NLP Course Project

Semen Kirsanov

May 2024

Abstract

My project is to study various translators from English to Russian and connect their own additional function to them for correct translation from a female person. Link to the project code: `https://github.com/anchsemen/NLP_project`.

1 Introduction

When translating English text into Russian, the entire text is translated from the male person, which in my other project became a serious problem, because it was necessary to do the translation from the female person.

My other project is as follows: when generating responses (based on the ChatGPT API), too many tokens are used in Russian (each letter is 1 token), and in English there are significantly fewer. Therefore, to save tokens, it is better to generate an answer in English and then translate it. But if it is necessary to generate text from a female face, then translators cannot cope and constantly translate the text from a male face.

For example, for the same text in Russian, 2 times more tokens are spent than in English (information is taken from the site `https://platform.openai.com/tokenizer`)

Text	**Number of tokens**
Oh, that sounds like a lot of fun! Sheregesh is a beautiful place, I'm sure you had a great time with your friends. Have you been skiing or snowboarding? It's nice to relax and spend time with friends. How was the trip overall? Have you had any memorable experiences?	63
О, это звучит очень весело! Шерегеш – красивое место, я уверен, вы отлично провели время с друзьями. Вы катались на лыжах или сноуборде? Приятно отдохнуть и провести время с друзьями. Как в целом прошла поездка? Были ли у вас какие-нибудь запоминающиеся впечатления?	128

Table 1: Distribution of the number of tokens depending on the language of the text

Analogs: There is no good translator (API/ML model) from English to Russian, which, when specifying which person the text should be from, therefore this new research will help at least not only my project, but maybe others.

1.1 Team

Semen Kirsanov prepared this document and did the whole project.

2 Related Work

Since there is no translator who takes into account the result of the translation from which person it should be made, I will give the works of translators that I use and supplement with my part for translating from a female person.

To check the functionality of the project for text translation, two translators were used - opus-mt-en-ru and t5-translate-en-ru-zh-large-1024 UtrobinMV (2024), some metrics on the TREX dataset are listed in the table.

name	params	bleu	bertscore_f1	meteor
Helsinki-NLP/opus-mt-en-ru	73M	25.0095	0.8476	0.4722
utrobinmv/t5-translate-en-ru-zh-large-1024	811M	**28,0225**	**0,8586**	**0,5272**

Table 2: Benchmark translation to TREX dataset

2.1 Helsinki-NLP/opus-mt-en-ru

The opus-mt-en-ru translator from Helsinki-NLP is a machine translation model developed based on the MarianMT architecture (Tiedemann and Thottingal, 2020)

Architecture

- **Transformer:** The model is based on transformers that use attention mechanisms to process sequences of words. Transformers consist of an encoder and a decoder, where the encoder converts the input sequence in English into contextual representations, and the decoder generates a sequence in Russian.
- **MarianMT:** This is a transformer implementation specifically optimized for machine translation tasks. MarianMT supports learning and inference in several languages, which makes it effective for translating from English to Russian.
- **Tokenization:** The SentencePiece tokenizer is used for text preprocessing, which breaks the text into smaller subwords. This helps the model to cope with morphologically rich text and words that were not found in the training sample.

Training

- **OPUS Corpus:** The model was trained on data from the OPUS corpus, which includes parallel texts in different languages. The corpus contains a variety of text types, including news, legal documents, and other sources, which provides the model with a broad linguistic base.
- **Data processing:** During the learning process, data is preprocessed and normalized. The texts are divided into sentences and tokens, then pairs of sentences in different languages are submitted to the input of the model for training.

2.2 t5-translate-en-ru-zh-large-1024

The utrobinmv/t5-translate-en-ru-zh-large-1024 translation model is based on the T5 (Text-To-Text Transfer Transformer) architecture (Raffel et al., 2020) and is designed for translation between English (en), Russian (ru) and Chinese (zh) languages. The model can perform direct translation between any of these languages using the prefix "translate to:" before the target text. The model is trained on the CCMatrix dataset, which is a large corpus of parallel texts for machine translation.

Architecture

The T5 model is a transformer using an all-in-one approach for natural language processing tasks. The main components of the T5 architecture include:

1. **Encoder-Decoder Architecture:** The T5 uses a symmetric encoder-decoder architecture where both components are built on the basis of attention mechanisms. The encoder converts the input text into hidden representations, which are then used by the decoder to generate the output text.
2. **Multitasking Mode:** The model is configured to work in multitasking mode, which allows it to translate between different language pairs using the target language identifiers as prefixes. For example, the prefix "translate to zh" is used to translate from Russian to Chinese: ".
3. **Generation:** To generate text, the model uses a multi-beam configuration (beam search) and prevents the repetition of n-grams, which improves the quality of translation.

3 Model Description

Two syntactic parsers are used to change texts: UDPipe and NewsSyntaxParser from the natasha library.

Next, I will describe in detail how parsers work.

3.1 UDPipe

UDPipe is a natural language processing tool that performs sentence segmentation, tokenization, POS tagging, lemmatization, and parsing (Straka and Straková, 2017)

UDPipe Architecture

- **Offer segmentation and tokenization**

  UDPipe 1.1 uses a single-layer bidirectional GRU network that predicts for each character whether it is the last in the sentence, the last in the token, or not the last in the token.

  Tokenization and segmentation of sentences are performed together, which allows you to more effectively divide the text into tokens and sentences.

  UDPipe 1.1 also uses automatically generated suffix rules to separate verbose tokens.

- Definition of parts of speech (POS tagging) and lemmatization

  The tagger uses the averaged perceptron method to disambiguate tags, and the hypothesis generator (guesser) generates possible triplets (UPOS, XPOS, FEATURES) for each word based on its last four characters.

  The lemmatizer works similarly, generating hypotheses for lemmas based on transformation rules that include removing and adding prefixes and suffixes.

- Syntactic analysis

  UDPipe 1.0 and 1.1 use a fast transient parser based on neural networks. It is based on a simple neural network with one hidden layer and no recurrent connections, using locally normalized estimates.

  The parser supports several transition systems, including design and non-design systems, and uses static and dynamic oracles to predict transitions.

  Embeddings of forms, UPOS, FEATURES, and DEPREL are used to improve parsing accuracy.

The work of the parser

- Tokenization

  The parser starts by splitting the text into tokens using a bidirectional GRU network. This network learns from training data and predicts the boundaries of tokens and offers.

- Definition of parts of speech and lemmatization

  The hypothesis generator creates possible part-of-speech tags and lemmas for each word using information about suffixes and prefixes. The perceptron tagger then selects the most likely tags based on the trained data.

- Syntactic analysis

  The parser uses a transitional model to build a syntactic sentence tree. It uses predicted parts of speech, lemmas, and other features to determine dependencies between words.

  UDPipe 1.1 supports various transition systems, which allows you to choose the most appropriate system for a specific language or task.

3.2 NewsSyntaxParser from natasha

NewsSyntaxParser from the Natasha library is used for parsing text in Russian. Natasha includes several components, each of which is responsible for certain text processing tasks, such as tokenization, morphological analysis, syntactic analysis, etc (Developers, 2020)

NewsSyntaxParser Architecture

NewsSyntaxParser is based on deep neural networks and modern NLP models.

- Tokenization and Sentence Splitting:

  The Razdel library is used to segment text into sentences and tokens. Razdel provides high segmentation accuracy and speed, which is critical for subsequent text processing.

- Embedding Layer:

  The models use pre-trained embeddings of words from the Navec library. These embeddings are dense vector representations of words that capture their semantic and syntactic properties.

- Morphological Analysis:

  The Slovnet module is used for morphological analysis. It defines the morphological properties of each token, such as part of speech, case, number, and other grammatical features.

- Syntactic Parsing:

  Slovnet is also used for parsing, which builds a dependency tree for sentences. This tree shows how words relate to each other at the sentence level.

The work of the parser

- Tokenization and segmentation:

  First, the text is split into sentences and tokens using Razdel. This allows the parser to work with individual sentences and words.

- Preprocessing and embeddings:

  Tokens are converted into vector representations using pre-trained embeddings from Navec.

- Morphological analysis:

  Morphological analysis is performed on each token using Slovnet, which allows you to determine the grammatical features of words.

- Syntactic analysis:

  A syntactic dependency tree is built based on morphologically marked tokens, which displays the syntactic structure of the sentence.

3.3 The pipeline of the project

1. Divide the sentences into parts to get a simpler syntactic structure;
2. Determine the syntactic structure using the UDPipe and NewsSyntaxParser parsers - they supplement information about a part of a sentence (root, subject and words depending on them), as well as morphological features of each word (from which person it comes, gender, part of speech);
3. Depending on the morphological features, we determine whether it is necessary to change the text to another gender using pymorphy2;
4. Combine the text and give the result.

4 Dataset

The dataset was assembled manually, because the primary goal of the project is to redo the texts generated by models, after translation from English into Russian. There were 30 examples in total, of which some of them consist of personally generated texts, as well as datasets:

- generated-reviews-enth - product reviews in English
- yelp-polarity - dataset for binary sentiment classification
- cefr-levelled-english-texts - a dataset containing about 1,500 English texts labeled with CEFR levels

Criteria for selecting texts for a dataset:

- It was necessary to choose texts with a lot of information from the first person;
- A text with a large number of adjectives dependent on the subject or predicate;

- Texts with a diverse sentence structure and complex sentences to ensure a variety of syntactic representations;
- Texts that include dialogues or direct speech to reflect spoken language patterns.

5 Experiments

5.1 Metrics

As a metric, the usual accuracy is used, where correctly translated words are calculated in the numerator, and all translated words are calculated in the denominator.

$$accuracy = \frac{N_{correct}}{N_{total}},$$

where $N_{correct}$ - the number of correctly translated words, N_{total} - the total number of translated words.

5.2 Experiment Setup

There was only one run, each part of the sentence was checked for a change of form. Everything is specified in the pipeline of the project 3.3.

5.3 Baselines

The simplest options are to translate the text without my part of translating the words into the feminine gender. Therefore, I will show the results of the translation without taking into account my part with the help of translators: **opus-mt-en-ru** and **utrobinmv/t5-translate-en-ru-zh-large-1024**.

An example from the dataset: I've been meaning to write for ages and finally today I'm actually doing something about it. Not that I'm trying to make excuses for myself, it's been really hard to sit down and write, as I've been moving around so much. Since we last saw each other I've unpacked my bags in four different cities.

opus-mt-en-ru	utrobinmv
Я давно собирался писать, и, наконец, сегодня я действительно что-то делаю с этим. Не то, чтобы я пытался оправдать себя, было очень трудно сидеть и писать, так как я так много переезжал. С тех пор, как мы виделись в последний раз, я распаковал свои сумки в четырех разных городах.	Я собирался писать в течение многих веков и, наконец, сегодня я на самом деле делаю что-то с этим. Не то чтобы я пытался оправдываться для себя, это было действительно трудно сидеть и писать, поскольку я так много двигался вокруг. С тех пор как мы последний раз видели друг друга, я распаковал свои сумки в четырех разных городах.

Table 3: Translation of an example from experiments by models

6 Results

As can be seen from the example in the section 5.3, if we need to translate from a female face, even such models cannot give us the desired result. Next, let's look at the results of my project using several examples.

The words that should not have been changed are marked in **red**, and those that were changed correctly are marked in **green**.

6.1 Example from the section 5.3

6.1.1 Model utrobinmv

utrobinmv	My result
Я собирался писать в течение многих веков и, наконец, сегодня я на самом деле делаю что-то с этим. Не то чтобы я пытался оправдываться для себя, это было действительно трудно сидеть и писать, поскольку я так много двигался вокруг. С тех пор как мы последний раз видели друг друга, я распаковал свои сумки в четырех разных городах.	Я **собиралась** писать в течение многих веков и, наконец, сегодня я на самом деле **делала** что-то с этим. Не то чтобы я **пыталась** оправдываться для себя, это было действительно трудно сидеть и писать, поскольку я так много **двигалась** вокруг. С тех пор как мы последний раз видели друг друга, я **распаковала** свои сумки в четырех разных городах.

Table 4: Differences between the translation of the model utrobinmv and the addition of my project from 5.3

6.1.2 Model opus-mt-en-ru

opus-mt-en-ru	My result
Я давно собирался писать, и, наконец, сегодня я действительно что-то делаю с этим. Не то, чтобы я пытался оправдать себя, было очень трудно сидеть и писать, так как я так много переезжал. С тех пор, как мы виделись в последний раз, я распаковал свои сумки в четырех разных городах.	Я давно **собиралась** писать, и, наконец, сегодня я действительно **делала** что-то с этим не то, чтобы я **пыталась** оправдать себя, было очень трудно сидеть и писать, так как я так много **переезжала** с тех пор, как мы виделись в последний раз я **распаковала** свои сумки в четырех разных городах.

Table 5: Differences between the translation of the model opus-mt-en-ru and the addition of my project from 5.3

As you can see from the results, the two models were translated approximately the same way and my project did well. However, this does not always work out, let's look at the following example.

6.2 Example 2

An example from the dataset: I'm writing this review to give you a heads up before you see this Doctor. The office staff and administration are very unprofessional. I left a message with multiple people regarding my bill, and no one ever called me back. I had to hound them to get an answer about my bill. Second, and most important, make sure your insurance is going to cover Dr. Goldberg's visits and blood work. He recommended to me that I get a physical, and he knew I was a student because I told him. I got the physical done.

6.2.1 Model utrobinmv

utrobinmv	My result
Я пишу этот обзор, чтобы дать вам понять, прежде чем вы увидите этого Доктора. Офисный персонал и администрация очень непрофессиональны. Я оставил сообщение с несколькими людьми относительно моего счета, и никто никогда не перезвонил мне. Мне пришлось поймать их, чтобы получить ответ о моем счете. Во-вторых, и самое главное, убедитесь, что ваша страховка будет покрывать визиты доктора Голдберга и анализ крови. Он порекомендовал мне сделать физическое обследование, и он знал, что я студент, потому что я сказал ему. Я сделал физическое лечение.	Я писала этот обзор, чтобы дать вам понять, прежде чем вы увидите этого Доктора. Офисный персонал и администрация очень непрофессиональны. Я **оставила** сообщение с несколькими людьми относительно моего счета, и никто никогда не **перезвонила** мне. Мне пришлось поймать их, чтобы получить ответ о моем счете. Во-вторых, и самое главное, убедитесь, что ваша страховка будет **покрывала** визиты доктора Голдберга и анализ крови. Он порекомендовал мне сделать физическое обследование, и он знал, что я студент, потому что я **сказала** ему. Я **сделала** физическое лечение.

Table 6: Differences between the translation of the model utrobinmv and the addition of my project example 2

6.2.2 Model opus-mt-en-ru

opus-mt-en-ru	My result
Я пишу этот обзор, чтобы предупредить вас прежде, чем вы увидите этого доктора. Офис и администрация очень непрофессиональны. Я оставил сообщение с несколькими людьми по поводу моего счета, и никто мне не перезвонил. Я должен был выследить их, чтобы получить ответ на мой счет. Во-вторых, самое главное, убедиться, что ваша страховка покроет визиты доктора Голдберга и анализ крови. Он рекомендовал мне пройти медосмотр, и он знал, что я студент, потому что я сказал ему. Я сделал физический анализ.	Я **писала** этот обзор, чтобы предупредить вас прежде, чем вы увидите этого доктора. Офис и администрация очень непрофессиональны. Я **оставила** сообщение с несколькими людьми по поводу моего счета, и никто мне не **перезвонила**. Я **должна была** выследить их, чтобы получить ответ на мой счет. Во-вторых, самое главное, убедиться, что ваша страховка покроет визиты доктора Голдберга и анализ крови. Он рекомендовал мне пройти медосмотр, и он знал, что я студент, потому что я **сказала** ему. Я **сделала** физический анализ.

Table 7: Differences between the translation of the model opus-mt-en-ru and the addition of my project example 2

As can be seen from the results of the translation of example 2, the models were translated differently and the text was changed in different ways using the add-in of my project. The words that should not have been changed are marked in red, and those that were changed correctly are marked in green.

The general accuracy of formatting text to text from a female person is as follows:

Model	Accuracy
opus-mt-en-ru	0.72
utrobinmv	0.7

Table 8: Accuracy of translation models

7 Conclusion

As part of my project:

- A small demo dataset was assembled
- It was possible to achieve formatting of the translation into text from a female person.

The models showed approximately the same results, but the correctness of the translation depends on how well my project will be able to work with formatting.

What can I do next:

- To achieve a more correct form replacement (take into account other morphological features)
- Indicate more general signs when highlighting words that need to be changed to words from a female person.
- Come up with a complete model for translating and replacing words from a female person

References

Developers, N. (2020). Natasha: Library for russian nlp. Available at `https://github.com/natasha/natasha`.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 21(140). `https://github.com/google-research/text-to-text-transfer-transformer`.

Straka, M. and Straková, J. (2017). Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, Vancouver, Canada.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

UtrobinMV (2024). Сравнение локальных моделей машинного перевода для английского, китайского и русского языков. Available at https://habr.com/ru/articles/791522/.