Natural Language Processing

Natural Language Processing

- Natural Language Processing is a **cross-disciplinary** research field that draws heavily from **artificial intelligence** (AI), **machine learning** (ML), mathematics, and linguistics.
- Personal assistants, recommender systems, fake news identification, financial stock analysis, chatbots, autocorrection, auto-completion, intelligent search engines, and automatic translation or captioning are just a few examples of how NLP and AI are helping us manage the flood of data. However, systems to process natural language are far from perfect, which leaves much space for research.
- Some of the areas we work are:
  - Natural language understanding
  - Paraphrase detection
  - Text summarization
  - Media bias/Fake news detection
  - Semantic analysis/extraction
  - Sentiment analysis

For a complete list of our research topics visit our website!

NLP02 CS-Insights – State of the art in Computer Science Publications

Background

DBLP is the largest open-access repository of scientific articles on computer science and provides metadata associated with publications, authors, and venues. We retrieved more than 6 million publications from DBLP and extracted pertinent metadata (e.g., abstracts, author affiliations, citations) from the publication texts to create the DBLP Discovery Dataset (D3). Now, on CS-Insights we devised a system (back- and front-end) to explore our dataset and uncover all the trends regarding computer science publications. As CS-Insights is an ongoing project we need to fix it's open issues and extend its functionalities.

Goal

- Solve existing issues in CS-Insights-Roadmap

Tasks

- Work on project roadmap for CS-Insights
  - Backlog and additional features
- Propose extension for CS-Insights
  - Authors features (e.g., h-index)

Jan Philip Wahle

wahle@gipplab.org

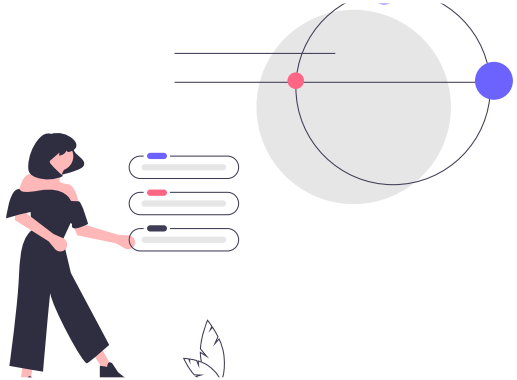
Terry L. Ruas

ruas@gipplab.org

NLP06 Semantic Feature Extraction for NLP Tasks

Background

The relationship between words in a sentence often have more semantic content than its actual words individually. Semantic analysis is arguably one of the oldest challenges in Natural Language Processing (NLP) and still present in almost all its downstream applications. However, the extraction of features that describe semantic aspects or the architecture of models/training tasks that capture intrinsic human characteristics is not a trivial task. We are interested in developing methods, training tasks, and architectures that can capture these underlying semantic features and use them in NLP tasks.



Goal

- Develop systems to solve NLP downstream tasks (or defined problems) using semantic features – Paraphrase generation/detection, text generation, language modeling, text summarization, etc.

Tasks

- Review the literature on selected task/problem;
- Extend devised approaches to recent state-of-the-art techniques (propose new ones);
- Evaluate your approach in specific datasets.

Jan Philip Wahle

wahle@gipplab.org

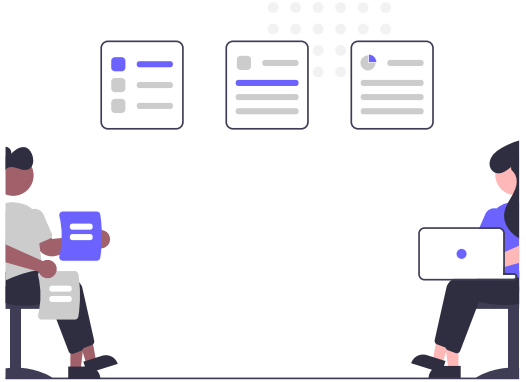
Terry L. Ruas

ruas@gipplab.org

NLP07 Meeting Summarization System Testbench

Background

The field of natural language processing has seen a significant amount of research in recent years on the task of meeting summarization. With the increasing availability of meeting transcripts, there is a growing need for efficient methods to automatically summarize the content of these meetings. As of now, due to the different formats of meetings and the dynamic, idiosyncratic nature, many domain- and problem-specific techniques have been introduced. However, the area lacks a standardized benchmark for evaluating these methods. Thus, it is difficult to compare and identify the strengths and weaknesses of the individual techniques.



Goal

- Design and develop a unified framework to test meeting summarization techniques (evaluation harness).

Tasks

- Design a solution to transform any kind of input format for models and datasets into one common form
- Develop a functionality to automatically add noise to the input text to assess models' robustness
- Implement a general applicable evaluation environment to test different models, datasets and metrics simultaneously
- Evaluate the most prominent techniques

Frederic Kirstein

kirstein@gipplab.org

Terry L. Ruas

ruas@gipplab.org

NLP13 Do Machines Have No Heart?

Background

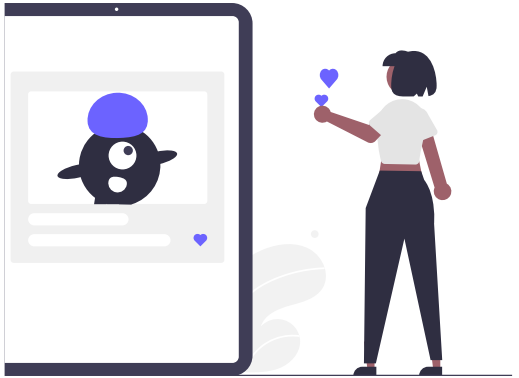
This project proposes an analysis of the sentiment embodied in text generated by large language models (LLMs), such as GPT-4. Using sentiment analysis methodologies, we aim to assess the sentiment polarity (positive, negative, neutral) and emotion classification (joy, anger, surprise) inherent in machine generated text across a diverse prompts and contexts. The proposed study will focus on understanding how LLMs, despite their lack of emotional states or personal perspectives, can potentially generate text embedding a wide spectrum of sentiment expressions. A significant aspect of our research will be identifying any sentiment inconsistencies in the model outputs, particularly in the face of ambiguous or complex prompts and comparing with existing human experiments

Goal

- Explore the sentiment embedded in LLM using machine-generated text and comparing it with human behavior

Tasks

- Literature review on sentiment/emotion analysis on language models
- Probe selected LLM to generate text following prompts/instructions
- Sentiment analysis and exploration of LLM's output
- Correlation between human and machine text



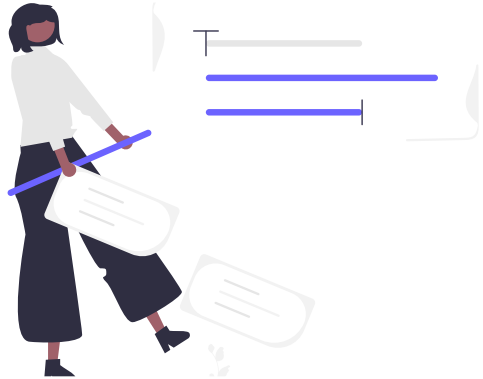
Jan Philip Wahle
wahle@gipplab.org

Terry L. Ruas
ruas@gipplab.org

NLP14 Paraphrase Types: Data and Task Generation

Background

Current paraphrase generation and detection systems are yet unaware of the lexical variables they manipulate. Generative models cannot be asked to perform certain types of perturbations, and detection models are unable to understand which paraphrase types they detect or learn limited language aspects (e.g., primarily syntax). The shallow notion of what composes paraphrases used by these systems limit their understanding of the task and makes it challenging to interpret detection decisions in practice. Thus, we need to leverage existing datasets and tasks used in Paraphrasing with more granular information so we can assess the problem better and develop more robust techniques.



Goal

- Extend current datasets used in paraphrase related tasks to include paraphrase types

Tasks

- Literature review on paraphrase types (atomic paraphrase types)
- Probe existing LLM to generate/classify pair sentences including selected paraphrase types (e.g., prompting, few-, or zero-shot) using the ETPC dataset as a reference
- Correlate (e.g., BLEU, similarity, ROUGE, BERTScore) generated paraphrase with existing data and select the best paraphrase types
- Extend the best paraphrase types to generate/classify new data from other paraphrase datasets
- Propose new tasks for the BIG-bench and/or GEM benchmarks based on Paraphrase Types

Jan Philip Wahle

wahle@gipplab.org

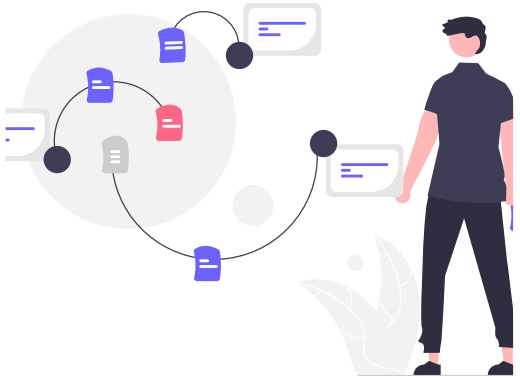
Terry L. Ruas

ruas@gipplab.org

NLP15 The Paraphrase Type Taxonomy

Background

This project proposes an extensive literature review to identify and critically evaluate various paraphrase types that have been proposed in linguistic, computational, and educational domains. By synthesizing these diverse perspectives, the project aims to develop a cohesive framework that categorizes paraphrase types based on linguistic features, context, and communicative intent. Through rigorous analysis and categorization, the project aims to establish a comprehensive taxonomy of paraphrase types. Furthermore, the research team plans to develop an open-access online repository, where the findings and the framework will be made available to the public, promoting collaboration and further research in this domain.



Goal

- Investigate and (re)organize available taxonomies and language models used in paraphrase types

Tasks

- Investigate available taxonomies used in paraphrase (types)
- Critical evaluation of existing ones (agreement and disagreements between them)
- Investigate available models used in paraphrase generation and detection
- Propose a new taxonomy (with definitions, examples, and instructions) for paraphrase types (generation and detection)

Jan Philip Wahle
wahle@gipplab.org

Terry L. Ruas
ruas@gipplab.org

NLP16 Meh-Tricks: Towards Reproducible Results in NLP

Background

In the rapidly evolving field of natural language processing (NLP), the measurement and evaluation of system performance are paramount to progress. Yet, the landscape of NLP metrics remains fragmented, with inconsistencies and variations in methodologies and libraries/wrappers which lead to potential misinterpretation, bias, incorrect, and irreproducible results. A great example of such a problem is ROGUE Scores, which evaluates the popular metric ROUGE in the last twenty years. This study's findings suggest that most reported scores are irreproducible, differences in evaluation protocol are common (affecting reported scores), and thousands of papers use nonstandard evaluation packages with software defects that produce provably incorrect scores. We will investigate in which other metrics the same issues are present.

Goal

- Investigate and analyze (selected) widely used metrics in natural language processing evaluation through their implementation, published papers, and baseline comparison.

Tasks

- Investigate and select widely used parametrized metrics in NLP downstream tasks
- Design and implement a semi-automated literature review (focus on metric) to retrieve relevant papers-metrics-libraries
- Identify decisions and aspects responsible for such discrepancies in their use/results
- Organize and analyze the landscape of such metrics and their use. A baseline evaluation should also be carried out.

Jan Philip Wahle

wahle@gipplab.org

Terry L. Ruas

ruas@gipplab.org

NLP17 LLMs and the Search for the Holy Prompt

Background

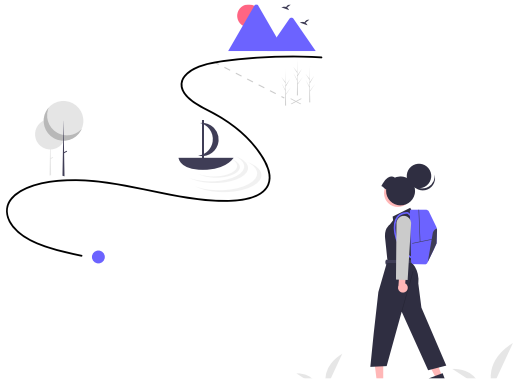
The advancements in the capabilities of large language models (LLMs) have ushered in a new era in artificial intelligence, with applications spanning diverse sectors (e.g., healthcare, education, entertainment). However, extracting precise and desired information from these models is not trivial. An emerging understanding of "prompt engineering" plays a crucial role in determining the efficacy, precision, and utility of the response from LLMs. Investigating the importance of prompt engineering is hence crucial, not only to improve the practical deployment of LLMs but also to delve deeper into understanding the intricacies of their internal representation and response mechanisms.

Goal

- Systematically explore and quantify the impact of prompt engineering on the performance of LLMs in paraphrase-related/text generation tasks and develop best practices for finding the best prompts

Tasks

- Literature Review: Examine existing literature on prompt engineering
- Empirical Study: Design experiments using various prompts across multiple tasks to measure the variability in LLMs' performance based on prompt differences.
- Framework Development: Construct a framework or guideline, based on empirical results, for crafting effective prompts that maximize desired outcomes when interacting with LLMs.
- Evaluate models, tasks, and prompts in selected tasks



Jan Philip Wahle

wahle@gipplab.org

Terry L. Ruas

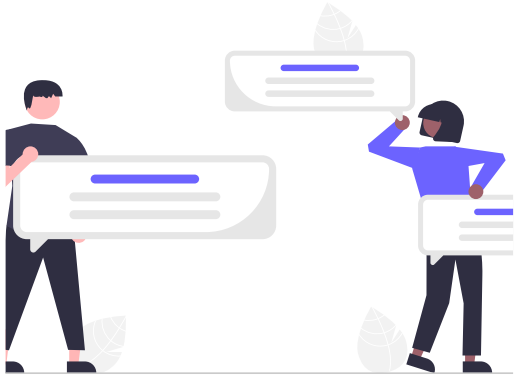
ruas@gipplab.org

NII

NLP18 It's not What you say it, but How you say it

Background

Large language models (LLM) have revolutionized Natural Language Processing (NLP) due to their ability to understand and generate human-like text. Their efficacy in producing meaningful outputs relies significantly on the way they are prompted. The same way as people, by slight alterations in prompts can lead to considerable differences in the generated content, which may affect both the quality and the nature of the response. This also raises the questions if specific models have a certain bias towards prompts. This phenomenon underscores the need for an in-depth analysis of the relationship between prompts, output and LLMS.



Goal

- Analyze how varying prompts influence the outputs of LLMs across selected NLP tasks and derive insights that can guide effective prompting strategies. Understand the differences between prompt alternation and selected LLM

Tasks

- Literature Review: Examine existing research and documented observations on how prompts influence large language models (select models and tasks)
- Experimental Design: Create a diverse set of prompts for selected NLP tasks. This set should include varied lengths, tones, styles, and implicit biases. (manual or auto)
- Data Collection: Use the selected prompts on consistent LLMs and collect/evaluate the outputs for each prompt (against gold standard)
- Analysis and Interpretation: Evaluate the data to discern patterns and relationships between prompt variations and model outputs.

Jan Philip Wahle

wahle@gipplab.org

Terry L. Ruas

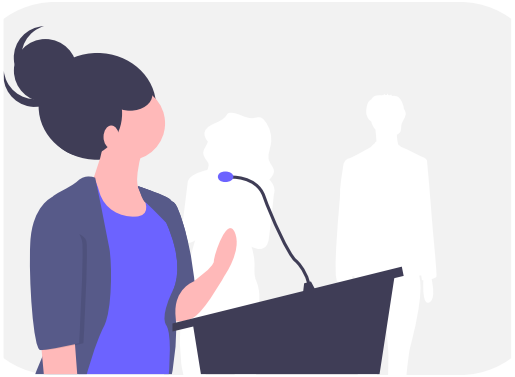
ruas@gipplab.org

NII

NLP19 What Are We Talking (and Doing) About in the EU?

Background

In an era with so much data available, knowing what to extract and how to structure it is essential for solving any problem. The structured compilation of extensively discussed topics at the European Union Parliament not only empowers policymakers, researchers, and analysts with a comprehensive overview of the legislative landscape but also grants citizens a clear overview of the issues that shape their continent. This project is not just a technical undertaking, but a venture that lays the foundation for transparency, accountability, and progress. This project focuses on the organization and exploration of the European Parliament's Open Data into meaningful structures so further investigations can be carried out.



Goal

- Analyze and organize the (selected) data of the European Parliament's into a more accessible way so specific investigations can be carried out.

Tasks

- Understand the structure of the European Parliament's Open Data Portal
- Identify major categories and topics we would like to gather and organize data (specific lexicons might be used to curate such data)
- Implement a solution to extract, categorize, and store data on selected topics from minutes, plenary sessions, speakers, etc;
- Propose sub-topic organization for the data
- Provide an initial (data science) analysis on selected topics

Jan Philip Wahle

wahle@gipplab.org

Terry L. Ruas

ruas@gipplab.org

PD05 Identifying Plagiarism of ChatGPT

Background

With the advent of advanced AI-powered language models like ChatGPT, the threat of machine-paraphrased plagiarism has become a serious concern. These models can generate text that is virtually indistinguishable from human writing, making it easy for individuals to commit plagiarism, but difficult for existing systems to detect. As these models become more accessible and widely adopted, the problem of plagiarism is expected to escalate, making it imperative for research institutions, publishers, and schools to have robust automated solutions in place.

Goal

- The goal of this project is to identify plagiarism of ChatGPT and other AI models in written works automatically.

Tasks

- Research and understand the inner workings of ChatGPT and other AI language models.
- Develop a method/training architecture for detecting generated and plagiarized text
- Evaluate the performance of the tool using quantitative and qualitative assessments.

Jan Philip Wahle
wahle@gipplab.org

Terry L. Ruas
ruas@gipplab.org

PD08: Authorship verification using LLMs

Background

Cloze test proved to be a useful tool for testing text comprehension. Some universities use it during a disciplinary procedure when a student is suspect from submitting a work authored by someone or something else (plagiarism, contract cheating, unallowed use of generative AI). Authors of the text are more likely to fill in correct words.

The project aims to find a method that identifies words to be masked such that the cloze test can reliably discriminate between authors and non-authors. LLMs are trained to predict the word in given context. Previous experiments showed that nouns that the model would not guess correctly are good candidates.

Goal

- To extend the existing project by conducting more experiments with LLMs and users
- To improve existing method (better discriminate between authors and non-authors)

Tasks

- Employ more language models to identify masked word (so far only MT-5 was used)
- Experiment with probability of the word in given context (so far only rank was used)
- Investigate the influence of language (English, German, etc.; native / non-native)
- Investigate the influence of time (authors forget their text and achieve lower scores)

The project aims to find a _____ that identifies words to be masked such that the cloze test can reliably _____ between authors and non-authors. LLMs are trained to predict the word in given context. Previous _____ showed that nouns that the model would not guess correctly are good candidates.

Tomáš Foltýnek
foltynek@fi.muni.cz

Terry L. Ruas
ruas@gipplab.org