NLP17: Annotation of a Dataset for Coreference Resolution with Active Learning

Background

Annotation of any dataset for NLP is a time-consuming task. Annotation of a dataset for cross-document coreference resolution (CDCR) is typically a more complex and time-consuming task due to a large volume of document to annotate and extensive training of human annotators. Moreover, various annotation schemes focus on different coreference relations: bridging or loose reference relations are harder to identity than strict identity. Recently, active learning has been gaining a lot of attention in assisting human annotators to create large datasets with higher quality.

Goal

Investigate how to design an active learning model for the coreference resolution task and implement an annotation system that would work with any exemplar annotated CDCR dataset.

Tasks

- Review literature review about active learning in NLP and CDCR datasets
- Design and implement CDCR model with active learning to annotate new collections of related articles given few annotated examples
- Evaluate model on multiple CDCR datasets

Anastasia Zhukova
zhukova@uni-wuppertal.de

Felix Hamborg
felix.hamborg@uni-konstanz.de

NLP18: Benchmarking Cross-document Coreference Resolution Models

Background

Cross-document coreference resolution (CDCR) is a research area in NLP that identifies chains of the mentions referring to the same actor, country, action, event, etc. across multiple related articles. A majority of the CDCR models are developed and trained on one single event-driven CDCR dataset. The most recent research on evaluation focuses on evaluating the existing CDCR models on other event-driven CDCR datasets. Unfortunately, the reported evaluations miss out the other CDCR datasets that focus, for example, on the concept-driven annotation schemes or mentions linked with more loose coreference relations.

Goal

Benchmark the most recent and state-of-art CDCR models on the CDCR datasets with various annotation schemes and tasks for CDCR.

Tasks

- Review the state-of-the art CDCR models and datasets
- Implement the publicly available code for the models into one benchmarking system
- Ensure the same input format for all datasets to evaluate
- Train and test the models on the same reproducible setup
- Evaluate the models on the CoNLL metrics

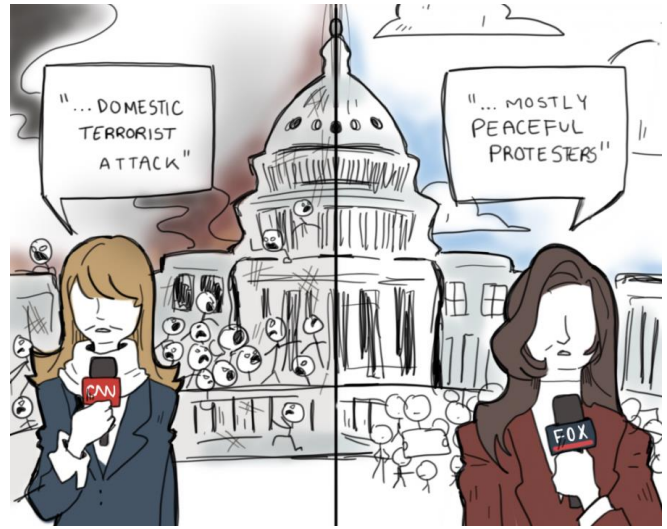
Anastasia Zhukova
zhukova@uni-wuppertal.de

Jan Wahle
wahle@uni-wuppertal.de

MB1: Coreference and bridging relations: linking facts and conveying bias

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.



Goal

Investigate types of coreference, bridging, and near-identity relations and derive how these relations convey facts or bias and influence 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.
- Provide a recommendation for annotating coreferential mentions and their relations to identify cases of bias by word choice and labeling.

Anastasia Zhukova
zhukova@uni-wuppertal.de

Felix Hamborg
felix.hamborg@uni-konstanz.de

MB2: Improving Visualization for Cross-document Coreference Resolution in News Articles

Background

Coreference resolution is a research area in NLP that identifies chains of the mentions of the same actor, country, action, event, etc. Cross-document coreference resolution (CDCR) identifies coreferential mentions across a set of articles and enables effective comparison of how often and in which wording news articles reported about entities or concepts. Although development of the CDCR models is a quickly advancing, very few research addresses effective visualization of the extracted and resolved mentions.

Goal

Expand and enhance a previously developed interactive visualization that enables exploration of the frequency, lexical diversity, and entity types of the identified concepts from related news articles of various political polarity.

Tasks

- Enhance an interactive visualizing to enable news readers to quickly analyze the text content through the extracted mentions;
- Design and train a weakly-supervised classification model for entity type identification
- Design and train a classification model for news article polarity identification
- Visualize phrasing diversity of the resolved mentions

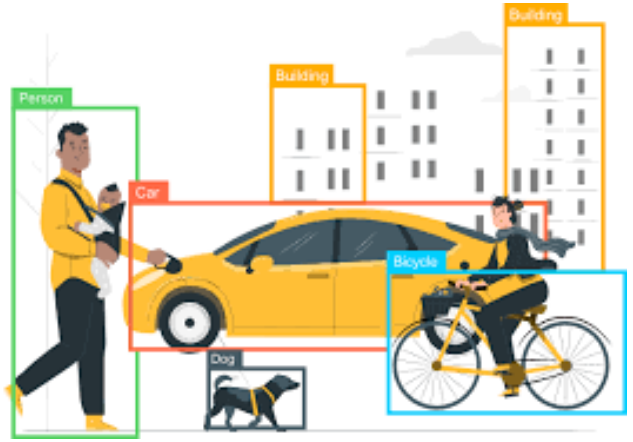
Anastasia Zhukova
zhukova@uni-wuppertal.de

Felix Hamborg
felix.hamborg@uni-konstanz.de

PA1: Annotation Tool for Creating Domain-specific Text Datasets

Background

Developing and training the NLP models requires annotated datasets to learn which tasks to perform. Such datasets, e.g., for named entity recognition, are created on general texts such as news articles or Wikipedia articles. Therefore, the models trained on such datasets fall short performing similar tasks on the texts from specific domains, e.g., chemistry or biology. Creation of a domain-specific dataset is required to train NLP models to learn patterns of these domains.



Goal

Develop a GUI for annotating and refining datasets for named entity recognition, text classification, and coreference resolution.

Tasks

- Design a GUI that takes as input text document and optionally pre-annotated text excerpts and allows refining these excerpts and annotating new ones
- Provide two views: in-text annotation and the overview of all annotations
- Ensure reusable software architecture between all three annotation tasks

Anastasia Zhukova
zhukova@uni-wuppertal.de

Norman Meuschke
meuschke@uni-wuppertal.de

PA2: Domain-specific Named Entity Annotation

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 such specific domains as biology, chemistry, or technology. Compared to models based on machine learning, the state-of-the-art NER models based on language models require smaller amounts of annotated data. Yet annotation even of small datasets is a time-consuming process for the untrained domain annotators.

Goal

Design and develop an approach that 1) automatically identifies entity categories from a set of domain-specific texts, 2) annotates spans of these domain texts into these categories.

Tasks

- Research literature about NER for domain-specific languages.
- Focus on fine-tuning of language models on unsupervised tasks.
- Identify which external resources can assist in bringing additional level of semantics to the domain data, e.g., Wiktionary.
- Use external resources and/or a small set of domain data to form learning objectives for named entity annotation

Anastasia Zhukova
zhukova@uni-wuppertal.de

Norman Meuschke
meuschke@uni-wuppertal.de

PA3: Acquisition of a Large Domain-specific Corpus for Fine-tuning of Language Models

Background

Language models (LM) learn probabilistic distribution of words in texts. In other words, LMs learn patterns, rules, and semantics of the language and allow reusing this learned knowledge in applications, e.g., text classification. LMs are trained on large general text collections such as news or Wikipedia articles and typically loose their accuracy and effectiveness when used in specific domains, e.g., technology. Fine-tuning of LM models with continuing training on the domain-specific corpora enables the LMs to capture relations that occur only in particular domains.

Goal

Design, implement, and evaluate a system that collects text data from large open sources, e.g., Wikidata, given keywords of a specific domain of interest and performs fine-tuning of the state-of-art language models on various languages.

Tasks

- Research a methodology to search and retrieve related information given keywords in large collections of text depending on the language of the keywords.
- Design and implement an approach to collect the domain-specific data.
- Fine-tune a state-of-the-art language model on the collected dataset and evaluate it on a direct task, e.g., classification

Anastasia Zhukova
zhukova@uni-wuppertal.de

Jan Wahle
wahle@uni-wuppertal.de

PA4: Incorporating structural and time dependencies into fine-tuning language models

Background

Language models (LM) learn statistical distribution of words in languages, thus, learning patterns and semantics of the languages. LMs are usually pretrained on very large text corpora and afterwards are fine-tuned to be used for specific domain languages and tasks. Fine-tuning LMs is performed by continuing training of a LM model on a domain-specific tasks to learn the domain patterns of the text. The task of our NLP-driven Plant Assistant is to learn 1) a domain language of a plant on which it will be deployed, 2) structural dependencies of the functional units of this plant, 3) temporal dependencies of the events logged into the journal of daily operations.

Goal

Design, implement, and evaluate a fine-tuning strategy that includes the tasks to learn temporal relations of the logged events and structural relations of the domain-specific terms.

Tasks

- Research strategies to fine-tune language models and focus on the task that learned temporal dependencies and structural dependencies between terms.
- Design and implement the fine-tuning tasks for learning these types of patterns in the domain data of logged events in a plant
- Evaluate the proposed fine-tuning tasks

Anastasia Zhukova
zhukova@uni-wuppertal.de

Jan Wahle
wahle@uni-wuppertal.de