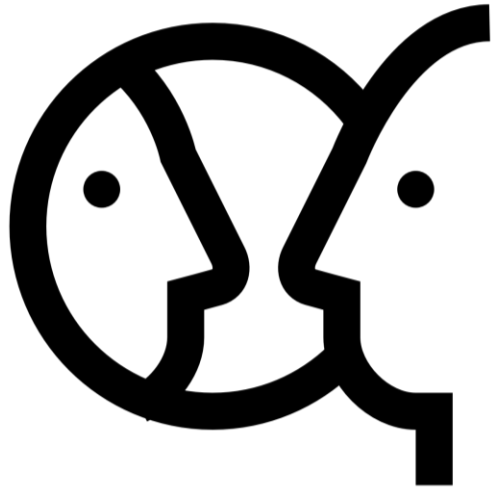
PD1: Neural Obfuscation-detectoR MAchiNe - NORMAN

Background

Paraphrasing tools are a severe threat to academic integrity. Dozens of cases of plagiarized documents appear on the media every day (e.g. Guttenberg) and their identification is a laborious task. One of the major bottlenecks in this domain is the acquisition of golden data that serve as ground truth to develop and test counter-measures to such problem. Paraphrasing tools (SpinBot, Spinner) charge users for spun-content generation, limiting even more to obtain valuable data. We want to develop a paraphrasing tool that can be used to produce machine generated text (for research purposes) so counter measures can be validated in real-life data/scenarios.



Goal

- Develop a paraphrasing tool based on recent language models in NLP (Transformers)

Tasks

- Investigate current Paraphrase systems
- Literature review on neural language models
- Create a Spun text generator using new Language models
- Compare NORMAN with baselines

Jan Wahle
jan.wahle@uni-wuppertal.de

Terry L. Ruas
ruas@uni-wuppertal.de

Norman Meuschke
meuschke@uni-wuppertal.de

LR4: Scientific Paper Recommendation via Topic-Chronological Content

Background

Our capacity to produce and store data has been continuously increasing in the last few years. In academia, this is no exception, and many academic repositories store valuable information for researchers. However, the process of organizing and recommending documents, from scientific repositories, is not a trivial task. We want to explore the new models and architectures in NLP to organize prospective suggestions based on their chronological order and their topic similarity producing a semantic timeline.

Goal

- Create a recommendation tool for scientific literature capable of organizing its suggestions by their topic-chronological relevance.

Tasks

- Review the literature on scientific paper recommendation;
- Prepare data from one or more repositories;
- Integrate NLP architectures to extract features from spec. dataset;
- Develop a system that organizes its recommendations in a topic-chronological manner;
- Evaluate your approach.

Corinna Breitinger
breitinger@uni-wuppertal.de

Terry L. Ruas
ruas@uni-wuppertal.de

NLP1: Semantic Feature Extraction for Text Classification

Background

The relationship between words in a sentence often tell us more about their 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 the semantic aspect of documents is not an easy exercise. We devised a group of approaches that are able to capture these underlying semantic features and use them in NLP tasks.

Goal

- Improve document similarity (other tasks also possible) using semantic features and recent advances in NLP/Word Embeddings/Transformers/Reformers

Tasks

- Review the literature on text classification using semantic features;
- Extend devised approaches to recent state-of-the-art word embeddings/transformers techniques;
- Evaluate your approach in specific datasets.

Jan Wahle
jan.wahle@uni-wuppertal.de

Terry L. Ruas
ruas@uni-wuppertal.de

NLP2: (In)stability and Usability of Word Embeddings

Background

With the increasing advances in word embedding models, their use is almost required in NLP downstream tasks or applications. However, how robust and stable pre-trained word embeddings models are? Even though the popularity of word embedding models makes them attractive for most NLP challenges there is little discussion about their semantic aspects concerning the semantic features they represent.

Goal

- Explore the (In)stability of recent word embeddings algorithms and pre-trained models towards NLP tasks and applications.

Tasks

- Review literature on (In)stability of word embeddings models;
- Select NLP tasks to evaluate similar/equivalent models;
- Present and discuss (possible) main aspects for the instability of word embeddings models;
- Propose and implement improvements to mitigate explored problems.

Jan Wahle
jan.wahle@uni-wuppertal.de

Terry L. Ruas
ruas@uni-wuppertal.de

NLP3: Transformers and NLP, a never-ending story?

Background

Neural Network-based models have gained much attention in the NLP community, mainly because of their success to capture latent semantic content and good results in tasks (WSD, sentiment analysis, translation). 2019 was elected the year of BERT. This was only possible because of the Transformer architecture, which is the foundation for many BERT-spawn approaches: SciBERT, RoBERTa, GlossBERT, Big BIRD, ERNIE, ELECTRA, etc. However, how many of them are disruptive? Which ones change the foundations? Where transfer- and multi-task learning fit in this scenario? Is there a trend? Taxonomy? Are they converging?

Goal

- Create a in depth literature review of Transformers applied to Natural Language Processing

Tasks

- Review the literature on Transformers (+Attention, +hierarchy, etc) mechanisms in NLP
- Propose taxonomy on the studied methods
- Propose/Categorize improvements for current methods

Jan Wahle
jan.wahle@uni-wuppertal.de

Terry L. Ruas
ruas@uni-wuppertal.de

NLP4: NLP-Land and its Secrets

Background

The ACL Anthology (AA) is the largest single repository of thousands of articles on Natural Language Processing (NLP) and Computational Linguistics (CL). It contains valuable metadata (e.g. venues, authors' name, title) that can be used to better understand the field. NLP Scholar, uses this data to examine the literature to identify broad trends in productivity, focus, and impact. We want to extend this analysis to specific components in NLP publications.

Goal

- Create an in-depth map of the publications in NLP and how their topics affect each other over time in their main venues

Tasks

- Review on system that analyze NLP Scholar dataset, and Google Scholar
- Incorporate semantic modeling to identify overlapping topics (e.g. embeddings, transformers, language models)
- Map how new and traditional topics behave over time in ACL venues
- Investigate the dynamic between venues and topics

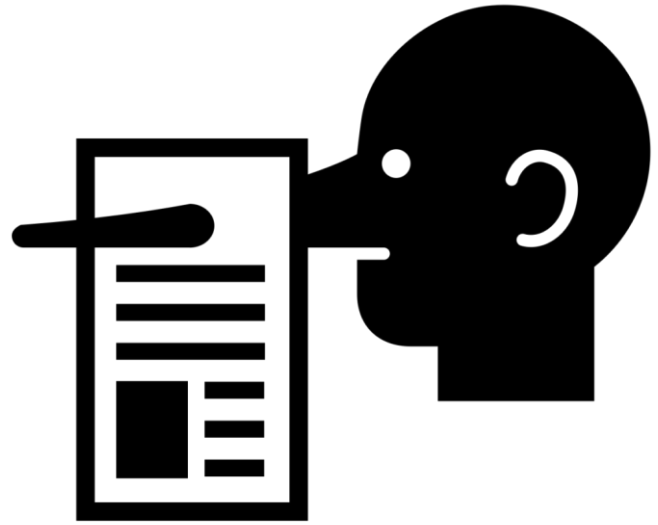
Jan Wahle
jan.wahle@uni-wuppertal.de

Terry L. Ruas
ruas@uni-wuppertal.de

NLP5: Is it true? A counter-attack to health-fake news (COVID)

Background

2020 will be remembered as the year that changed our modern way of life. Social distance, masks, hand-sanitizer, permanent lung problems, and loss of smell are some of the many things we hear and read about COVID-19. However, not every piece of information should be taken as truth. In fact, the amount and speed of which news are produced also contribute for the spread of misinformation. We want to help tackle this problem by creating a system that will identify fake-news in the health domain, more specifically related to COVID



Goal

- Create a system to combat health-related misinformation using recent NLP approaches.

Tasks

- Review the literature on recent NLP techniques used in fact-checking/misinformation in the health domain
- Propose taxonomy on the studied methods
- Apply recent techniques in benchmarks to identify the best classifier
- Propose a new approaches based on the artifacts provided by studied techniques

Jan Wahle
jan.wahle@uni-wuppertal.de

Terry L. Ruas
ruas@uni-wuppertal.de

NLP6: How Confident was Your Reviewer? Gauging Reviewer Confidence in Peer Review

Background

The peer-review system is the backbone of the scientific process, more specifically in conferences and journals. However, papers' evaluation processes are not always objective and are often open to issues such as bias, inconsistency, and arbitrariness. It is not uncommon to find reviewers providing shallow reviews to well well-written documents with high confidence scores. The International Conference on Learning Representations (ICLR) tries to mitigate the pipeline's subjectivity, as their reviews are openly available to anyone. We want to investigate the dynamic between the reviewers' report (e.g., confidence, text, sentiment) and the papers' final decision.

Goal

- Given a peer review text, the goal is to predict the reviewer confidence score and analyze its features.

Task

- Literature review on the peer review analysis
- Data curation from ICLR and ACL-2018 (Open and Closed review)
- Explore which neural language model is the most suitable for the problem
- Propose training architecture
- Analysis of the predicted output (e.g., ablation study)

Tirthankar Ghosal
tirthankar.pcs16@iitp.ac.in

Terry L. Ruas
ruas@uni-wuppertal.de

Topic SM1: REsearch Performance Analyzer (REPA)

Background

The process of evaluating a researcher is a laborious and manual task that many institutions rely on, especially for hiring prospective faculty members to their department. With many research-oriented social media platforms distinct perspectives are available to evaluate a scientist and the reach of her/his work. However, the task of evaluating researchers remains subjective and with little support. Current platforms allow a partial analysis of a scholar profile on pre-defined author-specific indicators. We want to take a step further and investigate the reach of researches through their publications and collaborators.

Goal

The main goal is to produce a detailed profile of a researcher about their areas of expertise, collaborators, and publications.

Tasks

- Review the literature on scraping scientific repositories and scientometrics;
- Provide the collaboration network of a researcher based on specific features;
- Analyze the similarity between two or more researches based on specific features.

Terry L. Ruas
ruas@uni-wuppertal.de

Norman Meuschke
meuschke@uni-wuppertal.de