Johannes Leveling
Intelligent Information and Communication Systems
FernUniversität in Hagen (University of Hagen)
Hagen, Germany

IRSAW – Towards Semantic Annotation of Documents for Question Answering

Motivation: More information becomes available and/but precise answers are difficult to find

→ IRSAW project (Intelligent information Retrieval on the basis of a Semantically Annotated Web)

- funded by the DFG (Deutsche Forschungsgemeinschaft) from 2006–2008
- general idea: get raw documents from the internet (relevant to a user's natural language question), semantically analyze and annotate the question and documents and apply deep linguistic methods for question answering on document content

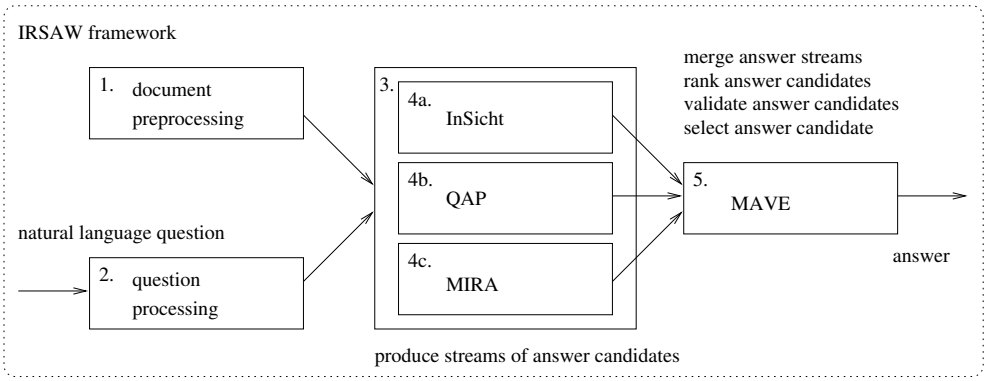
Project Goals: Development of a semantically based question answering (QA) framework integrating modules for different tasks, e.g.

- deep natural language analysis (for German language), producing a semantic network representation of questions and documents (based on the knowledge representation paradigm MultiNet)
  → allows a full semantic interpretation of questions and documents on which logical inferences are based (state-of-the-art: mostly shallow methods)

  also aims at investigating linguistic phenomena in questions and documents (e.g. idioms, metonymy, and temporal and spatial aspects)
- combination of different data streams containing answer candidates
  → applying different methods to produce answer streams increases recall and robustness
- logical answer validation
  → selecting validated answers from streams of answer candidates increases precision
- natural language generation
  → allows for rephrasing from text and combination of answer fragments from different documents (state-of-the-art: extracting snippets from the text)

Software: IRSAW will result in two software components accessible via the internet:

1. the question answering system IRSAW and
2. a web-service for the semantic annotation of documents



Question processing: Questions are processed in three phases, accessing web search engines, local databases, and a semantic network knowledge base.
First phase: The user question is transformed into an IR query, which is delivered to dedicated web search engines and web portals. Results from a web search typically consist of lists of URLs. The web documents referenced by these URLs are retrieved and converted into text.
Second phase: The text passages from the web are segmented and indexed in local databases. The local databases provide access to units of textual information of certain types (chapters, paragraphs, sentences, or phrases).
Third phase: Different modules are employed to produce answer streams, including QAP (Question Answering by Pattern matching), MIRA (Modified Information Retrieval Approach), and the InSicht system. InSicht uses a linguistic parser to analyze the text segments and returns the representation of the meaning of a text as a semantic network. Finding answers with InSicht is based on logical inferences and textual entailments on the annotation of questions and documents with semantic networks. Answer streams are merged and answer candidates are logically validated by a specialized module (MAVE).
Project status:

- implementation of a prototype is completed: QA system IRSAW including semantic annotation consisting of QAP, MIRA, InSicht, MAVE (see Figure)
- first evaluation of prototype on newspaper articles at the Cross Language Evaluation Forum 2006 (combining the answer streams produced by QAP and InSicht):
  one of the best results in the monolingual German question answering track
- second evaluation (published in proceedings of RIAO 2007; combining the answer streams produced by QAP, MIRA and InSicht):
  better results with more answer streams and logical answer validation

Outlook:

- fully implement semantic annotation of web pages as a web service
- investigate support for English language (requires a large English lexicon)
- installation of software in libraries (planned)