Identification of Media Bias

Gipp Lab
Scientific Infomation Analytics

Identification of Media Bias

- Automated identification of media bias is an applied NLP research.
- The project aims identifying text snippets, visual elements, or ideas conveyed in the texts that make news consumers perceive information in a biased way, i.e., word choice and labeling that portrayed politicians in a specific way.
- The main goal is to identify instances of bias in multiple forms and increase bias awareness of the news readers.
- Media bias goes hand-in-hand with fake news detection but focuses mostly on the official news outlets.
- Media bias detection is an interdisciplinary project between computer science, political science, digital humanities, phycology, computational linguistics.
- The project covers NLP areas such as:
  - Coreference resolution
  - Sentiment analysis
  - Hate speech detection
  - Dependency parsing
  - Information extraction (e.g., named entity recognition)

A complete list of Media Bias topics visit our website!

How to resolve "invade the country" = "cross the border“?!

Background

When reporting about events, journalists use different words to describe the same actors and entities, often based on personal or the outlet’s political or ideological views. News consumers are highly influenced by a non-objective reporting style. Current state-of-the-art cross-document co-reference resolution (CDCR) systems still lack robust approaches to resolve mentions of high lexical diversity and complex non-named-entity concepts.

CNN	Al Jazeera
UK soldiers cleared in Iraqi death	British murderers in Iraq acquitted
Seven British soldiers were acquitted on Thursday of charges of beating an innocent Iraqi teenager to death with rifle butts.	The judge on Thursday dismissed murder charges against seven soldiers, who are accused of murdering Iraqi teenager.

Goal

Develop a novel CDCR model based on Transformers by using transfer learning from paraphrase identification to identify coreferential mentions of high lexical diversity.

Tasks

- Review literature about CDCR;
- Review literature on Transformers, transfer learning, and datasets for CDCR and paraphrase identification;
- Design and train a neural network model to resolve noun and verb phrases referring to the same concepts;
- Evaluate the algorithm in a diverse collection of CDCR datasets.

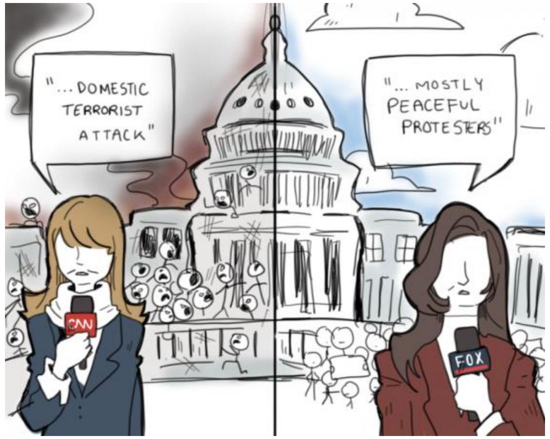
Anastasia Zhukova
zhukova@gipplab.org

Felix Hamborg
hamborg@gipplab.org

NewsWCL 2.0: CDCR Dataset with High Lexical Diversity

Background

When reporting about events, journalists use different words to describe the same actors and entities, often based on personal or the outlet’s political or ideological views. Identity coreference relations typically link mentions such as “Donald Trump” and “the president” that report about the true facts. On the contrary, loose coreference or bridging relations may convey bias, e.g., in “Donald Trump's 'Impulsive' Decision-making” a metonymy relation between “Trump” and “decision-making” frames Donald Trump as an impulsive person. Bias by word choice and labeling yield a non-objective reporting style and influences news consumers.


"...DOMESTIC TERRORIST ATTACK"
"...MOSTLY PEACEFUL PROTESTERS"
CNN
FOX

Goal

Create a new CDCR dataset by researching how phrases become coreferential with relations of varying strength, how these relations facilitate bias by word choice and labeling, and influence on news readers’ perception of the information.

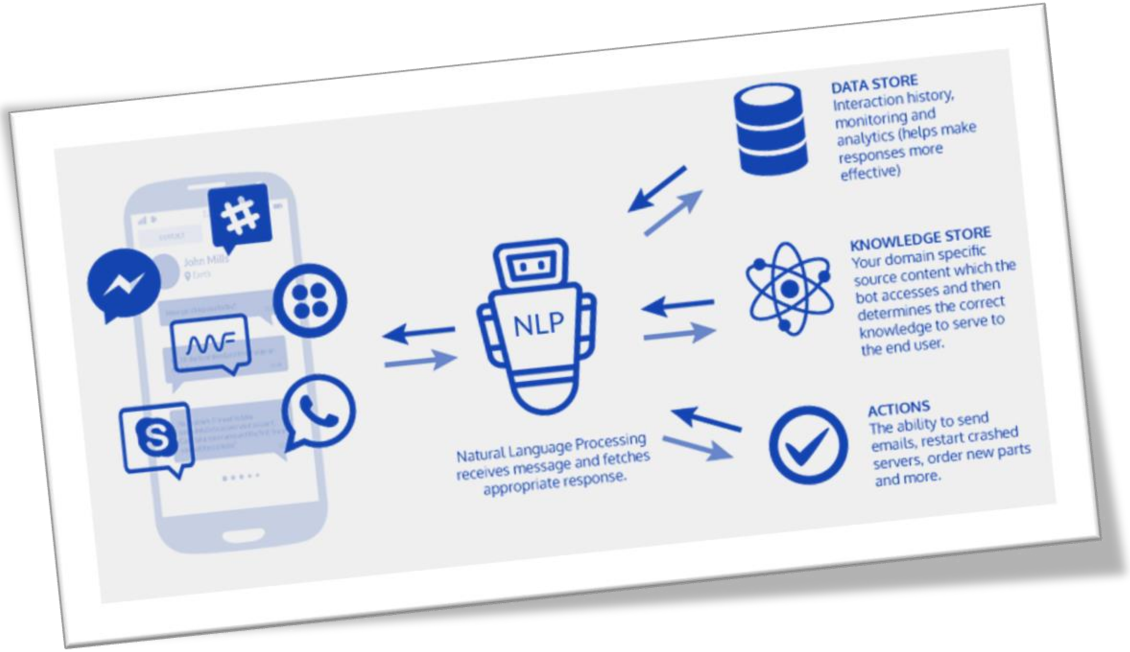
Tasks

- Review literature review about 1) types of coreference, bridging, and near-identity relations, 2) which types of relations do the datasets for (cross-document) coreference resolution include or excluded from annotations.
- Explore how coreference and bridging relations influence in news readers’ reasoning.
- Create a coding book to annotate coreferential mentions and their relations to identify cases of bias by word choice and labeling.
- Annotate a new CDCR dataset that focuses on high lexical diversity.

Anastasia Zhukova
zhukova@gipplab.org

Felix Hamborg
hamborg@gipplab.org

NLP-driven Plant Assistant



Natural Language Processing receives message and fetches appropriate response.

DATA STORE
Interaction history, monitoring and analytics (helps make responses more effective)

KNOWLEDGE STORE
Your domain specific source content which the bot accesses and then determines the correct knowledge to serve to the end user.

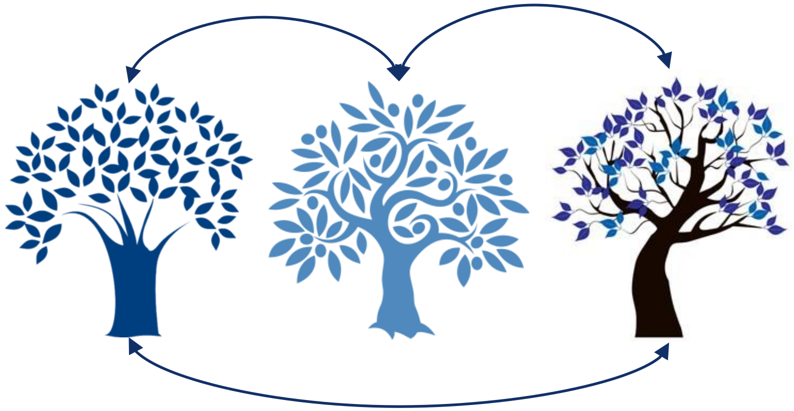
ACTIONS
The ability to send emails, restart crashed servers, order new parts and more.

NLP-driven Plant Assistant

- The project "NLP-driven Plant Assistant" aims at creating a tool that supports plant operators at their daily operations on a industrial processing plant.
- Plant Assistant aggregates knowledge and experience from logged plant operations and provides fast, efficient, and interactive feedback when users need solutions to encountered problems.
- A goal of the project is to leverage the recent advances in NLP to tailor language models towards domain-specific text data of multiple languages with uneven data quality, data scarcity, and lack of annotated sources.
- Plant Assistant is an applied research project conducted as a collaboration between GippLab and eschbach GmBH funded by ZIM (Zentrales Innovationsprogramm Mittelstand) run by the German Ministry of Economic Affairs and Climate Action.
- Plant Assistant addresses following NLP research areas/tasks:
  - Domain adaption of language models
  - Coreference resolution
  - Question & Answering systems
  - Information extraction
  - Automated annotation of data corpora
  - Semantic Information Retrieval

A complete list of Plant Asistant topics visit our website!

Resolution of Functional Location Trees of a Plant



Background

A functional location tree represents a structure of all machinery involved in production. A way to organize the components as nodes and leaves indicates the dependency between machines and their parts. During plant exploitation, a stored representation of the functional location tree may change due to a plant expansion, integration of another software, or adaption of a more efficient naming scheme. While changing to a certain level, the tree components remain recognized by the domain users as the same machinery. However, a problem arises when a software that needs to report about the same physical unit cannot recognize anymore that the name of that unit has changed.

Goal

Design and develop a system that will resolves functional locations between multiple naming versions of the functional locations, stores the resolved information in a graph database, and enables retrieving information from the resolved graph.

Tasks

- Read about text similarities, e.g., string and semantic, and graph databases (Neo4j).
- Design an algorithm how to incrementally build a graph and resolve functional locations to the previously added ones.
- Propose a search algorithm that retrieves nodes from a graph, given not identically matching 1) names of functional locations, 2) aliases of functional locations.

Anastasia Zhukova
zhukova@gipplab.org

Christian Matt
christian.matt@eschbach.com

Applying User Feedback to Improve Results of Semantic Search

Background

Recent directions of Information Retrieval focus on applying advances in Natural Language Processing (NLP) to retrieve relevant documents not only based on query-terms matching but also meaning of the query. Research of semantic search utilizes transformer-based language models that capture general semantics and syntactic of natural languages. When language models are applied to a narrow domain, e.g., processing industry, semantic search should be improved with a feedback of domain experts.

Goal

Design and develop a prototype of semantic search that employs language models, enables collecting user feedback, and fine-tunes a ranking transformer-based model on user feedback.

Anastasia Zhukova
zhukova@gipplab.org

Tasks

- Get familiar with Haystack framework for semantic search.
- Read about options to represent queries and documents with transformer models.
- Build a prototype using Haystack that retrieves and ranks results with two different models.
- Collect user feedback on the search results and fine-tune a ranking model on this feedback.

Christian Matt
christian.matt@eschbach.com

Domain-specific Named Entity Annotation with a Human-in-Loop

Background

Named Entity Recognition (NER) is a task in NLP for extracting and classifying spans of text into a set of predefined entity categories, e.g., person, organization, country, and datetime. NER with general categories is hard to apply to specific domains such as biology, chemistry, or technology. Annotation from scratch even of small datasets is a time-consuming process. An approach that combines an automated annotation of a silver-quality dataset and validation by a human-in-loop that corrects mistakes and adds new categories and corresponding terms speeds up an annotation process and increases data quality.

Goal

Design and develop an approach that 1) automatically identifies entity categories from a set of domain-specific texts, 2) collects user feedback with a GUI, 3) takes into account user feedback and annotates more text.

Tasks

- Research literature about NER and active learning for domain-specific languages.
- Design an approach that uses external sources, e.g., Wiktionary, and creates a small annotated dataset, and a model that incorporates feedback from annotators and annotates more domain texts.
- Improve GUI that collects user feedback.

Anastasia Zhukova
zhukova@gipplab.org

Christian Matt
christian.matt@eschbach.com