Natural Language Processing
NLP1: Semantic Feature Extraction for Text Classification

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 and recent advances in NLP (word embeddings, Transformers, Reformers, 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 Wahlle
wahle@uni-wuppertal.de

Terry L. Ruas
ruas@uni-wuppertal.de
NLP2: 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 and after that a flood of related papers followed. This was only possible because of the Transformer architecture, which is the foundation of many BERT-related 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
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 our current system and analyze 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

• Understand A3 architecture
• Extend A3 to other data repositories, e.g., openreview, arXiv, IEEE, ACM, Springer.
  • Include more data from AA
• Front- and backend issues of A3
Background

Literature reviews are usually composed of multiple tasks, such as data selection, extraction, keyphrase generation, topic categorization, feature extraction, etc. Many of these tasks are executed by humans through laborious manual work. However, some of these tasks can be translated into known NLP downstream tasks. By automating some (the more the better) of the tasks in the literature review process, we could cut down the required time to produce a literature review from 12 to 6 months (maybe less)! The automation should be conceived with the goal of assisting users not substituting them, so user interaction/feedback is a requirement.

Goal

• Design a system to automate specific tasks in the Literature Review process

Tasks

• Understand all aspects that compose a (systematic) Literature Review
• Mapping of aspects x NLP (sub) tasks
• Propose architecture for the System (in waves/versions)
• Organize dataset for exploration (ACL Anthology or arXiv)
• Implement proposed system
• Test/Validate system with users
Background

Human Trafficking is a lucrative organized crime that majorly includes sex trafficking and puts many human beings in misery with lifelong traumas. Human trafficking has found on the internet an ideal environment to cultivate more material and potential predators through today's digital communities. Every day prospective abusers find new ways to hide their actions through dubious online advertisements and hidden messages.

Goal

• Design an NLP pipeline for detecting forms of human trafficking through online ads.

Tasks

• Literature review on human trafficking and NLP methods (not many use recent Transformers)
• Provide an empirical evaluation of NLP baselines for the task (Transformers, Word2Vec, tf-idf + SVM)
• Find deficiencies in the baselines
  • Here we can analyze token processing, tokens, embedding relation in vector space, edge cases
  • (One example would include resolving indicators for hidden messages such as "c@$H" $\leftrightarrow$ "cash")
• Devise novel modules targeting these deficiencies
• Run an ablation study on the new modules
NLP6: Predatory Conversation Detection

Background

Every year, millions of people worldwide are victims of sexual abuse, and physical violence. According to police crime statistics, most (>50%) sexual abuse cases are committed by a known and trusted person, while strangers are rare. Therefore, the most prominent indicator for detecting sexual abuse lies in analyzing conversations between predators and victims. Online text conversations are frequently available while spoken conversations are rare. This project aims to target both domains, online textual conversations as well as spoken conversations to find a relation between the two which can be leveraged for transfer learning from the higher resource to the low resource case.

Goal

- Design and implement an NLP approach to detect potential sexual abuse from online conversations and transfer the acquired knowledge to a small set of real conversations.

Tasks

- Literature review on Predatory Conversation Detection.
- Provide an empirical evaluation of NLP baselines for detecting textual (online) conversations.
- Identify deficiencies in the baseline approach
- Iterate over the baseline approach targeting the known deficiencies and provide an ablation study for each component.
- Test the generalization of the trained models for spoken conversations.

Jan Wahlle
wahle@uni-wuppertal.de

Terry L. Ruas
ruas@uni-wuppertal.de
Background

Many machine learning algorithms and verification procedures rely on high-quality data. Supervised machine learning requires a human assessment of the task (the ground-truth), for example about the sentiment of Twitter posts. Language models use self-supervision and receive their ground-truth by removing words from natural text and try to predict the same words again. For both applications, the quality of text is crucial. One of the most recent language models, GPT-3, uses a machine learning classifier to find high-quality data using a set of known high-quality text documents.

Goal

- Develop a machine learning model to select high-quality data based on a gold-standard.

Tasks

- Review literature about NLP datasets and contextual word embeddings (Transformers)
- Establish a gold-standard dataset for the task of human trafficking.
- Train a machine learning classifier to find similar articles to the gold-standard.
- Select text from a larger pool of potential human trafficking text and do a sample analysis on the quality of the automatically selected data.

Jan Wahle
wahle@uni-wuppertal.de

Terry L. Ruas
ruas@uni-wuppertal.de
Background
The more data we have, the more we need to analyze them. Text Summarization is an NLP downstream task in which one creates a shorter version of a given input (or many) containing the most important aspects of it. Aside from being a scientific downstream task, text summarization can be applied to many scenarios such as meetings notes, scientific papers, bug reports, books, news, blog posts, TLDR, and transcripts. At a higher level, summarization can be extractive (unaltered) or abstractive (semantic). We want to investigate the summarization problem through new architectures and systems, possibly for low resource languages.

Goal
• Explore Text Summarization task (Extractive/Abstractive) [low resource languages]

Tasks
• Review literature about NLP approaches working with text summarization (DL, GA?)
• Establish which models and datasets are being used to exercise this task
  • Applied or Benchmark
• Identify (dis)similar between these approaches
• Propose training architecture, data, or paradigm to compete with SOTA systems
• Evaluate approach in known datasets

Jan Wahle
wahle@uni-wuppertal.de

Terry L. Ruas
ruas@uni-wuppertal.de
Background

An increase in the number of online meetings made clear that typically meetings only have few key topics and a limited amount relevant information for all participants. Therefore, the extraction of their key topics and their summarization became more obvious. Although text summarization has been long explored by the natural language processing (NLP) community, its application to meetings and dialogs is still incipient. Meetings differ from traditional text as their structure is often dynamic. The interaction between multiple participants (e.g., discussions), their deviant formats, irregular sequences, different semantic styles, and topics promote a complex scenario. Short meetings can easily reach thousands of tokens in just a few minutes of conversation. Thus, techniques that produce high quality meeting summaries, including the most important ideas discussed between its participants, are still necessary.

Goal

- Explore Text Summarization task (Extractive/Abstractive) [low resource languages]

Tasks

- Review literature about NLP approaches working with text summarization and meeting summarization.
- Establish which models and datasets are being used to exercise this task
- Propose training architecture, training data, or paradigm.
- Integrate solutions in our repository and contribute to state-of-the-art solutions.
- Evaluate approach in known datasets
NLP10: Transforming Speech Recognition

Background
The Transformer and self-attention architectures have become de facto paradigm in any NLP task. The results obtained via Transformers are now also reaching top positions in multiple CV tasks. It is only natural to also investigate the applicability of Transformers in another important aspect within NLP, namely Speech-to-Text (aka End-to-End Speech Recognition). In this project we want to explore how Transformers-based architectures and pre-training paradigms from well-established NLP tasks can be applied into (Automatic) Speech Recognition.

Goal
• Incorporate Transformer-based architecture (and possibly pre-training methods) into Speech Recognition.

Tasks
• Literature review on Transformers (+attention) x Speech Recognition
• Selection of prospective model(s) and benchmark(s)
• Selection of task(s) within Speech Recognition
• Incorporate the NLP pre-training architecture into Speech Recognition pipeline
• DeBERTa, Shortformer, T5, Linformer, Performer, etc
• Validate proposed model architecture

Jan Wahle
wahl@uni-wuppertal.de

Terry L. Ruas
ruas@uni-wuppertal.de
NLP11: Transforming Topic Modeling – Transformers + LDA

Background

Topic modeling allows us to produce reasonable explanations (topics) for unobserved text content. Latent Dirichlet Allocation (LDA) is a generative statistical model that allows us to obtain, from a collection of documents, topics that explain this collection. At the same time, LDA also provide how the words/documents are composing these topics. Transformer is a deep learning architecture (stacked encoder-decoder) used in NLP to extract/produce context dense vector representation of language in a semantic level. Its Self-Attention mechanism allows us to map the intra-relationship (importance) between words in a text modeling our language. Our job is to bring the probabilistic robustness of LDA to the context world of Transformers. The former was the model choice for many years in text mining, now the latter is the standard choice for all NLP tasks.

Goal

• Combine the LDA and Transformers to leverage the latent semantic structures through context embeddings

Tasks

• Review literature about LDA and context word embeddings (Transformers)
• Investigate how LDA and Transformers can be combined(concatenated/complementary)
• Set up a pipeline to validate known techniques combining LDA and Transformers (tasks)
• Propose alternative method to integrate both techniques
• Validate model in known downstream tasks

Jan Wahlle
wahle@uni-wuppertal.de

Terry L. Ruas
ruas@uni-wuppertal.de
NLP12: Graph Neural Networks in NLP?

Background
Is there a pattern in trend topics? What about Fake news? Do news articles about sensitive content connect in the same way? Do biased articles have something in common? How can we investigate the behavior and relationship of elements on the web? Graphs are powerful structures that help us to model relationships between objects. In NLP, Transformers and word embeddings have been used extensively to extract semantic features from text and use them to solve many tasks (e.g., document classification, QA, text summarization). Can we combine these two paradigms to investigate practical questions?

Goal
- Use Graph structures and Transformers to study the relationship and structure of Topics/News/FakeNews/UserComments/Rumors/NaturalDisasters/Etc

Tasks
- Literature review on Graphs in NLP (focused on news/text representation)
- Investigate how to apply Transformer models into Graph representations
- Select a target problem to apply Graph and Transformers
  - Downstream tasks or application
- Implement a solution combining Graphs + Transformers
  - Validate system either with a study case or benchmark

Jan Wahle
wahle@uni-wuppertal.de

Terry L. Ruas
ruas@uni-wuppertal.de
Scientometrics
SM1: REsearch Performance Analyzer—REPA

Background
Evaluating the performance of a researcher is a laborious manual task that many institutions face, especially for hiring prospective faculty members. Many research-oriented social media platforms offer basic metrics for evaluating a scientist and the impact of her/his work. However, the indicators available in current platforms are insufficient to enable a comprehensive performance evaluation. We want to improve the situation by providing a novel tool that allows in-depth comparisons of researchers based on their publication performance and collaboration network. A first system implemented in Python exists.

Goal
Extend the current platform, primarily by tapping additional data sources and implementing more performance metrics and result visualizations.

Tasks
• Review the literature on scraping scientific repositories and Scientometrics;
• Extend and improve the backend and frontend of the current system;
• Analyze the performance of researchers using specific features.