Sentence Simplification for Russian using Transfer Learning

Komleva E. P.
MIPT
Moscow
komleva.1999@inbox.ru

Anastasyev D. G.
Yandex
Moscow
dan-anastasev@yandex-team.ru

Abstract

In this article, we describe a way to perform sentence simplication. Our approach is largely based on solving this problem for the English language. We compare the different pretraining schemes in the application for this problem and show that the results obtained using machine translation of tests and fine tuning are approximately similar. mBART with control tokens is used as the main model for fine tuning. In the case of machine translation of tests into English, we use BART with control tokens.

Keywords: simplification, transfer learning, mBART, SARI, control tokens
DOI: 10.28995/2075-7182-2021-20-1075-1080

Симплификация на русском языке за счет transfer learning

Комлева Е.П.
ABBYY
Москва
komleva.1999@inbox.ru

Анастасьев Д. Г.
Яндекс
Москва
dan-anastasev@yandex-team.ru

Аннотация

В этой статье мы описываем способ симплификации предложений. Наш подход во многом основан на решении этой задачи для английского языка. Мы сравниваем различные схемы предварительного обучения в приложении к этой задаче и показываем, что результаты, полученные с помощью машинного перевода тестов и transfer learning примерно похожи. В качестве основной модели для тонкой настройки используется mBART с контрольными токенами. В случае машинного перевода тестов на английский используем BART с контрольными токенами.

Ключевые слова: Симплификация, transfer learning, mBART, fairseq, SARI, control tokens

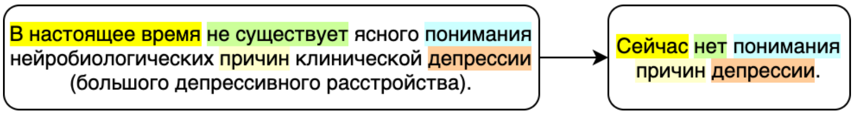
1 Introduction

Sentence simplification is the task of making a sentence easier to read and understand by reducing its lexical and syntactic complexity, while retaining most of its original meaning. Simplification has a variety of important societal applications, for example increasing accessibility for those with cognitive disabilities.

There is a lot of data for this task in English, for example Simple English Wikipedia -a large corpus whose texts are understandable to children and adults who are learning English. We also used the corpus to paraphrase sentences, since the simplification problem is a special case of the paraphrasing problem.

We can use machine translation to get data in Russian. Also for Russian language we have a large corpus of paraphrases and a small data set collected by the organizers of the competition on a crowdsourcing platform.

Below (figure 1) is an example simplification we would expect to get away with from our model.



В настоящее время не существует ясного понимания нейробиологических причин клинической депрессии (большого депрессивного расстройства).

Сейчас нет понимания причин депрессии.

Figure 1: Example

2 Related work

Data-driven methods have been predominant in English sentence simplification in recent years(Alva-Manchego et al., 2020b)[3], requiring large supervised training corpora of complex-simple aligned sentences. Methods have relied on English and Simple English Wikipedia with automatic sentence alignment from similar articles

Previous work on parallel dataset mining have been used mostly in machine translation using document retrieval , language models , and embedding space alignment to create large corpora .

We focus on paraphrasing for sentence simplifications, which presents new challenges. In "MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases"[4], the authors present control tokens to mining paraphrase data and training controllable simplification models on paraphrases. This idea worked well for the Russian model.

In article ASSET(Martin et al., 2020)[2] introduced such control tokens as length, lexical and syntactic complexity. We also assembled a case for an English-language model on a crowd sourcing platform.

3 Method

We use mBART[7] finetuned on the paraphrases from ParaPhraserPlus and automatically translated WikiSimple conditioned on specific control tokens.

The control tokens give us ability to train the model on everything that is semantically related and then to choose those control token values which work better for simplification (according to some metric).

For controllable generation, we use the small dataset collected by the organizers of the competition on a crowd-sourcing platform. We use the same control parameters as the original paper, namely length, Levenshtein similarity, lexical complexity, and syntactic complexity[2] We select the 4 hyperparameters using SARI on the validation set. We use the random search optimization . The hyper-parameters are contained in the [0.2, 1.5] interval. The 4 hyper-parameter values are then kept fixed for all sentences in the associated test set.

The following control tokens were implemented:

- Levenshtein similarity - how similar (by levenshtein metric) the result text should be;
- Chars fraction - how long the result text should be (that is, the you can specify the ratio between the result and original text lengths);
- Word rank - how simple the result text is expected to be (it's a ratio again between the ranks of the words in the texts in fasttext embeddings).
- Lexeme similarity - how similar by lexeme matching the result text should be. Wasn't used in the final model.

Figure 2 shows that SARI prefers small parameters. But such values sometimes do not preserve meaning. We believe that the best solution to the problem may be with other tokens, but then SARI is smaller.

4 Experimental Setting

4.1 Baseline

As a baseline model, we use its multilingual generalization mBART[7](from fairseq library), which was pretrained on 25 languages. BART[5] is a pre-trained seq2seq transformer, aimed at such tasks as text

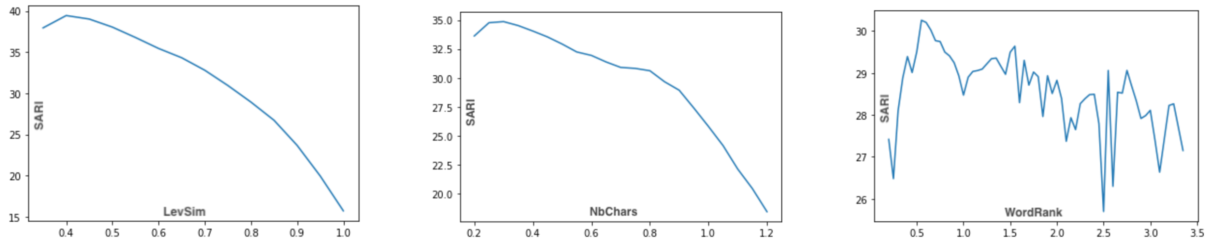


Figure 2: Dependence of the metric on control tokens

summarisation and text simplification. We pre-trained BART on first ParaPhraserPlus[6] and next on translated WikiLarge. In result we got a model giving SARI about 34 on our Gold Reference dataset[1]

Gold Reference We report gold reference scores for small data set collected by the organizers of the competition on a crowd-sourcing platform as multiple references are available. We compute scores in a leave-oneout scenario where each reference is evaluated against all others. The scores are then averaged over all references. Gold Reference gives SARI about 44.

4.2 Using English model

We can use the ready-made model for simplication in English[4]. To do this, you need to automatically translate the test data into English (we used https://github.com/nidhaloff/deep-translator). As a result, we get a model that gives similar results - SARI 39.49

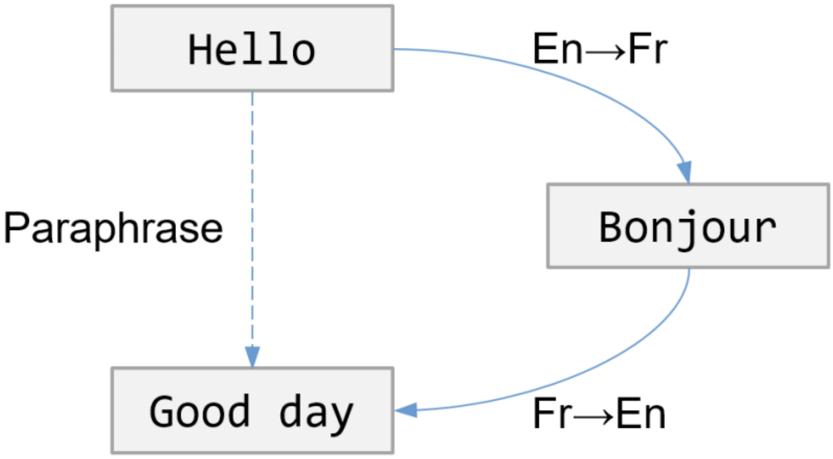


Figure 3: Translation scheme

4.3 Training data

- Automatically translated into Russian corpus WikiLarge. This corpus is twice noisy: there are bugs in the original dataset and machine translation
- The corpus assembled by the organizers on a crowdsourcing platform. Dev dataset (pairs complex - simple sentence)[1] Complex proposals from the public test for the first phase of the competition are available here. In this dataset, for each complex sentence, there are from 1 to 5 reference ones, which allows the SARI to be measured in the best way.
- Large corpus of ParaPhraserPlus header paraphrases clusters. [6]

4.4 Additional improvements

Usually in brackets people indicate information that complements the main text, so you can get a small increase in quality if you remove the expressions in brackets (about 0.2 SARI)

We also tried auto correct spelling and grammar suggestions, but it didn't give much result (0.01 SARI)

It was also possible to get an improvement in the model's performance (about 0.1) by trimming the result of each sentence to 300 characters

4.5 Evaluation

We evaluate our models against the dataset provided by the organizers. It consists of 1000 human-simplified sentences. It contains 1-5 references per source (dataset).

We evaluate with the standard metrics SARI.

SARI Sentence simplification is commonly evaluated with SARI (Xu et al., 2016), which compares model-generated simplifications with the source sequence and gold references. It averages F1 scores for addition, keep, and deletion operations. We compute SARI with the EASSE simplification evaluation suite

5 Results

In 1 are examples of how our model works.

Query	Базовая версия 172 появилась в ноябре 1955 года, как модель 1956 года и оставалась в производстве вплоть до замены моделью 172A в начале 1960 года.
Mined	Первый 172 был выпущен в ноябре 1955 года. Когда он вышел, он назывался моделью 1956 года, после, он назывался 172A.
Query	CD8 (Кластер дифференцировки 8) — трансмембранный гликопротеин, служащий корецептором Т-клеточных рецепторов (TCR).
Mined	CD8 представляет собой белок. Это трансмембранный гликопротеин, который содержится в крови.
Query	Административно-торговый центр Уфы сохранялся на территории исторического ядра города в течение длительного времени.TCR).
Mined	Долгое время административный и торговый центр Уфы находился в историческом центре г.орода

Таблица 1: Examples of Mined Simplificationx

Simplification, although sometimes not preserving the entire meaning, display various rewriting operations, such as lexical substitution, compression or sentence splitting.

The model is good at separating sentences by meaning and replacing complex words with simple analogs, but not ideal.

Let's give examples of the model's operation for different values of control tokens using the example of a sentence: "Python популярен среди индивидуальных разработчиков, но также используется крупными компаниями в достаточно серьёзных продуктах, ориентированных на получение прибыли. и пр." (table 2)

Finally, we present an overview of all the methods used and their results.(table 3)

The solution with the result 39.69 is presented in the repository. You can also test our solution using translation from the link tap here.

6 Future work:

In the process of working on the solution, we identified the following ways to improve:

- Remove noise from data.
- Simple wikipedia obtained with machine translation has many errors, you can fix these errors.
- Larger and more diverse dataset should be used in future, paraphrases should be mined similarly to MUSS;

Mined	Levenshtein similarity	Chars fraction	Word rank	SARI
Python - это программное обеспечение,используемое для разработчиков программного обеспечения.	0.4	0.95	1.6	36.47
Python популярен среди индивидуальных разработчиков, но также используется крупными компаниями.	0.9	0.5	0.9	26.20
Python популярен среди индивидуальных разработчиков, но также используется крупными компаниями в достаточно серьёзных продуктах, ориентированных на получение прибыли..	1	1.2	3.35	5.7
Python также используется для создания крупных программных приложений.	0.45	0.5	0.6	34.90

Таблица 2: Examples of Mined Simplification for different values of control tokens.

Method	Result
Baseline	34.04
Baseline+ fine tuning on ParaPhraserPlus and Wiki Simple	37.02
BART+control tokens+ test data translation	39.49
BART+control tokens + test data translation	39.49
BART+control tokens + test data translation + remove brackets	39.65
BART+control tokens + test data translation + remove brackets + auto correct	39.66
BART+control tokens + fine tuning on ParaPhraserPlus and Wiki Simple	39.69
Baseline+control tokens + test data translation + remove brackets + auto correct + trim characters after 300	39.74

Table 3: Overview of all the methods.

- Try to teach more different options with different hyperparameters.
- Another lexical complexity score instead of WordRank.
- Some other similarity metric instead of LevSim to handle semantics better.

7 Conclusion

We propose an approach to simplify sentences in Russian, which is based on the use of 4 changeable hyperparameters and fine tuning of the model on a simple wikipedia and ParaPhraserPlus.

Similar results can be obtained by translating the tests into English and using the English model, but it takes longer.

The results do not always preserve the original meaning of the sentence. In our opinion, SARI is not the best metric to evaluate in this task. This metric prefers a lower levenshtein similarity, but this often changes the meaning.

We showed that control-token-based simplification generation is simple yet effective method which is able to lead to the state-of-the-art results;

Source code, trained model, preprocessed data and demo are available here

Example of simplification with MUSS and round-trip translation is available here

References

[1] Andrey Sakhovskiy Alexandra Izhevskaya Alena Pestova. RuSimpleSentEval-2021 Shared Task: Evaluating Sentence Simplification for Russian.

[2] Fernando Alva-Manchego Louis Martin Antoine Bordes. ASSET: A dataset for tuning and evaluation of sentence simplification models with mul-tiple rewriting transformations.

[3] Fernando Alva-Manchego Carolina Scarton Lucia Specia. Data-DrivenSentence Simplification: Survey and Benchmark. Computational Linguistics.

[4] Louis Martin Angela Fan. MUSS: Multilingual Unsupervised Sentence Simplification by Mining Paraphrases // arXiv:2005.00352. — 2020.

[5] Mike Lewis Yinhan Liu Naman Goyal. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation,Translation, and Comprehension.

[6] Vadim Gudkov Olga Mitrofanova Elizaveta Filippskikh. Automatically Ranked Russian Paraphrase Corpus for Text Generation.

[7] Yinhan Liu Jiatao Gu Naman Goyal. Multilingual Denoising Pre-training for Neural Machine Translation.