Good morning dear committee. My name is Alexey Birshert, and I am about to demonstrate my research called "Adversarial attacks on cross-lingual models" to you.

Code-switching

I am researching natural language processing, more precisely the multilingual text corpora analysis and such a phenomenon as code-switching. Code-switching occurs in multilingual communities around the world and ordinary texts on the Internet, social media. It consists of mixing and using two or more languages within one phrase or sentence. In my work, I set myself to solve the problem of generating realistic texts with code-switching. Such texts' artificial creation would improve the existing datasets for classic multilingual tasks like recognizing user intent or question-answering systems. It would enhance language models' quality and make them more resistant to code-switching and more robust.

Texts with code-switching generation

For data generation, I would construct a language model based on Transformer architecture. I plan to utilize three different architectures - XLM, XLM-R and BERT. To train such a model, I would use adversarial attacks.

Training procces

As the primary model archetype, I have a neural network that fills "slots" in sentences. As an input, it takes a sentence with some tokens changed to the special mask token. As an output, it generates a probability distribution over tokens' vocabulary. I want to train the model in such a way that it would result in increasing other languages probabilities.

Training procces

Initially, these language models are trained on the data collected on the Internet, such as Wikipedia and books. These sources are primarily monolingual. The probability distribution over tokens that are not from the sentence's source language is very close to uniform, and every token probability is very close to zero. Here you can see a density plot of probabilities generated by the "slot" filling language model.

Training procces

To train a language model to generate tokens from the target language, I would use a complex loss function. It would consist of two components. First - I want to keep the meaning of the original sentence. I would calculate the semantic distance between the original sentence and the generated one. Second - I want to keep the classifier score for sentence classification. I would estimate the difference between the original score and the generated score.

Experiments

I project to experiment with changing the training process. In some experiments, I would try to determine how both loss function parts influence the overall performance. In other experiments, I would try to determine how to choose tokens to mask. Mainly all current approaches base on linguistic semantic constraints - the goal is to generate realistic code-switching data. The insertions could be words or more significant constituents, and they would comply with the grammatical frame of the language. However, random word insertions could lead to the formation of unnatural code-mixed sentences, which are very rare in practice.

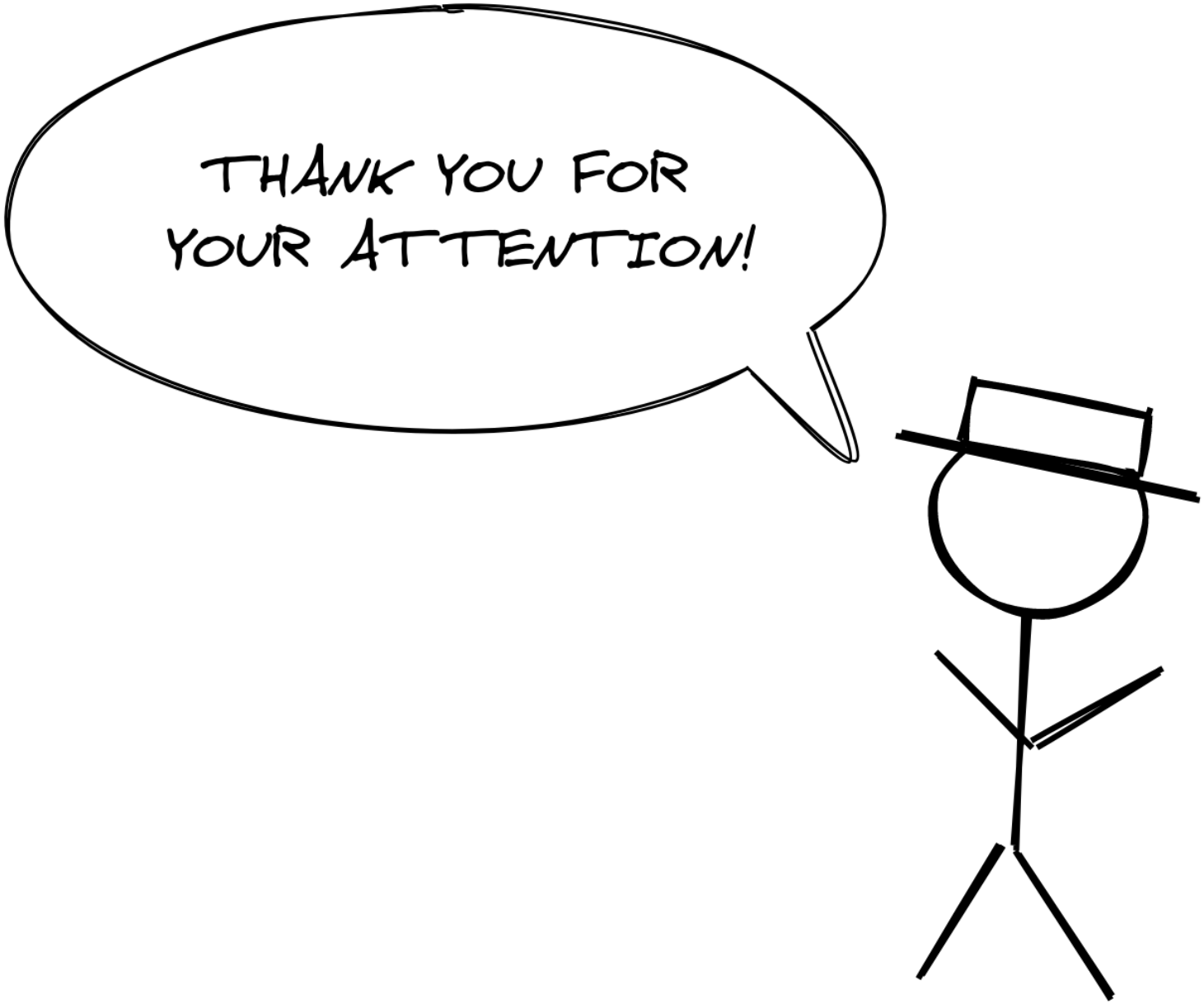
Evaluation

After training, I would evaluate my experiments and resulting models by two types of metrics. First - I would measure code-switching specific metrics that cover the quality and realism of the generated text. Second - I would finetune a language model for the user-intent classification task and compare the overall performance.

The research's main expected result is a training algorithm for the Transformer type neural networks to generate realistic code-switching data. A trained model should be capable of generating multilingual code-mixed training data for the question-answering task. It also should achieve higher metrics values than existing methods.

Future plans

This research should help classical language models become more robust and work better with multilingual texts. Further research on the topic of code-switching can use the resulting model for conducting various experiments.
In the future, I plan to experiment with classical language models fine-tuned on the generated data and improve their quality on particular benchmarks.


THANK YOU FOR
YOUR ATTENTION!

1. "Attention Is All You Need." arXiv preprint arXiv:1706.03762v5.
2. "Differentiable Language Model Adversarial Attacks on Categorical Sequence Classifiers." arXiv preprint arXiv:2006.11078.
3. "A Semi-supervised Approach to Generate the Code-Mixed Text using Pre-trained Encoder and Transfer Learning." Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2267–2280
4. "DialoGLUE: A Natural Language Understanding Benchmark for Task-Oriented Dialogue." arXiv preprint arXiv:2009.13570v2.
5. Guillaume Lample, Alexis Conneau. 2019. "Cross-lingual Language Model Pretraining." arXiv preprint arXiv:1901.07291.
6. "Unsupervised Cross-lingual Representation Learning at Scale." arXiv preprint arXiv:1911.02116v2.