Controllability for English-Ukrainian Machine Translation Based on Specialized Corpora

Presented by
Daniil Maksymenko

Multi3Generation
(June 15th, 2023)

Research Authors

Daniil Maksymenko
Software Engineering Department
NURE
Kharkiv, Ukraine

Nataliia Saichyshyna
Software Engineering Department
NURE
Kharkiv, Ukraine

Oleksii Turuta
Software Engineering Department
NURE
Kharkiv, Ukraine

Olena Turuta
Philosophy Department
NURE
Kharkiv, Ukraine

Maksym Yerokhin
Software Engineering Department
NURE
Kharkiv, Ukraine

Research area analysis

- machine translation models usually work as black boxes: they get input text and provide one or a few translations without any chance to interpret their results or modify them
- controllable machine translation solutions usually need to be retuned for each modification of output features
- translation options can be gathered by cheaper methods like frequency dictionary, which do not understand text context

Aims of research

- Determine methods to increase controllability of machine translation models by modifying existing solutions
- Propose a method for style transfer from a certain domain to input text translation (change translation sentiment, style, structure, used words based on passed domain features)
- Compare proposed solution with existing methods of control on multiple specialized corpora with different topics and styles

Available solutions for controllable machine translation task

Based on deep learning models

- tune pretrained model with a small specialized corpus to transfer necessary style;
- add special token, which should encode certain output features;
- mark topic or style with an external model and add this class to text automatically;
- concatenate text embeddings with vector-descriptor of features (length, politeness, officialness).

Metrics

$$BLEU(ref, cand, N) = \prod_{n=1}^{N} p_n^{w_n}$$

BLEU

Token metric, it gets calculated by number of matched n-grams of different sizes with even coefficients

$$METEOR = \frac{10PR}{R + 9P} = \frac{10 * \frac{m}{w_r} * \frac{m}{w_c}}{\frac{m}{w_c} + 9\frac{m}{w_r}}$$

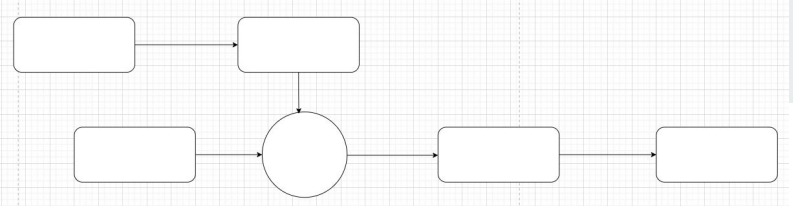
METEOR

Token metric, adds stemming and synonyms matching to the evaluation pipeline

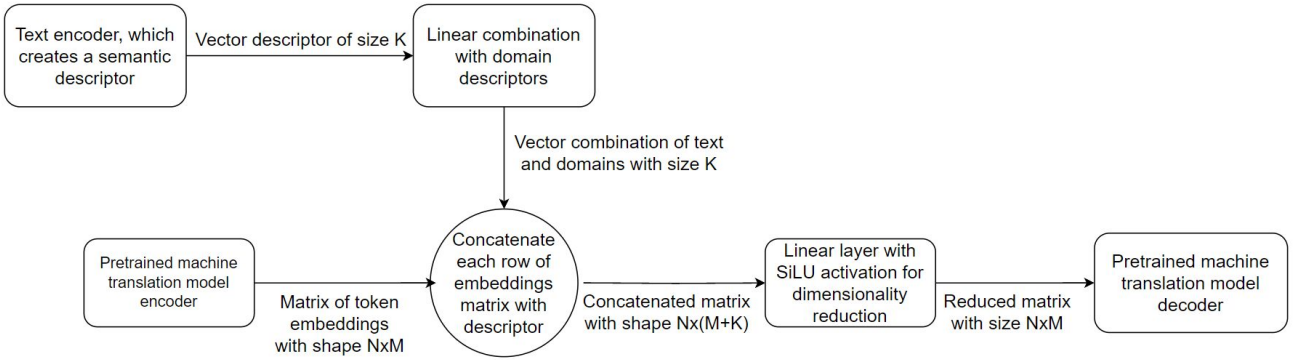
$$R_{BERT} = \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x_j} \in \hat{x}} x_i^T \hat{x}_j \quad P_{BERT} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j \quad F_{BERT} = 2\frac{PR}{P+R}$$

BERT Score

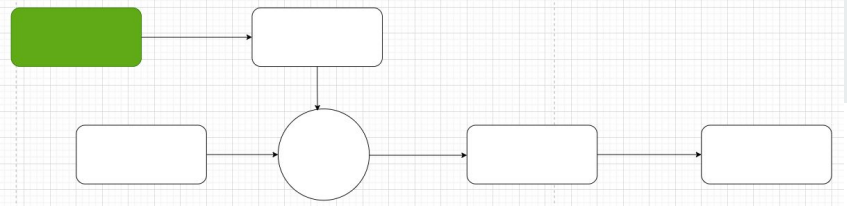
Embedding metric, evaluates texts by results of an external state-of-the-art model. It is Multilingual BERT in our case.



Proposed controllable machine translation architecture

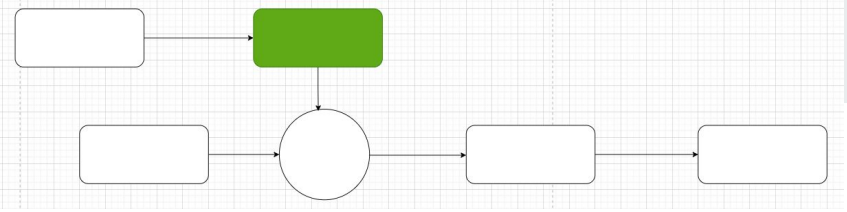


The architecture is based on **pretrained encoder-decoder MT model**. Between encoding and decoding stages we concatenate each row of tokens embeddings matrix with **vector-descriptor obtained from semantic search model**.



Text encoder: siamese BERT for semantic search

Model returns vector-descriptor for each text with 384 elements, which allow us to measure similarity of texts by using cosine similarity.



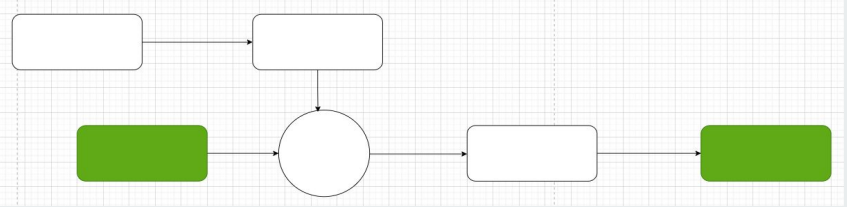
Translation controllability with linear combination of domain and text descriptors

We need to create a set of domain descriptors to transfer their respective styles and features. Let's calculate each element of domain descriptor as a mean value of corresponding texts descriptors elements:

$$feature_i = \frac{\sum_N^{j=0} vec_{ji}}{N}$$

Then we have to conduct linear combination of text descriptor and difference vector of text and domain with a certain transformation coefficient to combine both original features and features of chosen text cluster:

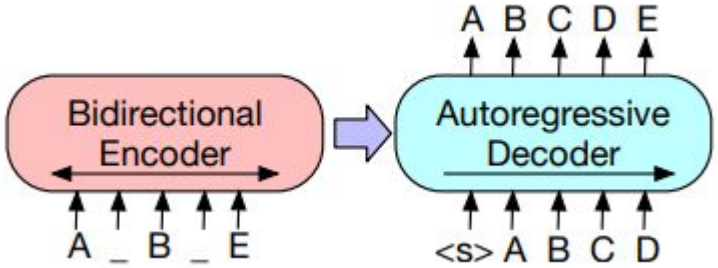
$$descriptor = V_{original} - \alpha * (V_{original} - V_{mean\ cluster})$$

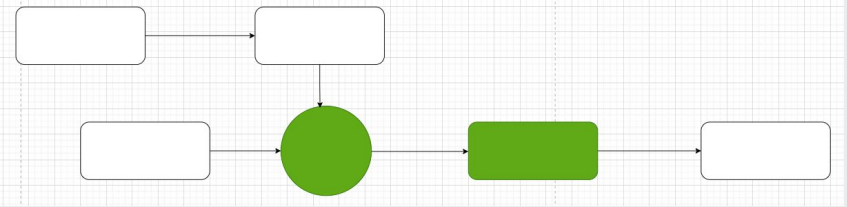


MarianMT implemented with BART interface

Results for OPUS MT checkpoint measured on original test set (TATOEBA Project subset):

BLEU: 0.5023
METEOR: 0.3917
BERT F1 Score: 0.9263





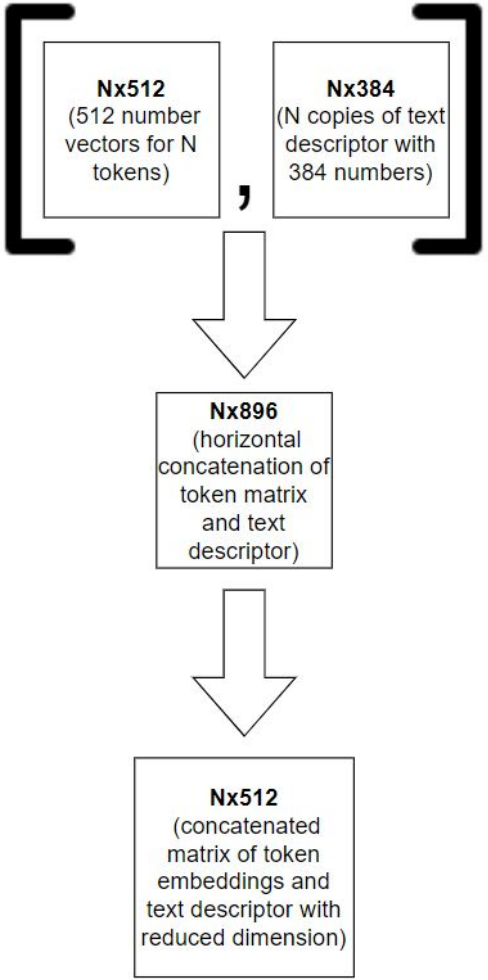
Concatenation block and dimensionality reduction

Text descriptor:
vector with 384 elements

Output of original MarianMT encoder:
matrix Nx512, where N - max number of tokens

Output of encoder matrix and text descriptor concatenation:
matrix Nx896 (512 + 384)

Output of dimensionality reduction layer:
matrix Nx512 (same as original)



Datasets

№	Dataset description	Style	Number of text pairs
1	OPUS corpora subset, which contains subtitles, TED talks, book reviews, etc.	–	1 942 849
2	Ukrainian translation of Multi30k [5]	general	30 000
3	Corpus of laws translation from Verkhovna Rada of Ukraine website	official	4 000
4	Scientific articles abstracts corpus from Google Scholar	scientific	2 000
5	Test set (25% of Multi30k, laws and abstracts)	–	9 000

Experiments plan step by step

Development environment:

Environment: Google Colab

RAM: 27 Gb

GPU: Nvidia T4

1. Train separate versions of MarianMT for each domain and one version trained with all gathered special corpora

2. Modify MarianMT according to described architecture diagram and train it from OPUS MT checkpoint with special corpora and OPUS subset

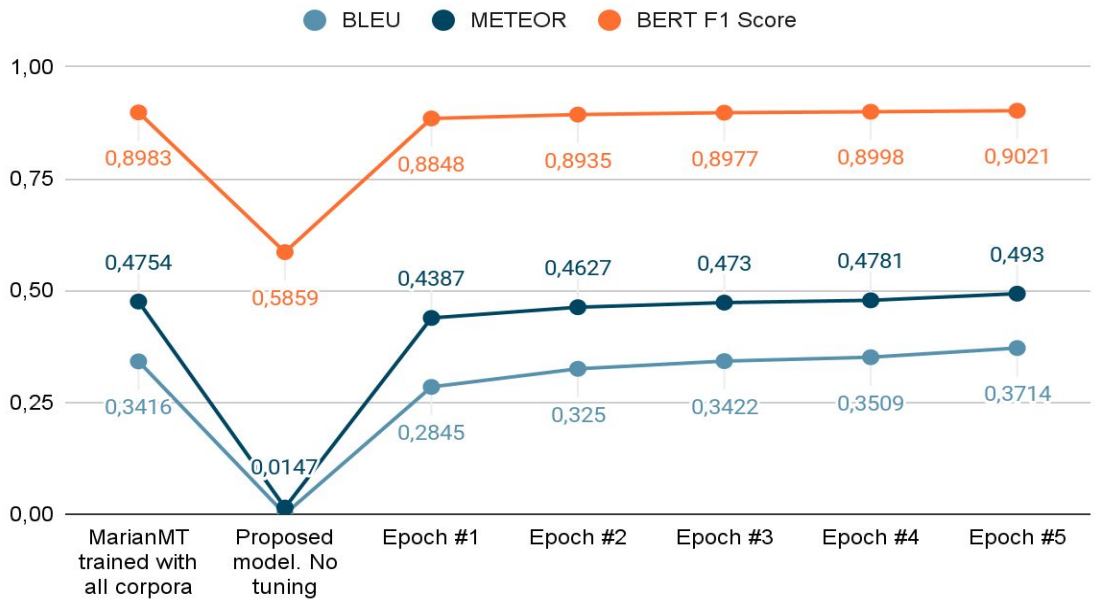
3. Measure and compare all obtained MarianMT versions

4. Check how model can modify translation by combining text with multiple domains

Results of MarianMT tuning with different specialized corpora and all texts

Model version/Metric	BLEU	METEOR	BERT F1 Score
Original	0.1120	0.2807	0.8115
Trained with Multi30k	0.1270	0.3034	0.8380
Trained with laws	0.2534	0.3861	0.8630
Trained with abstracts	0.1880	0.3347	0.8448
Trained with all corpora	**0.3416**	**0.4754**	**0.8983**

Proposed architecture training



BLEU: 0.3714 **METEOR:** 0.4930 **BERT F1 Score:** 0.9021

Model results before tuning

Original text:

“*He has to come back in the next movie*”

Correct translation:

“*Він має повернутися в наступному фільмі*”

Model generated translation before tuning:

“*Це означає, що ми маємо справу з іншими людьми, а не з ними.*”
(BERT Score F1: 0.6581)

Examples of same text translation with different transformation coefficients and chosen domains

Conclusions

- We proposed and tested architecture for controllable machine translation task with ability to use external context by passing vectors descriptors obtained from semantic search model
- We conducted a comparative research of efficiency of our proposed solutions and fine-tuned pretrained machine translation models to find out if it works better than easier for development ones
- Proved that BERT Score cannot be used as a standard benchmark for Ukrainian text generation tasks

Further developments

- We see prospects in further tuning to achieve style transfer from just one example instead of creation of predefined domain descriptors set
- We need to fully understand each descriptor element meaning and importance, so we can provide better control over certain features. It can be achieved with sparse embedding method
- We would like to modify architecture further by checking other text encoders or trying to make the model impact only certain sets of tokens
- We would like to gather more specialized corpora for Ukrainian language as it can help to improve our results and help further Ukrainian NLP research
- We would like to measure our models with COMET framework in order to check another embedding metric option instead of BERT Score

Thank you for the attention