



**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](https://fccheck.fel.cvut.cz)

Supervisor: **Ing. Jan Drchal, Ph.D.**

Field of study: **Informatics**

Subfield: **Natural Language Processing**

August 2023

Contents

1 Introduction	1
1.1 Motivation	2
1.2 Automated Fact Checking	2
1.3 A word on the Transformers	3
1.4 Dissertation minimum study outline	3
2 State of the Art	5
2.1 Pre-train + Fine-tune	5
2.1.1 BERT and derivatives	5
2.2 Large Language Models	5
2.2.1 GPT-3.5 and GPT-4	5
2.2.2 LLaMA-2 and derivatives	5
2.2.3 LoRA and other optimization	5
2.3 Fact checking approaches	6
2.3.1 FEVER and followups	6
2.3.2 Open-domain fact-checking	6
2.4 Claim generation	6
2.5 NLP Generative task benchmarking	7
2.5.1 BERTScore	7
2.5.2 AlignScore	7
3 Current Contribution	8
3.1 Datasets	8
3.1.1 CsFEVER	8
3.1.2 CTKFACTS	8
3.1.3 Other NLP datasets in West Slavic languages	8
3.2 Models	8
3.3 Publications	8
3.4 Applications	8
4 Dissertation plan	10
4.1 Current research agenda	10
4.1.1 Automated claim generation	10
4.1.2 Claim generation metrics	10
4.2 Data Collection	13
4.2.1 Human-in-the-loop grading of claim generators	13
4.2.2 Validation of the model outputs with human fact-checkers	13
4.2.3 Polish dataset scraping	13
4.2.4 Crowd-sourced fact checking platform	13
4.3 The grand scope	13
5 Conclusion	14
Bibliography	15
A Acronyms	19

1.1 A real world example of fact checking done by https://politifact.org	2
1.2 Automated fact-checking pipeline, reprinted from [Guo et al., 2022] . . .	3
1.3 Transformer model architecture, reprinted from [Vaswani et al., 2017] . . .	4
2.1 Proof-of-concept Czech fact-checking based on live-internet search (Bing API) and LLM prompting, based on the proposals of [Chen et al., 2023] in Czech, using a real-world claim that was fact-checked by demagog.cz in June 2023	6
3.1 FCheck – an open-source platform for fact-checking dataset collection I developed for TAČR project; collects data for claim generation, information retrieval and natural language inference tasks . . .	9
3.2 Factual claim extraction application done for the CEDMO project	9
4.1 LM-Critic – deciding text fluency viewed as finding local optima of Language Model output probability, reprinted from [Yasunaga et al., 2021]	11
4.2 A self-evaluating claim generation model based on GPT-4 [OpenAI, 2023] using the OpenAI API and a few-shot approach	12

Chapter 1

Introduction

My dissertation, as well my long-term research, centers around the field of *automated fact checking* through the means of Natural Language Processing (NLP) and its modern methods. The work consists of the analysis of the whole fact-checking process, its subdivision and simplification into tasks that can be efficiently addressed using the current state-of-the-art NLP methods, collection of data appropriate to benchmark such tasks, delivery of example solutions and their validation against similar research in other languages and related tasks.

The main focus of mine and of our research group are the fact-checking-related tasks in the West Slavic languages (Czech, Slovak and Polish) and secondarily in English. My contribution has so far been the collection and publication of novel datasets for the fact-checking task and its subroutines, models trained for the tasks and their debate, including the ongoing establishment of metrics that would explainably rate the model success and error rates in terms close to the human notion of *facticity* (which proves to be a challenge on its own, requiring another round of novel research [Koto et al., 2020, Wright et al., 2022]).

My doctoral aim is to cover every step on the path from gathering a factual claim – for example, extracting it from a political debate – to predicting its veracity verdict and justifying it rigorously with hard data. With the recent boom in NLP beginning with the advent of transformer networks and later the Large Language Models (LLMs) [Zhao et al., 2023], few-shot learning [Brown et al., 2020] and prompting [Liu et al., 2023] a significant part of the research is and has to be an appropriate and timely adoption of new ever-evolving sota NLP solutions, based on well-designed studies in our specific context.

Overall, my agenda is to follow up on my published research on fact checking in Czech with methods that reiterate on our results in other languages and evolving our previous methodology based on transformer *pre-training & fine-tuning* paradigm to a computationally feasible design based on LLMs, which are already exhibiting superiority tasks similar to ours [Chen et al., 2023] in English.

My recent focus within the whole grand fact-checking scheme is the step of *claim generation*, which I aim to establish among the other commonly benchmarked NLP tasks within the scientific community, adjacent to that of *abstractive summarization*. To benchmark the task, one would need a set of metrics that properly reflect phenomena such as *model hallucinations* – a common problem of modern day LLMs [Ji et al., 2023]. As the exact word-level metrics for NLP generative tasks do not correlate well with human judgement [Zhang* et al., 2020] and model-based metrics are hard to explain, my research also focuses on a delivery of a set of human-understandable model-based metrics.

The goal of this study is to show the directions I am taking to address these challenges, reasoning behind them, my research questions and current results that motivated them.



Figure 1.1: A real world example of fact checking done by <https://politifact.org>

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] and the Russo-Ukrainian war [Stănescu, 2022].

The recent advances in artificial intelligence have unwittingly contributed to the spread of misinformation on social media [Buchanan and Benson, 2019], as well as they hold a large potential for the false content generation [Sebastian, 2023].

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 (Figure 1.1) is a process of matching every information within a *factual 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 Automated Fact Checking

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

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



Figure 1.2: Automated fact-checking pipeline, reprinted from [Guo et al., 2022]

slower than that of generating false or misleading 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 3.2) [Guo et al., 2022]. This would reduce the fact-checker’s workload to mere adjustments of the proposed result and correction of mistakes on the computer side.

The goals of the ongoing efforts of FactCheck team at AIC CTU, are to explore and adapt the state-of-the-art methods used for fact verification or similar tasks in other languages, currate appropriate datasets for it and propose strong systems for such a task in Czech.

1.3 A word on the Transformers

For the past six 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 encoder [Devlin et al., 2019a] and its descendants, or the OpenAI’s GPT-3 decoder [Brown et al., 2020] and GPT-4 [OpenAI, 2023] that are used in the booming online AI service ChatGPT².

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 and successive ones, to some more than to the others.

The full Transformer architecture is depicted in Figure 1.3.

1.4 Dissertation minimum study outline

■ **Chapter 1** introduces the dissertation topic, motivates the research sets up our chal-

²<https://chat.openai.com>



Figure 1.3: Transformer model architecture, reprinted from [Vaswani et al., 2017]

lenges for the future research

- **Chapter 2** examines the most relevant research in the field and tries to highlight the recent paradigm shift from models trained for a single task to a single large models that perform well in everything
- **Chapter 3** explains our current contributions to the field of automated fact-checking and NLP in Czech
- **Chapter 4** describes our plan for the dissertation and justifies the directions we are taking
- Finally, **Chapter 5** concludes the study with a wrapup of its findings

Chapter 2

State of the Art

This chapter will first describe the originally popular models for NLP such as BERT and the recent paradigm shift from *pre-train+fine-tune* transfer learning framework popular since the original [Devlin et al., 2019b] paper to the currently booming LLMs which often outperform the smaller models even without the fine-tuning step [OpenAI, 2023, Touvron et al., 2023a, Vicuna, 2023]. We will then take a look at the performance optimization methods that enable training multi-billion parameter pre-trained models on a set of task-specific data on a single GPU and their potential for our research.

2.1 Pre-train + Fine-tune

Surprisingly well performing paradigm for small data, vague tasks

2.1.1 BERT and derivatives

Simple training task, scalable architecture, potential to evolve into something grand, czech, herbert, slovakbert

...Which currently arrives in form of:

2.2 Large Language Models

2.2.1 GPT-3.5 and GPT-4


Obscure, paid [Brown et al., 2020, OpenAI, 2023]

2.2.2 LLaMA-2 and derivatives

Open-source, freely usable. Often poor czech coverage Quantizace, 4b, 8b, zlomek parametrů [Köpf et al., 2023]

2.2.3 LoRA and other optimization

To be able to fine-tune multi-billion-parameter models such as LLaMA-2 [Touvron et al., 2023b] on a single TPU, multiple adjustments must be made to reduce the computational complexity. [Liu et al., 2022] [Hu et al., 2021]

 **Open-Domain Fact Checking (Bing API & GPT-3.5)**

This is a prototype CEDMO application to fact-check a real-world text.

Text to extract

Stráž se jednoznačně vyjádřil, že ministr spravedlnosti Pavel Blažek není prověřován.

Speaker

Eva Decroix

Affiliation


ODS

Date

20.08.2023

Clear


Submit

 **Částečně souhlasí**

Je Pavel Blažek ministr spravedlnosti?


[Pavel Blažek – Wikipedia](#)

Pavel Blažek (* 8. dubna 1969 Brno) je český právník, v letech 2012 až 2013 **ministr spravedlnosti ČR** ve vládě Petra Nečase a znovu od prosince 2021 **ministr spravedlnosti ČR** ve vládě Petra Fialy, v letech 2010 až 2014 místopředseda ODS, od října 2013 poslanec Poslanecké sněmovny PČR, v letech 2002 až 2014 zastupitel města ...

 **Souhlasí**

[Pavel Blažek | Vláda ČR - vlada.cz](#)

Pavel Blažek, ministr spravedlnosti. Pavel Blažek se narodil v Brně, kde vystudoval Právnickou fakultu Masarykovy univerzity. V letech 1992–2002 působil jako interní učitel na Právnické fakultě MU v Brně, **obor správní právo a správní věda**.

 **Úplný výmysl**

Je Pavel Blažek prověřován?

[Pavel Blažek - Wikipedia](#)

Prezident republiky jej členem Nečasovy vlády jmenoval 3. července 2012 v 10 hodin. Obvinění z trestného činu Dne 31. července 2012 publicista Jan Urban a obecně prospěšná organizace Kverulant.org zveřejnili prohlášení, ve kterém obvinili Pavla Blažka ze spáchání trestného činu podle § 209 **trestního zákoníku** – podvodu.


 **Částečně souhlasí**

Figure 2.1: Proof-of-concept Czech fact-checking based on live-internet search (Bing API) and LLM prompting, based on the proposals of [Chen et al., 2023] in Czech, using a real-world claim that was fact-checked by demagog.cz in June 2023

2.3 Fact checking approaches

2.3.1 FEVER and followups

Yields interesting benchmark with statistically quantifiable model success, oversimplifies the problem, as it uses Wikipedia for a trusted knowledge base and only reasons based on a data from a fixed period of time, focusing on “atomic” claims that do not match the complexity of real-world factoids.

2.3.2 Open-domain fact-checking

This paper with Bing for example uses the whole internet, but is that really what we want? Like, every lie can be backed with an internet – at the end of the day you do need to draw the line of what to trust somewhere, which directly conflicts this design.

2.4 Claim generation

- Approaches such as QACG exploit Question Answering
- The task of extreme summarization (XSum) focuses on summarizing a long body of text into a single sentence, focusing on its most relevant aspects and facts

- CLEF-CheckThat postulates the task of classifying *Checkworthiness* of different parts of a long texts, such as a political debate

2.5 NLP Generative task benchmarking

2.5.1 BERTScore

2.5.2 AlignScore

Chapter 3

Current Contribution

We have collected novel data, emulated and scraped inavailable datasets making them public or readying them for doing so, we have established numerous sota models and are currently working on establishing the topic of claim generation as a summarization-related NLP task. We are also readying metrics for fact-checking, experimenting with them and so on and soforth.

3.1 Datasets

3.1.1 CsFEVER

3.1.2 CTKFACTS

3.1.3 Other NLP datasets in West Slavic languages

1. **Translated NLI datasets** – SNLI, ANLI, MultiNLI,
2. SmeSum, CTKSum, CsFEVERSum
3. Polish summarization data

3.2 Models

3.3 Publications

3.4 Applications

Here we will show off the demonstration tools, as well as our open-source platform <https://fcheck.fel.cvut.cz> and currently running claim extraction tools.

Tvorba tvzení (Ú1a)

Zlatá pravidla extrakce tvzení

- Přidat pravidla tvzení k síti, pokud je potřeba.
- Tržba jednoduchá pravidla tvzení vychází ze zdrojového odstavce, který má smysl odlišovat.
- Pokud to zdrojový odstavec neumožňuje, nebo se Vám zdá nejasný, nechte se ho Přeskočit

Zdrojový článek

Přiznání vyhlášení drobné akcionářské teprmy; dle 1988 korun za akci 08.09.2005 09:12

Zdrojový odstavec

Z tohoto odstavce a příslušného článku vychází při tvorbě tvzení o jedné z požadovaných entit. (Přiznání Duchek, dle 1988 korun, 26. září, 1988)

Znalostní rámec

Rozhodnutí název článku pro zobrazení část, která byla vybrána jako relevantní pro daný zdrojový odstavec.

Články ve znalostním rámci byly vybrány podle frekvence výskytu slovních spojeních mezi (jedna osoba, akcie, firm, akcie), nebo pomocí sémantického vyhledávání odstavců z původního článku.

Pravdivé tvzení

Snažte se vybrat přibližně 2 minuty tvzení 1 až 5 tvzení z tohoto zdrojového odstavce.

Výsledná tvzení odstavce koncem řádku (-/-).

Pokud není zdrojový odstavce použitelný, stiskněte tlačítko Přeskočit

Navržené tvzení

Přiznání zastupitelé rozhodli, že město vyhlásí téměř 700 drobných akcionářů.

Obměny tvzení (Ú1b)

Původní tvzení

Přiznání zastupitelé rozhodli, že město vyhlásí téměř 700 drobných akcionářů.

Obměny tvzení

- Cílem je stávk odtá 1 minuty při generování každé obměny tvzení.
- Pokud není možné obměnit vygenerovat, nechte textové pole prázdné.

Parafraze

Parafráze je obměna tvzení, která má stejný význam, ale je formulována jinak.

Nahrzení podobnou entitou nebo vztahem

Nahrzení entitu, vztah nebo oběj jakoukoli entitou či vztahem. Vyberte se parafrázování původního tvzení.

Nahrzení odlišnou entitou nebo vztahem

Nahrzení entitu, vztah nebo oběj jakoukoli entitou či vztahem. Vyberte se parafrázování původního tvzení.

Zužení

Změňte formulaci tak, aby bylo nové tvzení specifické. Příklad: "všechny akcionáře" nahraďte "všechny akcionáře".

Zobecnění

Změňte formulaci tak, aby bylo nové tvzení obecnější. Příklad: "všechny akcionáře" nahraďte "všechny akcionáře".

Negace

Vyvoďte negaci původního tvzení. Vyvarujte se negace pomocí jednoduchého předání záporné.

Anotace správnosti vlastního tvzení (Ú2a)

Tvzení

08.09.2005 09:12 Zastupitelé z Přiznání vyhlášení jakýchkoli drobných akcionářů.

Podmínka anotace

Sem můžete napsat informaci chybějící v úplnosti důkazu.

Důkazy potvrzující/vyvracející tvzení

Zdrojový článek: Přiznání vyhlášení drobné akcionářské teprmy; dle 1988 korun za akci 08.09.2005 09:12 **Důkaz #1 #2**

PLZEN & září (ČTK) - Přiznání zastupitelé dnes rozhodli, že město vyhlásí téměř 700 drobných akcionářů. Přiznání teprmy (PT). Ti vlastní téměř 9700 akcií firmy, což představuje 0,82 procenta základního jmění. Za jeden cenový papír dostanou 1988 korun. Potvrdit to musí ještě mimořádná valná hromada 26. září. ČTK to řekl ekonomický náměstek primátora Vladimír Duchek.

„Item přiznání, že akcionář dostává za své akcie přiměřené protiplnění a opírá se o posudek soudního znalce“ uvedl Duchek. Město uhradí drobným vlastníkům zhruba 20 milionů korun, což není z hlediska rozpočtu žádný problém.

Ovšem PT koresponduje se zadáním zastupitelstva z roku 1996 - trvale zvyšovat podíl města v PT. Od července to navíc podpořila novela obchodního zákoníku. Umožňuje, aby vlastníci alespoň 90 procent základního kapitálu firmy mohli na valné hromadě rozhodnout, že vyplatí drobné akcionáře.

Výkupní cena velkých neobchodovatelných listovních akcií PT na město je podle Duchka vyšší, než za kolik město vykupovalo loni v říjnu 16 procent akcií PT od nímehodně energetického koncernu E.ON, a o 400 korun vyšší než cena, za níž město jako hlavní akcionář loni vykupoval cenové papíry po stažení z veřejných trhů.

„Něsem vůbec nadšen. Mám akcie na delší dobu, a pokud taková firma udělá zisk 300 korun na akci, tak je to zhruba investice, která je jiné nedostane“ řekl Vít Ryjáček, vlastník téměř 1500 akcií PT. Navíc přiznaná miliónová investice do kolie na komasu capital akcionářů zvýšení částky zisku, který loni činil 258 milionů korun a meziročně se zvýšil o 28 procent.

Znalostní rámec: Voda v Přiznání od ledna nezdrží, město chce státní peněz z nájmů 08.09.2005 09:12

Znalostní rámec: Stock Přiznání vyhlášení drobné akcionářské teprmy; dle 1988 korun za akci 08.09.2005 09:12

Znalostní rámec: Přiznání teprmy; dle 1988 korun za akci 08.09.2005 09:12

Znalostní rámec: Město vyhlásí v posledních letech své rozpočty prodávám akci 08.09.2005 09:12

Znalostní rámec: Vodárna Přiznání loni snížila ztrátu při rozvodu pitné vody 08.09.2005 09:12

Figure 3.1: FCheck – an open-source platform for fact-checking dataset collection I developed for TAČR project; collects data for claim generation, information retrieval and natural language inference tasks

Central European Digital Media Observatory

Factual Claim Extraction

This is a prototype CEDMO application to extract factual claims from an arbitrary text.

Text to extract

He said this became possible after Kyiv liberated the "strategically important" village of Robotyne in the Zaporizhzhia region. Russia's military says its forces are still holding on to the village. Ukraine seeks to cut the land corridor to Crimea, which Russia seized in 2014. By doing so, Kyiv would split in two the land Russia occupies in southern Ukraine, making Moscow's supply lines more complicated. Speaking at a recent meeting of French ambassadors in Paris, Mr Kuleba said: "Having entrenched on its [Robotyne's] flanks, we are opening the way to Tokmak and, eventually, Melitopol and the administrative border with Crimea."

Model

- EN: LLaMa-2 (BBC)
- CS: MBart (CTK, Sumeczech+CNC)

Max number of claims

2

Amount of randomness

11

Clear **Submit**

Claims

```
[
  0: "Ukraine's foreign minister has said his country is "on the verge" of liberating the southern city of Melitopol from Russian control." ,
  1: "Ukraine is seeking to reopen a land corridor to Crimea in its south, Kyiv's foreign minister Dmytro Kuleba has said."
]
```

Flag

Figure 3.2: Factual claim extraction application done for the CEDMO project

Chapter 4

Dissertation plan

4.1 Current research agenda

4.1.1 Automated claim generation

TODO:

4.1.2 Claim generation metrics

The common problem with generative tasks in NLP is that of explaining model reasoning in human-understandable manner and troubleshooting the prediction faults, such as the *model hallucination*.

For the task of claim generation, where we also face the challenge of the *relevance* of the information extracted by the model, we postulate the following metrics:

1. **Fluency** – *is the claim grammatically correct and intelligible?*

Currently, we are working with two emulations of claim fluency, challenge that is similar to a standard NLP task of Gramatical Error Detection (GED): LM-Critic (Figure) [Yasunaga et al., 2021] perturbs the claim words and characters to find local optima in output probability of its tokens, using a language model such as GPT-2 as its reference. GPTScore [Fu et al., 2023] uses prompting a LLM (such as GPT-3.5) to obtain a model-inferred score using few- or zero-shot learning.

Both can be adapted for Czech and the latter is demonstrated in Figure 4.2.

2. **Decontextualization** – *can the claim be correctly interpreted without any additional context from the source document or elsewhere?*

A common problem with machine-extracted factual claims is reusing excerpts from source document along with inexplicable contextual pronouns (“President won’t sue them”) and relative referencing (“*Last year*, CTU had 23K students”).

[Choi et al., 2021] proposes decontextualization as a sequence to sequence task with two texts on input (s, c) – sentence and context. T5 model [Raffel et al., 2019] is then trained on machine-generated gold data from Wikipedia to output sentence s' such that the truth-conditional meaning of s' in an empty context is the same as that of s in c .

[Mohri et al., 2023] improves upon this, altering the problem formulation to minimizing surrogate loss, rejecting with a fixed predictor and claiming to get as close as $\sim 3\%$ away from the theoretical limit for the task.

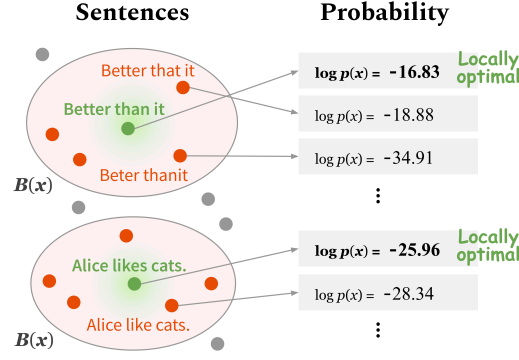


Figure 4.1: LM-Critic – deciding text fluency viewed as finding local optima of Language Model output probability, reprinted from [Yasunaga et al., 2021]

The approaches are reproducible using Czech Wikipedia corpus and appropriate for further examination.

3. **Atomicity** – does the claim describe a single entity, relation or process?

can be checked using the Relationship Extraction methods such as LUKE [Yamada et al., 2020]. To put it simply, the RE task is to identify the entities of a text (persons, institutions,...) and the relations between them (such as (“study at”, Herbert, CTU)). The atomicity evaluation can be converted to a RE task by attempting to extract such fact triples and mark the claim as atomic if there is at most one such triple found (after removing symmetries)

4. **Faithfulness** – does the claim only contain information that is consistent with the source document?

This metric is crucial to pinpoint *model hallucinations* – parts of claim where the model outputs stray from the information present in source text and begin to just “make stuff up”. We proceed to use two alternative metrics – a score proposed within the FFCI evaluation framework [Koto et al., 2020] as:

$$\text{AvgTop-}n_{s_j \in X, t_i \in Y'}(\text{BERTSCORE}(t_i, s_j))$$

Where the AvgTop- n simply averages across the top n (say, 5) highest scores, X, Y' are the sentences of source document and output (so, in the claim generation scenario $|Y'| = 1$) and BERTSCORE [Zhang* et al., 2020] is a recently popular similarity score between two sentences that doesn’t compare the texts on a verbatim level (like e.g., ROUGE [Lin, 2004] which correlates poorly with human judgement) but expresses the sentence similarity as a sum of cosine similarities between their tokens’ embeddings – this should capture semantical relations rather than the word-for-word similarity, which could be beneficial in highly inflected languages such as Czech.

Similar metric called ALIGNSCORE was proposed in [Zha et al., 2023], looking for optimum alignment of output and input parts, in terms of a RoBERTa model [Liu et al., 2019] trained to detect inconsistencies on 4.7M training examples adapted from various tasks (inference, question answering, paraphrasing,...) and while relatively small (355M parameters), outperforms metrics based on GPT-4 that is orders of magnitude larger.

Empirically, the models work encouragingly well on spotting hallucinations and inconsistencies in English, and while the transduction of BERTSCORE is trivial, using

Central European Digital Media Observatory

Claim extraction and its metrics using GPT

This is a prototype CEDMO application to fact-check a real-world text.

Text to extract

A global search has been launched to find one of the world's most iconic instruments - Paul McCartney's original Höfner bass guitar. McCartney bought the instrument for £30 (\$38) in Hamburg, Germany, in 1961, but it disappeared eight years later. The hunt began after McCartney urged manufacturers Höfner to track down his beloved instrument. The bass features in The Beatles' music of those years, including the hits Love Me Do and She Loves You. Nick Wass is heading Höfner's search project and has joined forces with two journalists in trying to solve the "greatest mystery in the history of rock and roll."

Number of claims to extract

3

Type of text

news article

Model

gpt-3.5-turbo

Clear Submit

Claims

1. Paul McCartney's original Höfner bass guitar, which he bought for £30 in 1961, has been missing since 1969.
2. McCartney has urged Höfner, the manufacturer of the instrument, to launch a global search to find his beloved bass guitar.
3. Nick Wass, along with two journalists, is leading Höfner's search project to solve the "greatest mystery in the history of rock and roll."

Multi-metrics

- **focus = 5** All three claims accurately represent the relevant information presented in the news article, which is about the global search effort launched to find Paul McCartney's missing Höfner bass guitar and the involvement of Nick Wass and two journalists in this project.
- **coverage = 3.5**, since the claims cover the main information regarding the missing bass guitar and the search project, but do not mention the role of the two journalists nor the significance of the instrument to The Beatles' music.

Claim metrics

1. *Paul McCartney's original Höfner bass guitar, which he bought for £30 in 1961, has been missing since 1969.*
 - **fluency = 5** The claim is grammatically correct and well-formed, presenting a factual statement about the missing original Höfner bass guitar that Paul McCartney bought in 1961 for £30.
 - **decontextualization = 4.5** The claim provides the key information about the missing Höfner bass guitar, including its ownership, purchase date, and disappearance, but lacks details about the ongoing global search and the efforts involved.
 - **atomicity = 5** This claim is highly atomic as it presents a single specific fact about Paul McCartney's missing bass guitar, including the purchase date, price, and the year it went missing.
 - **faithfulness = 4.5** The claim is mostly faithful to the news article. The article states that Paul McCartney's original Höfner bass guitar has been missing since 1969, and the hunt to find it began after McCartney urged Höfner to track it down. However, the claim does not mention the involvement of Nick Wass and two journalists in the search project, which slightly deviates from the article.

Figure 4.2: A self-evaluating claim generation model based on GPT-4 [OpenAI, 2023] using the OpenAI API and a few-shot approach

a Czech embedding model such as CZERT [Sido et al., 2021] or FERNET [Lehečka and Švec, 2021], reproducing the success of ALIGNSCORE will require more research and data.

5. **Focus@k** – if we generate k claims using this model, what will be the proportion of gold (relevant) information among all the information listed in the generated claims?

The metric is analogous to the concept of *precision* in the common machine learning applications, however, its deciding gets more ambiguous in the natural language settings, where we are dealing with synonyms and endless number of possible wordings for every piece of information.

An elegant and functional perspective on the problem has been brought around in QAGS¹ evaluation protocol [Wang et al., 2020], where the idea is to use a Question Generation model (QG) to formulate questions in natural language based on all k predicted claims. The questions are then twice answered using a Question Answering (QA) model, giving it knowledge from (i.) the predicted claims (ii.) the gold claims written by a human. The focus is then defined as the proportion of questions with the same answers extracted from the gold and predicted claims, among all questions model can generate from the predicted claims.

6. **Coverage@k** – if we generate k claims using this model, what proportion of gold (relevant) information from the source text will be covered?

¹Pronounced “kags”, stands for “Question Answering and Generation for Summarization”

Anologous to *recall@k* in general machine learning, QAGS proposes to generate questions using gold claims and try to answer them using the predicted claims, much like in the *focus* scenario, but vice versa.

4.2 Data Collection

4.2.1 Human-in-the-loop grading of claim generators

4.2.2 Validation of the model outputs with human fact-checkers

4.2.3 Polish dataset scraping

Will be first of its kind for NLP purposes

4.2.4 Crowd-sourced fact checking platform

TODO: Cite *boys* [Bútorá, 2023]

4.3 The grand scope

1. Claim extraction metrics proposal based on factuality of summarization
2. Claim extraction paradigm that benchmarks best in the newly given metrics
3. Systems for NLI built on top of LoRA paradigm to score best in the task, as showed promising by Daniil



Chapter 5

Conclusion



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.
- [Bútorá, 2023] Bútorá, R. (2023). Crowd-sourcing platform frontend for fact-checking. <https://dspace.cvut.cz/handle/10467/109505>.
- [Chen et al., 2023] Chen, J., Kim, G., Sriram, A., Durrett, G., and Choi, E. (2023). Complex claim verification with evidence retrieved in the wild.
- [Cheng et al., 2016] Cheng, J., Dong, L., and Lapata, M. (2016). Long short-term memory-networks for machine reading. *CoRR*, abs/1601.06733.
- [Choi et al., 2021] Choi, E., Palomaki, J., Lamm, M., Kwiatkowski, T., Das, D., and Collins, M. (2021). Decontextualization: Making sentences stand-alone. *Transactions of the Association for Computational Linguistics*, 9:447–461.
- [Devlin et al., 2019a] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019a). Bert: Pre-training of deep bidirectional transformers for language understanding.
- [Devlin et al., 2019b] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019b). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

- [Fu et al., 2023] Fu, J., Ng, S.-K., Jiang, Z., and Liu, P. (2023). Gptscore: Evaluate as you desire.
- [Guo et al., 2022] Guo, Z., Schlichtkrull, M., and Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.
- [Hu et al., 2021] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.
- [Ji et al., 2023] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- [Koto et al., 2020] Koto, F., Baldwin, T., and Lau, J. H. (2020). Ffci: A framework for interpretable automatic evaluation of summarization.
- [Köpf et al., 2023] Köpf, A., Kilcher, Y., von Rütte, D., Anagnostidis, S., Tam, Z.-R., Stevens, K., Barhoum, A., Duc, N. M., Stanley, O., Nagyfi, R., ES, S., Suri, S., Glushkov, D., Dantuluri, A., Maguire, A., Schuhmann, C., Nguyen, H., and Mattick, A. (2023). Openassistant conversations – democratizing large language model alignment.
- [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.
- [Lehečka and Švec, 2021] Lehečka, J. and Švec, J. (2021). Comparison of czech transformers on text classification tasks. In Espinosa-Anke, L., Martín-Vide, C., and Spasić, I., editors, *Statistical Language and Speech Processing*, pages 27–37, Cham. Springer International Publishing.
- [Lin, 2004] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- [Liu et al., 2022] Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., and Raffel, C. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.
- [Liu et al., 2023] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).
- [Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- [Mohri et al., 2023] Mohri, C., Andor, D., Choi, E., and Collins, M. (2023). Learning to reject with a fixed predictor: Application to decontextualization.
- [OpenAI, 2023] OpenAI (2023). Gpt-4 technical report.

- [Raffel et al., 2019] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.
- [Sebastian, 2023] Sebastian, G. (2023). Exploring ethical implications of chatgpt and other ai chatbots and regulation of disinformation propagation.
- [Sido et al., 2021] Sido, J., Pražák, O., Příbáň, P., Pašek, J., Seják, M., and Konopík, M. (2021). Czert – czech bert-like model for language representation.
- [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.
- [Stănescu, 2022] Stănescu, G. (2022). Ukraine conflict: the challenge of informational war. *SOCIAL SCIENCES AND EDUCATION RESEARCH REVIEW*, 9(1):146–148.
- [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.
- [Touvron et al., 2023a] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023a). Llama: Open and efficient foundation language models.
- [Touvron et al., 2023b] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. (2023b). Llama 2: Open foundation and fine-tuned chat models.
- [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.
- [Vicuna, 2023] Vicuna (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. <https://vicuna.lmsys.org/>. Accessed: 2023-09-04.
- [Wang et al., 2020] Wang, A., Cho, K., and Lewis, M. (2020). Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

- [Wright et al., 2022] Wright, D., Wadden, D., Lo, K., Kuehl, B., Cohan, A., Augenstein, I., and Wang, L. L. (2022). Generating scientific claims for zero-shot scientific fact checking. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2448–2460, Dublin, Ireland. Association for Computational Linguistics.
- [Yamada et al., 2020] Yamada, I., Asai, A., Shindo, H., Takeda, H., and Matsumoto, Y. (2020). Luke: Deep contextualized entity representations with entity-aware self-attention.
- [Yasunaga et al., 2021] Yasunaga, M., Leskovec, J., and Liang, P. (2021). LM-critic: Language models for unsupervised grammatical error correction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7752–7763, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- [Zha et al., 2023] Zha, Y., Yang, Y., Li, R., and Hu, Z. (2023). Alignscore: Evaluating factual consistency with a unified alignment function.
- [Zhang* et al., 2020] Zhang*, T., Kishore*, V., Wu*, F., Weinberger, K. Q., and Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- [Zhao et al., 2023] Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. (2023). A survey of large language models.



Appendix A

Acronyms

BERT Bidirectional Encoder Representations from Transformers

GPT Generative Pre-trained Transformer

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