Using RuGPT3-XL Model for RuNormAS competition

Anton Emelyanov\textsuperscript{1,2} Oleh Shliazhko\textsuperscript{1} Nadezhda Katricheva\textsuperscript{1} Tatiana Shavrina\textsuperscript{1,3,4}

\texttt{login-const@mail.ru, olehshliazhko@gmail.com}
\texttt{n.katricheva@gmail.com, rybolos@gmail.com}

\textsuperscript{1}SberDevices, Sberbank, Moscow, Russia
\textsuperscript{2}Moscow Institute of Physics and Technology, Moscow, Russia
\textsuperscript{3}National Research University Higher School of Economics, Moscow, Russia
\textsuperscript{4}ANO «AI Research Institute», Moscow, Russia

Abstract

The paper presents a fine-tuning methodology of the RuGPT3-XL (Generative Pretrained Transformer-3 for Russian) language model for the normalization of text spans task. The solution is presented in a competition for two tasks: Normalization of Named Entities (Named entities) and Normalization of a wider class of text spans, including the normalization of different parts of speech (Generic spans).

The best solution has achieved 0.9645 accuracy on the Generic spans task and 0.9575 on the Named entities task. The presented solutions are in the public domain at https://github.com/RussianNLP/RuNormAS-solution

Keywords: text normalization, text generation, evaluation track, ruGPT-3, generative pretrained transformer

DOI: 10.28995/2075-7182-2021-20-204-212

1 Introduction

The task of normalization is indispensable in Natural Language Processing (NLP) because it allows both to obtain a connection between the wordforms of the same paradigm and to reduce vocabulary size while preserving lexical meaning. Text classification, clusterization, topic modeling, style detection, and many more NLP tasks depend on normalization as a basic stage in the text processing pipeline. Regarding an isolating, fusional or agglutinative morphology type, normalization comes in two basic
wordform procedures: stemming or lemmatization. As a more simplistic approach, stemming only chops word endings from the stem, and thus it often attributes the same stem to cognates or different stems to the same lexeme. Lemmatization, in contrast, aims to bring tokens to lexemes. There are other types of normalization, too, such as expanding contractions and abbreviations, but in this paper, we understand normalization differently due to the entities it is applied to. To normalize a named entity or a phrase means to reduce it to its so-called «initial» form representing the semantic core which stays the same no matter what inflections its constituent parts may adopt to provide the syntactic integrity of a sentence. Most of the time it includes lemmatization, like in the case of the named entity (and noun phrase) “группы компаний ЛУКОЙЛ”, which is normalized into “группа компаний ЛУКОЙЛ”: the head of the noun phrase, “группа компаний ЛУКОЙЛ”, becomes a lexeme after normalization. Proper nouns as parts of named entities can be normalized and plural at the same time, like in the case of “Сердца России”, and there are other proper nouns which remain inflected and should not be changed through normalization.

Normalization methods include the usage of lexical databases, where word forms are linked to their lexemes. The result improves if part-of-speech (POS) tags are attributed to the word forms. Another common approach is rule-based, and of course words, as well as named entities and phrases, can be normalized using Neural Networks (NNs) (not only through using NNs for POS-tagging).

This paper is structured as follows: in section 2, we present the already existing research works related to the topic under discussion; section 3 gives a general overview of the competition; section 4 is devoted to our solution of the RuNormAS competition; section 5 provides error analysis, and the paper is concluded in section 6.

2 Previous Work

The English language was traditionally the first to undergo normalization algorithms, in particular, became the object for the first stemmer algorithm [1]. The analytical morphological structure was the best suited for this type of algorithms (for example, [2]), which, together with the growing needs of information retrieval, pushed their development — this happened, in particular, in the works [3], [4]. Nevertheless, it was the fusional and agglutinative languages, with their more productive morphology, that pushed normalization technologies to new levels and made them a subject of competition. Thus, the CONLL competition held in 2016-2018 [5,6, 7] set the task of complete grammatical annotation, from raw text to syntax, which comprised lemmatization for 103 languages, including "surprise languages" in the private test set. For the Russian language, the quality of word inflexion in context achieved 94.4% accuracy.

As for the Russian language separately, normalization technologies are also actively developing for it as a language with a developed morphology. The needs for information retrieval [9] prompted the use of the rich heritage of morphological description [8].

In 2010, for the first time, a shared task was held for automatic Russian part-of-speech tagging, lemmatization, and morphological analysis, including the subtask of annotating rare words [10]. The participants achieved 98.1% accuracy on lemmatization, the test set being not very large. At the MorphoRuEval-2017 shared task [11], a 96.91% accuracy score in lemmatization was achieved on a balanced set of data from various sources (news, social networks, fiction, etc.). And in the GramEval-2020 shared task [12] the track became even more complicated since data from social media, poetry and historical texts of the 17th century were added to the test sample: the best overall lemmatization score being 98% on fiction texts, 98.2% on the news, 95.3% on poetry, 96% on social media, 93% on wiki and 78.3% on historical texts. It became manifest that it is technically possible for the Russian language to pose more complex challenges, especially for notoriously "difficult-to-process" groups of words and lexical categories.

3 Dialog Evaluation 2021 Track

Within the framework of the RuNormAS (Russian Normalization of Annotated Spans) competition [13], the normalization problem is proposed — bringing a part of the text (a named entity, a phrase) to its
normal (initial) form. The main part of the task is to correctly normalize the words from the group that
need normalization without changing the other ones (dependent, etc.) while using the given context to
the benefit of this task. The latter is especially important since the initial form for many words can be
determined only in context — for example, the the word "Иванова", depending on the surrounding
context, can have either the normal form "Иванова" or "Иванов".

The competition offers two tracks:
1. Normalizing Named Entities
2. Normalization of a wider class of text spans, including the normalization of different parts of speech.

The data for the first track were collected from the articles of the «ВЗГЛЯД» newspaper, for the second
one — from the documents of the Ministry of Economic Development. Both samples were labeled
manually.

The quality metric for the task is the percentage of exact matches between the normalization result
and the reference.

3.1 Dataset
Both tasks have the same data format. The text_and_ann folder contains files with texts (.txt)
and files with span markup (.ann). In the file with the markup, the indices of the beginning and end
of the entity in the text are written on each line. If the entity has breaks, then one line is written with
the start and end indices for each chunk (and the chunks may be unordered). For example, if an entity
has two breaking chunks, then the annotations on the corresponding line will contain start1 end1
start2 end2 or start2 end2 start1 end1. In the folder norm, on each line, there is the
result of normalization of the corresponding span. The match is made by the filename up to a dot. Also,
for best model additional data was used. We add the "lenta news" dataset to the train data. This is a
corpus of Russian News for the year 2019. That corpus was annotated automatically and is a part of
Taiga corpus [14].

4 Approach
4.1 Baseline
The competition presents a baseline obtained using normalization tools from the Natasha library¹. This
solution is completely rule-based.

4.2 Neural Language modeling
The idea of finetuning a pretrained Language Model (LM) is at the core of our approach. All the exper-
iments were carried out using RuGPT3XL². The main difference is connected with data preparation for
the RuGPT3XL LM finetuning procedure and model inference strategy. We do not separate data for two
tasks and train one model at each approach on the whole set of train data.

The main algorithm for making predictions consists of three steps (all of them are described below):
1. Prepare data for LM using one of data preparation approaches;
2. Make predictions with LM using one of the inference strategies;
3. Apply the post-processing pipeline;

Each approach differs from the other one only in a specific template for generation, which is fed to the
input of the LM. We tested the following approaches of data preparation (for the first step):
1. Model0 — only left context LM;
2. Model1 — only left context LM with <start> special token;
3. Model2 — left and right contexts LM with <start> and <end> special tokens;
4. Model3 — left and right contexts LM with <start> and <end> special tokens and additional
   training data;

For each approach, we apply two inference strategies:

¹https://github.com/natasha/natasha
²https://huggingface.co/sberbank-ai/rugpt3xl
• **“argmax” inference strategy** is the decoding strategy of LM. We select the next token by applying ‘argmax’ operation over probability distribution that is produced by LM on each decoding step.

• **“beam search” inference strategy** is the standard beam search algorithm with the number of beams equal to 10.

For each approach, we apply the same post-processing pipeline.

### 4.2.1 Post-processing pipeline

The post-processing pipeline should correct errors that occur while generating with LM (after the second step). We have categorized the errors as follows:

1. extra special tokens — model generates extra special tokens that should be removed;
2. letters case errors — model generates words in different cases;
3. extra or removed punctuation — model generate additional punctuation marks or remove some punctuation;
4. different word count in annotation and prediction;
5. symbol intersection error — this error occurs if the following condition is met:
   \[
   \| \text{set(\text{annotation})} \cap \text{set(\text{generated})} \| < 0.6
   \]
   here, the annotation and generation are strings; the 0.6 parameter is selected with some greed search on a subsample of the training data.

For steps 4-5 of this pipeline, we get prediction from the baseline model if errors occurred. Other steps of post-processing are also implemented in our repository.

### 4.2.2 Model0 — only left context LM

For each line in files with span markup (.ann), we find a substring in the text that should be normalized. For example, we have a text in the file of the test set with the name "723362" at Named Entities task:

"Пушков назвал выход Греции из еврозоны «сильнейшим ударом по ЕС» Греция — ключ к будущему ЕС, ее выход из еврозоны станет сильнейшим ударом по ЕС за всю его историю, заявил глава комитета Госдумы по международным делам Алексей Пушков. «Что бы ни говорили в Берлине, выход Греции из еврозоны станет сильнейшим ударом по ЕС за всю его историю. Сегодня Греция — ключ к будущему ЕС», — написал Пушков в Twitter. Накануне пресс-секретарь Еврокомиссии (ЕК), отвечая на вопросы о Греции, заявил, что участие стран в еврозоне согласно законодательству Евросоюза не подлежит отмене. Ранее в издании Der Spiegel сообщалось, что канцлер ФРГ Ангела Меркель и министр финансов страны Вольфганг Шойбле допускают выход Греции из еврозоны при необходимости. Позже заместитель официального представителя кабмина ФРГ Георг Штрайтер заявил, что позиция Германии по вопросу выхода Греции из еврозоны не изменилась. не изменилась".

For the 12th line, in the annotation file we extract the subtext that needs to be normalized: “Вольфганг Шойбле”. For training LM, we construct a training record with the help of the following template:

\[
<s>{\text{left\_context}}}{\text{to\_norm}}<\text{answer}>{\text{norm}}</s>

Here the \(<s>\) token denotes beginning of text; \(\text{left\_context}\) denotes all text before subtext that should be normalized \((\text{to\_norm})\); the \(<\text{answer}>\) token separates input text prefix and answer that LM should learn; \(\text{norm}\) is the normalized text; and \(</s>\) token denotes the end of text. For our previous example, we have the following training record:

\(<s>\text{Пушков назвал выход Греции из еврозоны “сильнейшим ударом по ЕС” Греция — ключ к будущему ЕС, ее выход из еврозоны станет сильнейшим ударом по ЕС за всю его историю, заявил глава комитета Госдумы по международным делам Алексей Пушков. “Что бы ни говорили в Берлине, выход Греции из еврозоны станет сильнейшим}\</s>\)
ударом по ЕС за всю его историю. Сегодня Греция — ключ к будущему ЕС», — написал Пушков в Twitter. Накануне пресс-секретарь Еврокомиссии (ЕК), отвечая на вопросы о Греции, заявил, что участие стран в еврозоне согласно законодательству Евросоюза не подлежит отмене. Ранее в издании Der Spiegel сообщалось, что канцлер ФРГ Ангела Меркель и министр финансов страны Вольфганг Шойбле не подлежит отмене. Ранее в издании Der Spiegel сообщалось, что канцлер ФРГ Ангела Меркель и министр финансов страны Вольфганг Шойбле в Twitter. Накануне пресс-секретарь Еврокомиссии (ЕК), отвечая на вопросы о Греции, заявил, что участие стран в еврозоне согласно законодательству Евросоюза не подлежит отмене. Ранее в издании Der Spiegel сообщалось, что канцлер ФРГ Ангела Меркель и министр финансов страны Вольфганг Шойбле написал Пушков в Twitter. Накануне пресс-секретарь Еврокомиссии (ЕК), отвечая на вопросы о Греции, заявил, что участие стран в еврозоне согласно законодательству Евросоюза не подлежит отмене. Ранее в издании Der Spiegel сообщалось, что канцлер ФРГ Ангела Меркель и министр финансов страны Вольфганг Шойбле написал Пушков в Twitter. Накануне пресс-секретарь Еврокомиссии (ЕК), отвечая на вопросы о Греции, заявил, что участие стран в еврозоне согласно законодательству Евросоюза не подлежит отмене. Ранее в издании Der Spiegel сообщалось, что канцлер ФРГ Ангела Меркель и министр финансов страны Вольфганг Шойбле написал Пушков в Twitter. Накануне пресс-секретарь Еврокомиссии (ЕК), отвечая на вопросы о Греции, заявил, что участие стран в еврозоне согласно законодательству Евросоюза не подлежит отмене. Ранее в издании Der Spiegel сообщалось, что канцлер ФРГ Ангела Меркель и министр финансов страны Вольфганг Шойбле написал Пушков в Twitter. Накануне пресс-секретарь Еврокомиссии (ЕК), отвечая на вопросы о Греции, заявил, что участие стран в еврозоне согласно законодательству Евросоюза не подлежит отмене. Ранее в издании Der Spiegel сообщалось, что канцлер ФРГ Ангела Меркель и министр финансов страны Вольфганг Шойбле написал Пушков в Twitter. Накануне пресс-секретарь Еврокомиссии (ЕК), отвечая на вопросы о Греции, заявил, что участие стран в еврозоне согласно законодательству Евросоюза не подлежит отмене. Ранее в издании Der Spiegel сообщалось, что канцлер ФРГ Ангела Меркель и министр финансов страны Вольфганг Шойбле.
Table 1: Evaluation results

<table>
<thead>
<tr>
<th>Model name</th>
<th>Generic spans</th>
<th>Named entities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model0 + argmax + not_fixed</td>
<td>0.6953</td>
<td>0.7513</td>
</tr>
<tr>
<td>Model0 + argmax</td>
<td>0.7507</td>
<td>0.7891</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.7732</td>
<td>0.8881</td>
</tr>
<tr>
<td>Model0 + beam search</td>
<td>0.8454</td>
<td>0.8828</td>
</tr>
<tr>
<td>Model1 + argmax</td>
<td>0.9059</td>
<td>0.9306</td>
</tr>
<tr>
<td>Model1 + beam search</td>
<td>0.9483</td>
<td>0.9455</td>
</tr>
<tr>
<td>Model2 + beam search</td>
<td>0.9592</td>
<td>0.9570</td>
</tr>
<tr>
<td>Model3 + beam search + not_fixed</td>
<td>0.9636</td>
<td>0.9522</td>
</tr>
<tr>
<td>Model3 + beam search</td>
<td><strong>0.9645</strong></td>
<td><strong>0.9575</strong></td>
</tr>
</tbody>
</table>

4.2.5 Model3 — left and right contexts LM with the `<start>` and `<end>` special tokens and additional training data

The main difference from the previous approach is using additional training data. We add the “lenta news“ corpus with normalization markup and finetune the model on this corpus joint with the training data. After that, we finetune the model only on the training data. The data for LM finetuning was prepared as described in the previous section.

4.2.6 Training details

Each model was trained on 16 GPU with distributed training for around 12 hours. We use the Adam optimizer from [18] with the decoupled weight decay regularization 1e-2 [19]. We use a constant learning rate, 0.000015 on 20000 train iterations with fp16 precision and deepspeed code optimizations[20]. The final perplexity on all models is around 1.0002-1.0005.

5 Error Analysis and Results

5.1 Results

The results of our experiments on test set are presented in Table 1. The best result (Accuracy Generic spans = 0.9645 and Accuracy Named entities = 0.9575) was obtained for “Model3 — left and right contexts LM with `<start>` and `<end>` special tokens and additional training data“ approach with the beam search inference strategy.
The fourth approach “Model3 — left and right contexts LM with <start> and <end> special tokens and additional training data” with the beam search inference strategy obtains the best accuracy for the RuGPT3XL model in this competition. Also, we can see the difference provided by the post-processing pipeline on “Model0” and “Model3”. For the last model, the impact is minor because the LM model has very strong results and sees more data.

5.2 Error analysis

5.2.1 Evaluation errors

Here, we describe errors that are connected with the incorrect data in the evaluation set and markup. We have categorized errors into the following classes:

1. word count errors — these errors denote different counts of words in gold prediction and annotation. For example: “Костромская область” and “область”, here our best model predicted “области”.

2. titled errors — these errors denote difference between word cases in gold prediction and annotation. For example: “Генпрокуратура Украины” and “генпрокуратура Украины”, here our best model predicted “генпрокуратура Украины”.

3. symbol errors — these errors denote the difference between some symbols in gold prediction and annotation. For example: “город Антрацит” and “город Антрацит”, here our best model predicted “город Антрацит”.

4. punctuation errors — these errors denote the difference between the punctuation in gold prediction and annotation. For example: “ООО «Первая топливная компания»” and “ООО Первая топливная компания»”, here our best model predicted “ООО Первая топливная компания»”.

5. word start errors — these errors denote the difference between the starting symbols in gold prediction and annotation. For example: “расти” and “будет расти”, here our best model predicted “будет расти”. Also these errors denote encoding mismatch or truncated markup.

If these errors are not taken into account, then the model obtained 0.9767 accuracy on the Generic spans task and 0.9810 accuracy on the Named entities task.

5.2.2 Model errors

Here we describe model errors. We divide the errors into categories:

1. word count errors. An example of prediction and gold prediction: “дорога Артемовск Луганское Дебальцево” and “дорога Артемовск Луганское Лозовое Дебальцево”.

2. word position errors. An example of prediction and gold prediction: “Киевская городская государственная администрация” and “Киевская государственная городская администрация”.

3. word ending errors. An example of prediction and gold prediction: “Верховная рада” and “Верховная рада”.

4. word case errors. An example of prediction and gold prediction: “«взрослый» Арктический Совет” and “«взрослый» Арктический совет”.

5. consistency errors. An example of prediction and gold prediction: “Южно-Русский газоконденсатный месторождение” and “Южно-Русское газоконденсатное месторождение”.

6. errors with foreign words. An example of prediction and gold prediction: “Укрнафта” and “Укртанснафта”.

7. errors with POS tags mismatches. An example of prediction and gold prediction: “Новороссийский” and “Новороссийск”.

Some of the described errors can be avoided by finetuning the model with more extra data.

6 Conclusion and Future Work

We present the results of our participation in the DE2021: RuNormAS (Russian Normalization of Annotated Spans) task. The implemented methods in both subtracks are based on RuGPT3XL LM. As future work, we plan to finetune RuGPT3XL LM on more extra data.
The best model was presented in the paper is available open-source. We hope that our developments will be useful to the community since all the presented prototypes are easily portable to new domains and tasks.

References