important thing that can be done is to get more data.

Also now exist better approaches for this kind of tasks that don't rely on recurrent neural networks but use attention mechanism instead. It will be good to try such Transformer architecture [10] and see how well it will do in comparison to this approach.

Another thing that can be tried is to use a pre-trained model such as BERT [1] or OpenAI GPT [8] and then fine-tune it for our specific task.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *Bert: Pre-training of deep bidirectional transformers for language understanding*, arXiv preprint arXiv:1810.04805, (2018).
[2] V. Jakimovski and M. Makeska, *Zbirka poezija*, (2016).
[3] A. Lai, *Writing like shakespeare with machine learning in pytorch*, (2019).
[4] V. Lozanov and A. Kovacevski, *Koco racin. beli mugri*, (2006).
[5] LyricsTranslate, *Lyrics translate*.
[6] MakedonskiJazik, *Makedonski jazik*.
[7] pelister.org, *Blaze koneski. poezija*.
[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, *Language models are unsupervised multitask learners*, (2019).
[9] S. J. Reddi, S. Kale, and S. Kumar, *On the convergence of adam and beyond*, in International Conference on Learning Representations, 2018.
[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is all you need*, 2017.
[11] Wikibooks, *Makedonski narodni pesni*, (2018).
[12] Wikisource, *Motorni pesni*, (2018).

VII. Appendix

Below is a link to the Github repository where the code for this project is stored:

Github repository link