Czech
Technical University
in Prague

F3

Faculty of Electrical Engineering
Department of Computer Science

NLP Methods for Automated Fact-Checking

Dissertation Minimum Study of
Ing. Herbert Ullrich

FCHECK.FEL.CVUT.CZ

Supervisor: **Ing. Jan Drchal, Ph.D.**
Field of study: **Informatics**
Subfield: **Natural Language Processing**
August 2023

Chapter 1
Introduction

1.1 Motivation

The spread of misinformation in the online space has a growing influence on the Czech public [STEM, 2021]. It has been shown to influence people's behaviour on the social networks [Lazer et al., 2018] as well as their decisions in elections [Allcott and Gentzkow, 2017], and real-world reasoning, which has shown increasingly harmful during the COVID-19 pandemic [Barua et al., 2020].

The recent advances in artificial intelligence and its related fields, in particular the recommendation algorithms, have contributed to the spread of misinformation on social media [Buchanan and Benson, 2019], as well as they hold a large potential for automation of the false content generation and extraction of sensational attention-drawing headlines – the "clickbait" generation [Shu et al., 2018].

Recent research has shown promising results [Thorne et al., 2019] in false claim detection for data in English, using a trusted knowledge base of true claims (for research purposes typically fixed to the corpus of Wikipedia articles), mimicking the *fact-checking* efforts in journalism.

Fact-checking is a rigorous process of matching every information within a *factic claim* to its *evidence* (or *disproof*) in trusted data sources to infer the claim veracity and verifiability. In exchange, if the trusted *knowledge base* contains a set of "ground truths" sufficient to fully infer the original claim or its negation, the claim is labelled as **supported** or **refuted**, respectively. If no such *evidence set* can be found, the claim is marked as **unverifiable**[1].

1.2 Challenges

Despite the existence of end-to-end fact-checking services, such as `politifact.org` or `demagog.cz`, the human-powered approach shows weaknesses in its scalability. By design, the process of finding an exhaustive set of evidence that decides the claim veracity is much slower than that of generating false or misguiding claims. Therefore, efforts have been made to move part of the load to a computer program that can run without supervision.

The common research goal is a fact verification tool that would, given a claim, semantically search provided knowledge base (stored for example as a *corpus* of some natural language), propose a set of evidence (e. g. k semantically nearest paragraphs of the corpus) and suggest the final verdict (Figure **??**). This would reduce the fact-checker's workload to mere adjustments of the proposed result and correction of mistakes on the computer side.

[1] Hereinafter labelled as `NOT ENOUGH INFO`, in accordance to related research.

The goal of the ongoing efforts of FactCheck team at AIC CTU, as addressed in the works of [Rýpar, 2021,Dědková, 2021] and [Gažo, 2021] is to explore the state-of-the-art methods used for fact verification in other languages, and propose a strong baseline system for such a task in Czech.

1.2.1 Challenge subdivision

In order to maximize our efficiency and the depth of our understanding of every relevant subproblem, we have divided the fact-checking task according to the Figure **??** among the members of our research group.

The works of [Rýpar, 2021] and [Dědková, 2021] focus on the Document Retrieval task and compare the performance of the numerical methods, s.a the *tf–idf* search and the *bag-of-words*, to the neural models, most notably the state-of-the-art *Transformer networks* [Vaswani et al., 2017]. [Gažo, 2021] is proposing the methods of their scaling for long inputs, such as full news reports.

1.2.2 Our contribution

Our part is to provide the needed datasets for the fact verification tasks in the *low-resource* Czech language. We examine both major ways of doing so – localizing the large-scale datasets available in the high-resource languages, typically in English, and collecting a novel dataset through human annotation experiments.

Our second task is to establish a baseline for the final task of the fact-checking pipeline: the *Natural Language Inference*, which is a decisioning problem of assigning a veracity verdict to a claim, given a restricted *set of evidence* in the Czech natural language.

In continuation with research funded by TAČR, experiments are to be made using the archive of the Czech News Agency (hereinafter referred to as ČTK[2]) for a knowledge base, exploring whether a corpus written using journalistic style can be used for such a challenge.

1.3 A word on the Transformers

For the past four years, the state-of-the-art solution for nearly every Natural Language Processing task is based on the concept of *transformer networks* or, simply, *Transformers*. This has been a major breakthrough in the field by [Vaswani et al., 2017], giving birth to the famous models such as Google's BERT [Devlin et al., 2019] and its descendants, or the OpenAI's GPT-3 [Brown et al., 2020].

In our proposed methods, we use Transformers in every step of the fact verification pipeline. Therefore, we would like to introduce this concept to our reader to begin with.

Transformer is a neural model for *sequence-to-sequence* tasks, which, similarly e.g. to the *LSTM-Networks* [Cheng et al., 2016], uses the Encoder–Decoder architecture. Its main point is that of using solely the *self-attention* mechanism to represent its input and output, instead of any sequence-aligned recurrence [Vaswani et al., 2017].

In essence, the *self-attention* (also known as the *intra-attention*) transforms every input vector to a weighted sum of the vectors in its neighbourhood, weighted by their *relatedness* to the input. One could illustrate this on the *euphony* in music, where every tone of a song relates to all of the precedent ones, to some more than to the others.

The full Transformer architecture is depicted in Figure 1.1.

[2]Which stands for "Česká Tisková Agentura", the original name of Czech News Agency

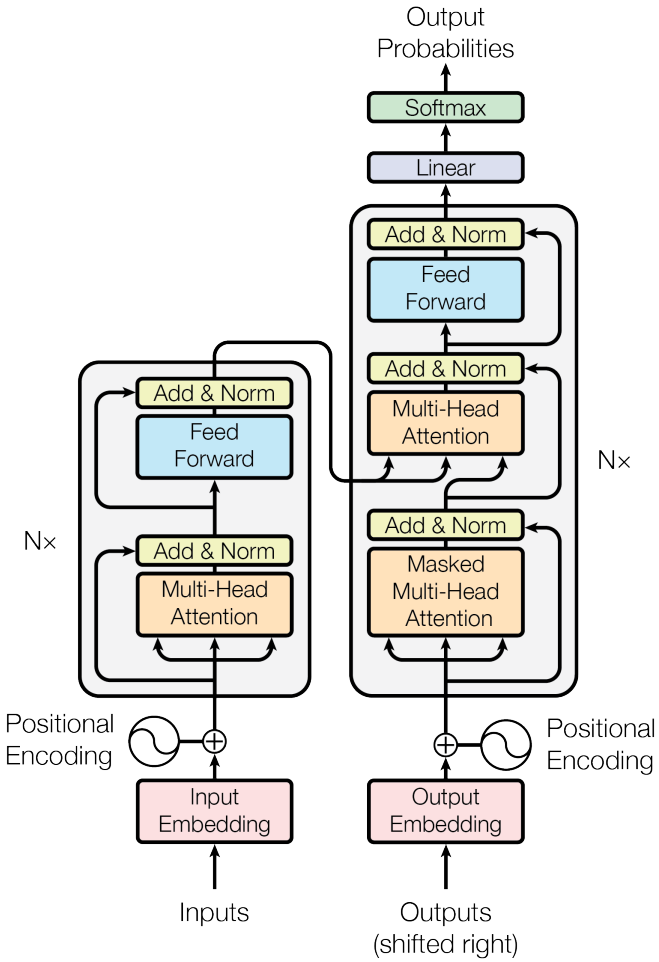

Figure 1.1: Transformer model architecture, reprinted from [Vaswani et al., 2017]

1.4 Thesis outline

Due to the bipartite nature of our thesis assignment, we have divided the chapters to follow into two parts. The **Part ??** presents our Czech datasets and the methods of their collection, and the **Part ??** makes the initial experiments for the NLI task.

- **Chapter 1** introduces the problem, motivates the research on the topic and sets up the challenges of this thesis
- **Chapter ??** examines the most relevant research in the field, with an emphasis on the methods of dataset collection, it introduces the two subsequent chapters on the topic
- **Chapter ??** lists and justifies our methods of generating the *localized dataset*, i. e. the methods of transferring the learning examples from a high-resource Natural Language to Czech
- **Chapter ??** describes our methods of collecting a novel fact-checking dataset using the non-encyclopædically structured knowledge base of ČTK news reports
- **Chapter ??** introduces the resulting dataset, as collected during three waves of annotation with Václav Moravec and students of the Faculty of Social Sciences

- **Chapter ??** briefly introduces the full fact-checking pipeline we have established with the FactCheck team at AIC using the collected data and a couple of real-world applications stemming from it
- **Chapter ??** explores the state-of-the-art methods of Natural Language Inference and their potential for our system, and it proceeds to make preliminary experiments on our dataset using these methods
- Finally, **Chapter ??** concludes the thesis, summarises the results we have achieved and proposes directions for future research

Bibliography

[Allcott and Gentzkow, 2017] Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–36.

[Barua et al., 2020] Barua, Z., Barua, S., Aktar, S., Kabir, N., and Li, M. (2020). Effects of misinformation on covid-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. *Progress in Disaster Science*, 8:100119.

[Brown et al., 2020] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. *CoRR*, abs/2005.14165.

[Buchanan and Benson, 2019] Buchanan, T. and Benson, V. (2019). Spreading disinformation on facebook: Do trust in message source, risk propensity, or personality affect the organic reach of "fake news"? *Social Media + Society*, 5(4):2056305119888654.

[Cheng et al., 2016] Cheng, J., Dong, L., and Lapata, M. (2016). Long short-term memory-networks for machine reading. *CoRR*, abs/1601.06733.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Dědková, 2021] Dědková, B. (2021). Multi-stage methods for document retrieval in the czech language.

[Gažo, 2021] Gažo, A. (2021). Algorithms for document retrieval in czech language supporting long inputs.

[Lazer et al., 2018] Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., and Zittrain, J. L. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

[Rýpar, 2021] Rýpar, M. (2021). Methods of document retrieval for fact-checking. `https://www.overleaf.com/read/thbvcjvvvfjp`. [Online; accessed 21-May-2021].

[Shu et al., 2018] Shu, K., Wang, S., Le, T., Lee, D., and Liu, H. (2018). Deep headline generation for clickbait detection.

[STEM, 2021] STEM (2021). Mýtům a konspiracím o covid-19 věří více než třetina české internetové populace | stem.cz. https://www.stem.cz/mytum-a-konspiracim-o-covid-19-veri-vice-nez-tretina-ceske-internetove-populace/. Accessed: 2021-05-03.

[Thorne et al., 2019] Thorne, J., Vlachos, A., Cocarascu, O., Christodoulopoulos, C., and Mittal, A. (2019). The FEVER2.0 shared task. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 1–6, Hong Kong, China. Association for Computational Linguistics.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.

Appendix A
Acronyms

BERT Bidirectional Encoder Representations from Transformers

FEVER Fact Extraction and Verification – series of Shared tasks focused on fact-checking

CLI Command-Line Interface

NLI Natural Language Inference

ČTK Czech Press Agency