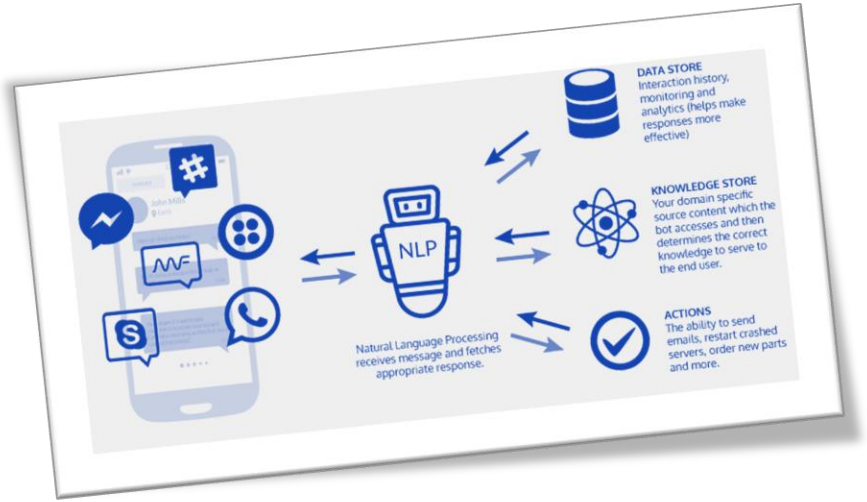


NLP-driven Plant Assistant

Gipp Lab
Scientific Infomation Analytics

NLP-driven Plant Assistant

GippLab
Sci. Info. Analytics

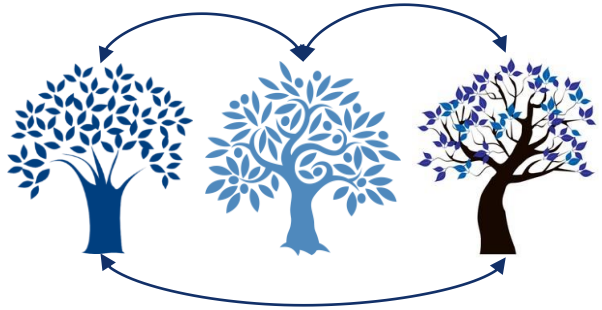
- The project "NLP-driven Plant Assistant" aims at creating a tool that supports plant operators at their daily operations on a industrial processing plant.
- Plant Assistant aggregates knowledge and experience from logged plant operations and provides fast, efficient, and interactive feedback when users 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

Resolution of Functional Location Trees of a Plant

GippLab
Sci. Info. Analytics

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

GippLab
Sci. Info. Analytics

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.

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.

Anastasia Zhukova
zhukova@gipplab.org

Christian Matt
christian.matt@eschbach.com

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

GippLab
Sci. Info. Analytics

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