University of Ljubljana
Faculty of Computer and
Information Science

FRI Natural language processing course 2024
Project report

Natural language inference dataset

Lea Cvetko, Arja Hojnik, and Katarina Plementaš

Abstract

This project aimed to construct and evaluate a Slovenian Natural Language Inference (NLI) dataset based on the Gigafida corpus, utilizing Large Language Models (LLMs) to generate pairs of paragraphs exhibiting different relationships (entailment, neutral, contradictory). In our evaluation, we compared performance metrics between ChatGPT-3.5 and ChatGPT-4, both before and after manual corrections, using key indicators such as entailment, neutrality, contradiction, and overall accuracy. Our statistical analysis revealed that ChatGPT-4 generally outperforms ChatGPT-3.5, demonstrating higher mean values across various metrics. Additionally, ChatGPT-4 exhibited lower standard deviations in some categories, indicating more consistent performance compared to ChatGPT-3.5. Thus, the upgraded version of the ChatGPT model shows a clear improvement in performance.

Keywords

Natural Language Processing (NLP), Natural Language Inference (NLI), ChatGPT, GigaFida, Slovenian Language

Advisor: Aleš Žagar

Introduction

Natural Language Processing (NLP) continually evolves to bridge the gap between human language understanding and machine interpretation. Despite significant advancements, challenges remain, notably in processing less-resourced languages like Slovenian. This project focuses on constructing a Slovenian Natural Language Inference dataset, a critical resource for testing machine understanding of text in entailment, contradiction, and neutrality contexts. With an interest in making AI systems more accessible and effective across diverse linguistic landscapes, our project aims to push the boundaries of what's currently achievable with Slovenian language processing.

By utilizing Large Language Models (LLMs) to generate text pairs, we aim to produce a diverse and challenging dataset that reflects the complexity of natural language reasoning. Recognizing the foundational work done on the Slovenian Natural Language Inference (SI-NLI) dataset, we embark on an ambitious initiative to further enrich this dataset.

Related Work

Current approaches to NLI rely heavily on datasets such as SNLI and MultiNLI, which are predominantly in English. While these datasets have driven significant advancements in NLP, the reliance on English limits the applicability of derived models to other languages. Even though recent efforts, like XNLI (The Cross-Lingual NLI Corpus), attempt to bridge this gap by providing multilingual extensions, the quality and quantity of data for Slovenian remain insufficient. On the LLM front, models like GPT-3 and BERT also show promising results in understanding and generating natural language.

Dataset

For the purpose of our project, we chose to work with the GigaFida (ccGigaFida 1.0) corpus, which is the largest corpus of written Slovene. It contains texts from newspapers, magazines, books, and web publications, which is why it presents a comprehensive resource that reflects contemporary usage of Slovene, making it valuable for generating text passages that capture current linguistic trends. This rich linguistic foundation is crucial for developing a Slovenian Natural Language Inference dataset aimed at enhancing machine understanding of textual entailment, neutrality, and contradiction.

Methodology

In this section, we focus on the methods used for developing such dataset.

1. **Preprocessing:** We chose to download the corpus encoded in TEI-like format with annotations in Slovenian, for easier processing in Python. It consists of 39,427 .xml files,

each containing one excerpt from various different fields of public interest. Then we performed some basic preprocessing. The text in original files is written in the form of lemmas, so the first step included extracting relevant textual content, ignoring metadata, headers and footers. We also removed non-textual elements and any unnecessary whitespace and saved the processed paragraphs into a new folder. Upon realizing that the whole process takes quite a lot of time and RAM, we decided to fix the code and divide the documents into three sections, with each of the members analyzing around 13,000 files.

Document Similarity: To make our dataset more representative, we decided to identify the most common themes that appear in the corpus and then extract the paragraphs from there. Keep in mind that the code was implied three separate times, as we divided the corpora into three sections, so the search for the common themes was also carried out three times (once for each section). The process included removing punctuation, numbers, and common stopwords (commonly used words that may not add significant semantic value to the text). First, we tried to make our own list of stopwords, but later on realized it is better to use a pre-set collection of stopwords integrated in the Python environment because our own set was too large and made it more difficult to use the dataset for further analysis. We also put the words into their basic form (lemmatization). Then we defined a term-frequency-inverse document frequency (TF-IDF), which helped us determine the importance of specific words in the document. We decided to use K-means clustering, to divide the documents into 4 clusters with the help of previously calculated TF-IDF measure. For better visualization, see the analysis of one cluster below. The specific words that appeared in said cluster, show that Cluster 2 focused on the topic of sports.

Cluster 2: *nizozemci bežijo reprezentance amsterdam mesecev nizozemski hitrostni zoi v naganu osvojili*

2. **Paragraph Selection:** We have determined that the optimal approach is to manually review all clusters and select the 13 most representative paragraphs from each cluster. Each team member generated 50 samples, with each sample consisting of an original paragraph, followed by three corresponding paragraphs, each illustrating a clear relationship with the original (entailment, neutral, contradictory). These corresponding paragraphs were produced using creatively designed prompts, which were subsequently input into the large language model.

3. **Designing Prompts for LLMs:** For each selected paragraph, we crafted tailored prompts designed to guide LLMs in generating hypotheses that are either entailed by, neutral to, or contradict the paragraph's content. This creative step is vital for ensuring the generated hypotheses are relevant and diverse. We iteratively refined the creative prompts multiple times to ensure the relationships were unequivocal and left no room for ambiguity. We then employed ChatGPT-4 to produce hypotheses. See example for creating a contradictory paragraph to one of the examples below:

"Napiši paragraf, ki opisuje, kako je avantgarda šestdesetih let močno cenila zgodovinske umetniške prakse in kako je ta zavezanost preteklosti vplivala na naslednjo generacijo mladih komponistov. Opisujte, kako so ti mladi umetniki našli navdih v tradicionalnih tehnikah in zvočnih materialih, ki so jih raziskovali njihovi predhodniki, kar je privedlo do renesanse klasičnih stilov v sedemdesetih letih. Poudarite, kako je ta kontinuiteta tradicije in spoštovanje preteklosti privedlo do večjega zanimanja in vrednotenja klasične glasbe med mlajšimi generacijami, kar je postopoma omililo potrebo po nenehnem iskanju novih tehnik ali radikalnih idej."

4. **Manual Review and Annotation:** Each NLI pair was then manually reviewed by all three members to validate the logical relationship between the original paragraph and the outputs for each of the relationships that ChatGPT provided. This step ensures the accuracy and reliability of the dataset, with multiple reviewers involved to mitigate subjectivity.

5. **Dataset Compilation:** After review and annotation, the pairs were compiled into a comprehensive dataset, which can be accessed through our repository. We strived for a balanced representation of logical relationships, ensuring the dataset's utility across various NLP applications.

7. **Preliminary Evaluation:** Finally, we conducted an evaluation of the dataset's effectiveness by training baseline NLI models, ChatGPT-3.5 and ChatGPT-4. Our primary objective in employing these two models was to conduct a comparative analysis between an older version and its newer, upgraded counterpart, in order to determine which one demonstrates superior performance. To avoid the models predicting the relationship between paragraphs through repetition of the same order, we randomly selected one relationship for each paragraph and queried both models on the nature of their relationship with the original paragraph. Based on their responses, we created two tables, one for each model, where results were recorded. When the model correctly identified the relationship, we noted a 1 for True, and when the model failed to identify the relationship, we recorded a 0 for False. This approach enabled us to generate a statistical report, quantifying the number of correctly predicted relationships and determining which model performed better.

For ChatGPT-4, we aimed for a perfect score. Consequently, when this model incorrectly identified a relationship, we revised the specific prompt to make the relationship more explicit. Table 3 presents the results after these corrections. However, for ChatGPT-3.5, we did not revise the prompts where mistakes occurred, as we anticipated this model to perform slightly worse from the outset.

Results

The group's initial expectations were that ChatGPT-4 would generally outperform its predecessor, given its advanced architecture and suitability for such tasks. However, ChatGPT-3.5 surprised us in many ways. First, let us examine the statistics from both models.

Phasellus in ligula nunc. Vivamus sem lorem, malesuada sed pretium quis, varius convallis lectus. Quisque in risus nec lectus lobortis gravida non a sem. Quisque et vestibulum sem, vel mollis dolor. Nullam ante ex, scelerisque ac efficitur vel, rhoncus quis lectus. Pellentesque scelerisque efficitur purus in faucibus. Maecenas vestibulum vulputate nisl sed vestibulum. Nullam varius turpis in hendrerit posuere.

Table 1. ChatGPT-3.5 statistics

Metric	Entailment	Neutral	Contradictory	Accuracy
Count	52.0	54.0	50.0	152.0
Mean	0.628	0.528	0.866	0.707
Std Dev	0.465	0.499	0.338	0.455
Min	0.0	0.0	0.0	0.0
Median	1.0	1.0	1.0	1.0
Max	1.0	1.0	1.0	1.0

ChatGPT-3.5 has shown relatively high performance in identifying contradictory relationships (86.8However, it is noteworthy that even if ChatGPT-3.5 struggled the most with identifying neutral relationships, it still exceeded our expectations by correctly identifying several such instances, where interestingly, these same examples posed challenges for ChatGPT-4, necessitating adjustments to the creative prompts to ensure accurate identification. See ChatGPT-4 statistics below.

Table 2. ChatGPT-4 statistics

Metric	Entailment	Neutral	Contradictory	Accuracy
Count	52.0	54.0	50.0	152.0
Mean	0.981	0.669	0.962	0.876
Std Dev	0.139	0.465	0.184	0.327
Min	0.0	0.0	0.0	0.0
Median	1.0	1.0	1.0	1.0
Max	1.0	1.0	1.0	1.0

ChatGPT-4 shows very high performance in identifying entailment (98.1The overall accuracy of 87.6

Model Comparison

Even though ChatGPT-3.5 surprised us in some aspects, our initial expectation that it will perform worse than ChatGPT-4 was quickly confirmed. The newer model consistently outperforms its predecessor in identifying entailment, neutral, and contradictory relationships, while also the overall accuracy of ChatGPT-4 is significantly higher. ChatGPT-4 demonstrates more consistent performance, especially in identifying entailment and contradictory relationships. The counts are consistent across both versions, indicating similar sample sizes. The mean values for ChatGPT-4 are generally higher than those for ChatGPT-3.5, suggesting an enhancement in performance metrics. The standard deviations are lower for ChatGPT-4 in some categories, signifying more consistent performance. Although ChatGPT-4 demonstrates superior overall performance, neutral relationships continue to pose a challenge for both models, highlighting an area for further improvement. ChatGPT-4 seems to aim for a more sophisticated analysis when giving an answer, which often makes ChatGPT-3.5 more accurate at handling neutral relationships since it doesn't delve as deeply. The newer model often tries to correct itself when asked if it's sure about its answer, but we wanted it to be fully confident in its choices. That's why we decided to correct the prompts that were challenging for ChatGPT-4.

ChatGPT-4 Corrections

To ensure the ChatGPT-4 model achieved complete accuracy in identifying relationships, we refined those creative prompts that posed challenges. Given that the model struggled most with neutral relationships, this refinement often involved instructing the model to generate a paragraph entirely unrelated to the original topic, thus describing a completely different subject.

Here is a brief example of a paragraph where the model incorrectly identified it as having an entailment relationship with the original, while it should have been classified as neutral.

Example 1: Original Paragraph

> Vloga Vere Drake v istoimenskem filmu, ženske, ki v petdesetih letih minulega stoletja opravlja nezakonite splave, je popolnoma spremenila življenja 49-letne Britanke Imelde Stauton, ker je bila doslej predvsem gledališka igralka. Po Benetkah, kjer je film zmagal in je Imelda dobila pokal volpi za najboljšo igralko, je bila nominirana še za zlati globus in oskarja. Znašla se je v središču pozornosti, novinarji so jo obletavali z vseh strani.

Example 2: Incorrectly Marked Paragraph (Entailment Instead of Neutral)

> Med snemanjem filma ”Vera Drake” je Imelda Staunton svoj delovni dan začela z zgodnjim prebiranjem scenarija, da bi popolnoma razumela globine in subtilnosti svojega lika. Sodelovala je v tesnih pogovorih z režiserjem, da bi uskladila svoje interpretacije in namere z vizijo filma, ter aktivno sodelovala z drugimi igralci, da bi izboljšala medsebojno kemijo in pristnost odnosov na zaslonu.

Example 3: Corrected Paragraph and Correctly Guessed Relationship (Neutral)

> Po uspehu filma ”Vera Drake” je Imelda Staunton doživela pomembno preobrazbo v svoji karieri. Njena vloga Vere Drake, ki ji je prinesla zmago na filmskem festivalu v Benetkah in nominacije za Zlati globus ter Oskarja, jo je izstrelila v središče mednarodne pozornosti. Ta prepoznavnost je Imeldi odprla vrata do novih priložnosti v filmski industriji, saj so ji ponudbe za vloge začele prihajati bolj pogosto in iz različnih koncev filmskega sveta, ne samo iz gledališkega okolja. Te nove vloge so bile bolj raznolike in zahtevne, kar ji je omogočilo, da je razširila svoj igralski repertoar. Njena zmaga v Benetkah in nadaljnje nominacije so ne le okrepile

njeno samozavest, temveč so ji tudi odprle vrata k sodelovanju v večjih filmskih projektih, kar je močno obogatilo njeno igralsko kariero.

Discussion

We observed that the performance of the models is highly dependent on the inputted creative prompt, necessitating careful tailoring for each model individually. Consequently, the group encountered difficulties in crafting prompts for both entailment and neutrality. Often, entailment was merely an abstract of the original paragraph, while neutral prompts incorrectly discussed the same event in detail and were consequently classified as entailment. However, once we developed a system and adhered to the SI-NLI guidelines, the process became more manageable. As mentioned initially, we included three corresponding paragraphs for each original paragraph in the dataset (entailment, neutral, and contradiction). Although only one was randomly chosen for use, this approach ensured a representative dataset and set the stage for any future project extensions.

Conclusion

Through this project, we constructed a Slovenian NLI dataset and conducted a thorough evaluation of two versions of ChatGPT models. Our findings indicate that ChatGPT-4 not only enhances mean performance metrics but also delivers more consistent results compared to ChatGPT-3.5. These improvements are evident both before and after manual corrections, highlighting ChatGPT-4's capability in handling natural language inference tasks. While ChatGPT-4 demonstrated an overall improvement in performance metrics, it also introduced challenges, particularly with neutral relationships, due to its tendency towards more sophisticated analysis.

Overall, while ChatGPT-4 shows promise with its advanced capabilities, it requires carefully tailored prompts to fully leverage its potential. These findings underscore the importance of prompt engineering in natural language processing and pave the way for future improvements and extensions in our projects.

References

1. Bowman, S. R., Angeli, G., Potts, C., Manning, C. D. (2015). A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326.
2. Klemen, M., Žagar, A., Čibej, J., Robnik Šikonja, M. (2022). Označevalne smernice: Ustvarjanje slovenske množice SI-NLI za sklepanje o pomenskem sosledju besedil. CLARIN.SI. Version 1.0, Last updated: 2022-07-08.
3. Yu, F., Zhang, H., Wang, B. (2023). Nature language reasoning, a survey. arXiv preprint arXiv:2303.14725.