Transformer Methods and Their Application in the Field of Text Analysis

Martin Bača
Supervisor: RNDr. Šimon Horvát, PhD.
Consultant: doc. RNDr. Ľubomír Antoni, PhD.

FACULTY OF SCIENCE
PAVOL JOZEF ŠAFÁRIK UNIVERSITY IN KOŠICE

Motivation and Problem

Modern society produces enormous amounts of text every day, which makes effective access to information increasingly challenging.
Main challenges are:

- Rapid growth of digital text requires fast and accurate information retrieval.
- Traditional keyword-based search is often insufficient – users expect direct, reliable answers.
- Question Answering (QA) systems based on Transformer models are state-of-the-art, but datasets are often monolingual and task-specific.

Objectives

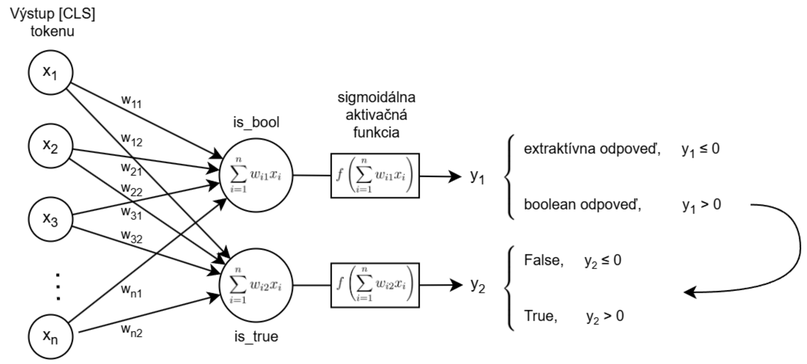
The thesis sets out to improve QA systems by creating new sythetic dataset and adapting models to multilingual and mixed tasks.
Specifically, it aims to:

1. Create a multilingual QA dataset (English, German, French).
2. Cover extractive and binary QA tasks.
3. Extend Transformer models with a classification head for handling question types.
4. Evaluate model performance across sizes and computational costs.

Methods

The methodology combines data generation, model training, and systematic evaluation.
Particular methods are:

- Automatic dataset generation using large generative language models → eliminates manual annotation.
- Preprocessing and tokenization tailored for multilingual input.
- Training BERT-based architectures (RoBERTa, XLM-R) in monolingual and multilingual settings.
- Evaluation with accuracy, F1, and precision/recall metrics.



Results

Experiments confirm that larger models achieve higher accuracy, but with greater computational demands.
Key findings include:

- Best performance: XLM-RoBERTa Large
  - 90.6% accuracy on binary QA
  - 87.7% F1-score on extractive QA
- Compact models provide efficiency but lower accuracy.
- Trade-off identified between model size and performance.

Outcomes

These results demonstrate the potential of multilingual QA systems in practical applications.
Concretely, the work:

- Demonstrated feasibility of automatically generated multilingual datasets.
- Provided a scalable QA approach adaptable to different languages and question types.
- Results applicable in:
  - Customer support chatbots
  - Legal/medical document search
  - Multilingual digital assistants

Conclusion

The thesis highlights how Transformer models can be tailored for multilingual QA, balancing performance and efficiency. This work advances multilingual QA by combining dataset generation and model extension. It shows how Transformers can be optimized for both precision and efficiency, with potential impact on how people worldwide access information.