Enterprise Platform and Integration Concepts
Research Group of Prof. Dr. Hasso Plattner
Master Project Summer Term 2015

Ask your Database: Natural Language Processing using In-Memory Technology

The goal of this Master project is to develop and implement novel concepts for a Question Answering system for the biomedical domain. Question Answering (QA) is one of the more complex applications of Natural Language Processing (NLP) and consists on processing questions in natural language, instead of the usual keyword-based query, and providing exact answers in return, instead of potential relevant documents. NLP includes a variety of tasks such as tokenization (delimitation of words), part-of-speech tagging (assignment of syntactic categories to words), chunking (delimitation of phrases) and syntactic parsing (building a syntactic structure for a sentence). It also involves semantic-related tasks such as named-entity recognition (delimitation of predefined entity types, e.g., person and organization names), relation extraction (identification of pre-defined relations from text) and semantic role labeling (assignment of pre-defined semantic roles to phrases).

QA systems involve integration of many of the NLP components in their three main steps as described below:

- question processing: processing of questions and construction of queries;
- passage retrieval: retrieval of sentences or short text passages relevant to the question and based on the derived query;
- answer processing: extracting the exact answer(s) and/or building summaries for the provided question.

The current textual data deluge, e.g., scientific publications, Web pages or messages in the social media, demands fast and real-time processing to support various NLP applications, specially for QA. There are currently three

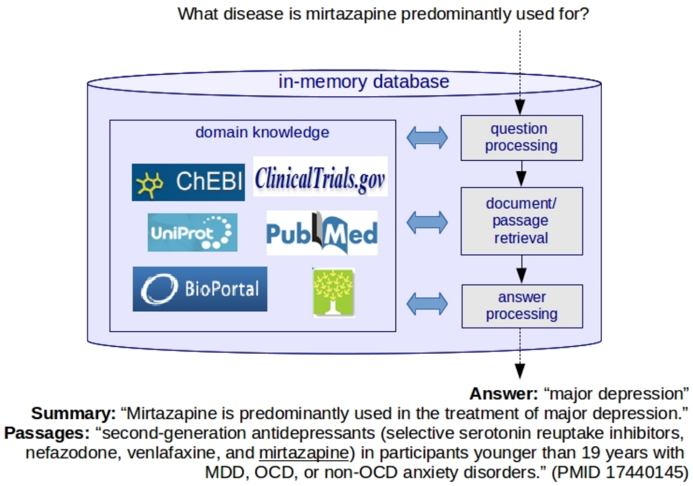
What disease is mirtazapine predominantly used for?
in-memory database
domain knowledge
ChEBI
ClinicalTrials.gov
UniProt
PubMed
BioPortal
question processing
document/ passage retrieval
answer processing
Answer: “major depression”
Summary: “Mirtazapine is predominantly used in the treatment of major depression.”
Passages: “second-generation antidepressants (selective serotonin reuptake inhibitors, nefazodone, venlafaxine, and mirtazapine) in participants younger than 19 years with MDD, OCD, or non-OCD anxiety disorders.” (PMID 17440145)

Figure: Architecture of a IMDB-based QA system for Biomedicine.

QA systems for Biomedicine, but none of them provides fast and reliable answers to the users. A recent comparison of their results and time response for 40 randomly selected questions from the EU-funded BioASQ dataset (`http://bioasq.org`) shows that a correct answer was returned only for 5 of these questions when merging answers returned by all three systems. Further, average time response varied from 10 seconds to a maximum of 100 seconds, after which no answer was returned for many of the questions.

Building real-time applications that integrate many of these NLP components is a challenge as these are time-consuming processes. In-memory database (IMDB) technology comes as an alternative given its ability to quickly process large document collections in real time. Our system runs on the top of SAP HANA database and has been scored first for passage retrieval in the last edition of the BioASQ challenge.

This Master project will include implementation of NLP components which are still missing in HANA (e.g., chunking), adaptation of existing methods to Biomedicine (e.g., part-of-speech tagging), development of missing steps in our QA system (e.g., answer extraction, question classification, summarization) and evaluation of the system on the BioASQ benchmark. Implementation will be based on supervised and semi-supervised learning methods and will make use of training datasets made available in BioASQ and existing biomedical corpora.

Project goals

- Participate in the development of a Question Answering system for the biomedical domain
- Implement new NLP functionalities in SAP HANA database
- Adapt current NLP features in SAP HANA to the biomedical domain
- Evaluate the application on the BioASQ dataset

Technology and skills

Participants should have knowledge of SQL and at least one programming language (Python, Java, preferably C++), and interest in database technologies (in-memory database, stored procedures), natural language processing and machine learning (supervised and semi-supervised learning). No previous knowledge on Biomedicine is necessary, domain knowledge will be integrated through the use of available resources (ontologies, dictionaries, corpora, etc).

Group structure and project start

The team will consist of 3-5 students and the project will start on April 13th. There will be an initial meeting with all participants on April 9th for presentation of the details of the project and to get familiar with supervisors and colleagues.

Contact

Your are welcome to contact or visit us in the Villa, HPI Campus II, room V1.02:

Dr. Mariana Neves
`mariana.neves@hpi.de`

Cindy Fähnrich, MSc.
`cindy.faehnrich@hpi.de`

Dr. Matthias Uflacker
`matthias.uflacker@hpi.de`