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 complex non-named-entity related concepts.

Goal

Develop a novel CDCR approach to identify mentions of the same, entities, events, objects, and abstract concepts with identity, border, and near-identity relations.

Tasks

- Review the literature about CDCR;
- Review literature on the application of the context-based word embeddings and language models (e.g., ELMo, BERT, etc.) to the relatively small text corpora;
- Design and train a neural network model to resolve noun and verb phrases referring to the same concepts of varying complexity of the coreference relations;
- Evaluate the algorithm in a structured pipeline of Newsalyze
MB2: Concept Trees of News Stories

Background
When reporting about the same event, media outlets tend to use different, often bias-prone, word choices to refer to the same actors, entities or concepts. When resolved, these concepts require a tree structure to represent the cross-concept relations, e.g., „USA“ contains „Donald Trump.“ The tree structure will help users to overview quickly the major and minor concepts of the news stories.

Goal
Develop an algorithm that will perform a structured content analysis and arrange the previously identified entities from a set of news articles in a tree structure with part-whole relations and relations that are extracted as event-concepts.

Tasks
• Research literature about knowledge graphs, e.g., WikiData, ontologies and lexical bases, e.g. ConceptNet;
• Develop an approach that will identify relations between the entities in a set of related and weakly related news articles;
• Evaluate results on a provided dataset as a part of Newsalyze system
MB3: Concept Type Identification

Background

Named Entity Recognition, a branch in NLP research, focuses on extracting phrases that mainly refer to the four standard types: persons, locations, time, and organizations. While these types are essential for text content understanding, many other entity types are overlooked by the established NER types, e.g., groups of persons or events.

Goal

Develop a machine learning (ML) model to annotate texts mainly in the news domain.

Tasks

• Research literature about (1) recent models for NER and other extraction & labeling tasks in NLP, (2) senses of synsets in WordNet;
• Develop a ML model that will annotate noun and verb phrases into the set of the “silver” quality preidentified concept types;
• Evaluate results as a part of Newsalyze system
Background
The news stories are dynamic: they originate with one event or discussion, get modified over the time, may end in a few days, develop into frequent reporting its substory or evolve into a similar yet different event. The established cross-document coreference resolution (CDCR) systems typically analyze the preselected apriori related texts. Unfortunately, such systems ignore open story “environment”, i.e., do not resolve news articles cross-related only to some degree. Tracking of the news storylines and its components, e.g., involved actors and concepts, assists in analyzing the evolution of the reported events and the rapidly changing news agenda.

Goal
Develop an approach to identify the storylines, i.e., the semantically and event-related news articles, and visualize how the content composition and the wording evolved over time.

Tasks
• Research literature about text similarity;
• Extract storylines and break them into narrowly-related sets of articles;
• Create an interactive visualization for the analysis of the story flow and its aspects;
• Evaluate results as a part of Newsalyze system

Anastasia Zhukova
zhukova@uni-wuppertal.de

Felix Hamborg
felix.hamborg@uni-konstanz.de