

IRSAW – Towards Semantic Annotation of Documents for Question Answering

Johannes Leveling

CNI Spring 2007 Task Force Meeting,
Apr. 16–17, 2007, Phoenix, AZ

IRSAW: *Intelligent Information **R**etrieval on the Basis of a
Semantically **A**nnnotated **W**eb*

funded by the DFG (Deutsche Forschungsgemeinschaft)

Outline

- 1 Introduction
- 2 IRSAW
 - Basic Architecture
 - The MultiNet paradigm
 - Processing Phases
 - Modules
- 3 Results



Motivation

More information becomes available on the internet
and/but precise answers are difficult to find

- Develop a semantically based question answering (QA) framework in the IRSAW project
- General idea:
 - Retrieve documents from the internet
 - Semantically analyze and annotate the question and documents, and
 - Apply deep linguistic methods for question answering on document content



Background and Embedding in a General Strategy

- Semantically based natural language processing
- Knowledge representation MultiNet (concept oriented)
- Support by large semantically oriented computational lexicon → word sense disambiguation
- Homogeneous representation of lexical knowledge, general background knowledge (world knowledge), dialogue context, and meaning of sentences and texts
- Emphasis on semantic relations in all applications (not just concepts or descriptors)



NLI-Z39.50: Beyond Descriptor Search

Predecessor: Natural language interface for the Z39.50 protocol

- Natural language interface to libraries and information providers on the internet
- Transformation of semantic structures into expressions of formal retrieval languages
- Includes features such as de-duplication of results, phonetic search, decomposition of compounds, query expansion with additional concepts
- Example query: *Where do I find books by Peter Jackson which were published in the last ten years with Springer and Addison-Wesley?*



Natürlichsprachliche Expertensuche in Bibliothekskatalogen - Konqueror <@ki212>

Location Edit View Go Bookmarks Tools Settings Window Help

Natürlichsprachliche Expert... Natürlichsprachliche Expert...

Benutzerkennung: **Gast** [Benutzer abmelden](#)

Natürlichsprachliche Anfrage: **Wo finde ich Bücher von Peter Jackson, die in den letzten zehn Jahren bei Springer und Addison-Wesley veröffentlicht wurden?**

Interpretation der Anfrage: (Die Zeichen '(), ']' und '+' wurden zur besseren Lesbarkeit eingefügt)

Materialart : b (Bücher)
Person : jackson, peter
Verlag : springer oder addison-wesley
Erscheinungsjahr : nach oder in 1996

Die Anfrage wurde an die folgende Datenbank geschickt: [GBV](#)

2 Ergebnisse werden angezeigt

1. [GBV]

Titel: Introduction to expert systems
Verfasserschaft: [Jackson, Peter](#)
Verlag: [Addison-Wesley](#)
Verlagsort: Wokingham, Engl. [u.a.]
Erscheinungsjahr: 1996
Umfangsangabe: XVII, 526 S
ISBN: 0-201-17578-9
Ausgabe: 2. ed., reprinted
Schlagwort: International computer science series
Verfügbarkeit: GBV / UB Rostock <28> / I28/BB11 ST 302 J13(2) / 2000.16138

2. [GBV]

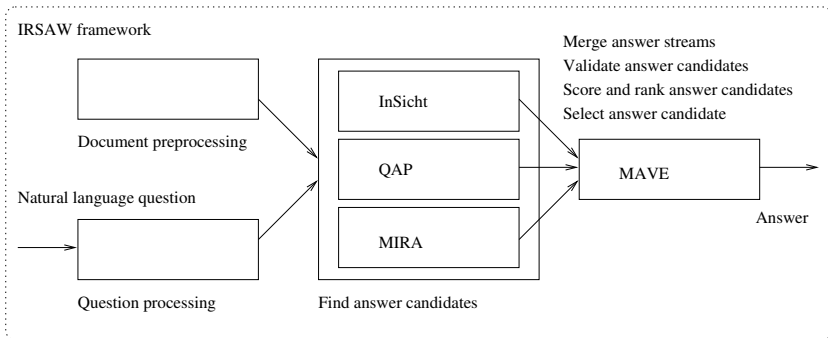
Titel: Geographies of consumption
Verfasserschaft: [Jackson, Peter](#)
Thrift, Nigel
Erscheinungsjahr: 1996
Verfügbarkeit: GBV / MPI ethnol. Forschung <Ha 163>

IRSAW

- Multi staged approach
- Questions are processed in three phases, accessing web search engines, local databases, and a semantic network knowledge base
- Web documents → semantic annotation



Architecture of IRSAW



IRSAW – Methods and Modules (1/3)

- Apply parser (for German language) to produce a semantic network representation of texts (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)



IRSAW – Methods and Modules (2/3)

- Combine different data streams containing answer candidates
- Use different methods to produce answer streams to increase recall and robustness
- Logically validate answers
- Select validated answers from streams of answer candidates to increase precision



IRSAW – Methods and Modules (3/3)

- Natural language generation of answers
 - Allows for rephrasing from text and combination of answer fragments from different documents (state-of-the-art: extracting snippets from the text)
- IRSAW also aims at investigating linguistic phenomena in questions and documents (e.g. idioms, metonymy, and temporal and spatial aspects)



IRSAW – Software

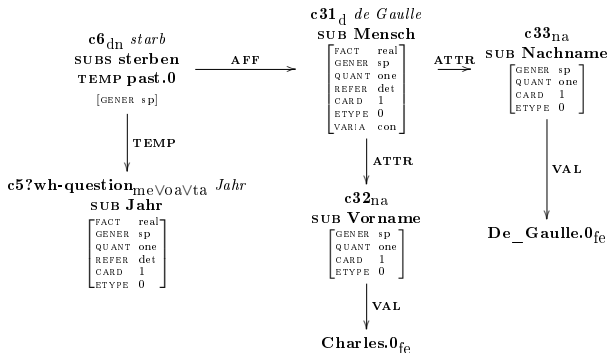
The IRSAW project will result in two software components accessible via internet:

- 1 The question answering system IRSAW
- 2 A web service for the semantic annotation of (web) documents



MultiNet: Example for Semantic Network

*In which year did Charles de Gaulle die?/
In welchem Jahr starb Charles de Gaulle?*



MultiNet: Tools and Resources

- Syntactic-semantic parser:
[WOCADI](#) (Word Class Controlled Disambiguating Parser)
- Large semantic computational lexicon:
[HaGenLex](#) (Hagen German Lexicon)
- Workbench for the computer lexicographer:
[LiaPlus](#) (Lexicon in action)



IRSAW: First Processing Phase

- Transform user question into IR query
- Preselect information resources (→ Broker)
- Send IR query to web search engines and web portals (→ lists of URLs)
- Retrieve web documents referenced and convert them



IRSAW: Second Processing Phase

- Segment and index text passages from the web in local database
- Access to units of textual information of certain types (chapters, paragraphs, sentences, or phrases)



IRSAW: Third Processing Phase

- Employ different modules to produce data streams containing answer candidates:
 - **QAP** (Question Answering by Pattern matching),
 - **MIRA** (Modified Information Retrieval Approach), and
 - **InSicht**
- Merge, rank, logically validate answer candidates and select best answer (**MAVE**)



InSicht

- Analyze text segments (question, texts) with WOCADI and return the representation of the meaning of a text as a semantic network
- Expand queries with semantically related concepts
- High recall
- Paraphrase answer node in semantic network (generate answer)
- Match semantic networks
- High precision
- + Co-reference resolution, logical inference rules/textual entailments



InSicht Logical Entailment

```
( (rule
  (
    (subs ?n1 "ermorden.1.1") ;; kill
    (aff ?n1 ?n2)
    ->
    (subs ?n3 "sterben.1.1") ;; die
    (aff ?n3 ?n2)
  ) )
(ktype categ)
(name "ermorden.1.1_entailment"))
```



InSicht Example

- **User question:** *In which year did Charles de Gaulle die?*
In welchem Jahr starb Charles de Gaulle?
- **Text passage:** *France's chief of state Jacques Chirac acknowledged the merits of general and statesman Charles de Gaulle, who died 25 years ago.*
Frankreichs Staatschef Jacques Chirac hat die Verdienste des vor 25 Jahren gestorbenen Generals und Staatsmannes Charles de Gaulle gewürdigt.
(SDA.951109.0236)
- **Answer:** 1970 (deictic temporal expression resolved; document written in 1995)



QAP - Question Answering by Pattern Matching

- Create patterns by processing known question-answer pairs
- Search for text passages containing keywords from question
- Apply pattern matching on answer candidates
- Extract answer string from variable binding
- + Robustness, high precision for a small class of questions
- No logical inferences possible



QAP Example

- **User question:** *In which year was the Russian Revolution?*
In welchem Jahr fand die russische Revolution statt?
- **Text passage:** *The satire inspired by the Russian revolution 1917 lets the dream of liberty and equality fail because of humans.*
Die von der Russischen Revolution 1917 inspirierte Satire läßt den Traum von Freiheit und Gleichheit an den Menschen scheitern. (FR940612-000533)
- **Answer:** 1917 (pattern matching subsystem ignores metonymy and ellipsis)



MIRA - Modified Information Retrieval Approach

- Train tagger on answer classes (LOC, PER, ORG)
- Search for text passages containing keywords from question
- Use tagger on answer candidate sentence and select most frequent word sequence
- + Highly recall-oriented
- Very low precision, works only for a small class of questions (with answer type LOC, PER, ORG)



MIRA Example

- **User question:** *Who was the first man on the moon?*
Wer war der erste Mensch auf dem Mond?
- **Text passage:** *Twenty-five years ago Neil Armstrong was the first man to step onto the moon, but today manned space flight stagnates.*
Vor 25 Jahren betrat Neil Armstrong als erster Mensch den Mond, doch heute stagniert die bemannte Raumfahrt. (FR940724-001243)
- **Answer:** Neil Armstrong (PER)



MAVE - MultiNet-based Answer Verification

- Validate answer candidates
- Test logical validity of answer candidate (using inferences, entailments)
- Added heuristic quality indicators as fallback strategy
- Select most trusted answer



Evaluations

- InSicht evaluation: best performance for monolingual German question answering task at Cross Language Evaluation Forum 2005 (QA@CLEF 2005)
- IRSAW evaluation at QA@CLEF 2006: combination of InSicht and QAP answer stream: one of the best results in the monolingual German QA track; best results for answer validation task with MAVE
- IRSAW evaluation (for RIAO 2007): InSicht, QAP, MIRA answer streams, and logical validation with MAVE → better results with more answer streams and logical answer validation



Evaluation Results

Results for answer validation of answer candidates for 600 questions (InSicht:I, MIRA:M, QAP:Q; c=correct, i=inexact, w=wrong)

QA streams	c	i	w
IRSAW: I	199.4	10.9	15.7
IRSAW: I+M+Q	244.4	16.9	255.7
IRSAW: I+M+Q (Optimum)	290.0	15.0	215.0



Project Status

- Implementation of modules for QA system IRSAW completed
- Evaluations are now based on corpus of newspaper articles / Wikipedia (corpora of millions of sentences)
- Semantic annotation \Rightarrow Semantic Web



Future Work

- **Next:** Evaluation on web pages (even bigger corpus, dynamic content)
- Add more robustness
- Add acoustic interface (speech input)
- Create English prototype (?)



Selected References

- Ingo Glöckner, Sven Hartrumpf, and Johannes Leveling. Logical validation, answer merging and witness selection – a case study in multi-stream question answering. *To appear in: Proceedings of RIAO 2007*, 2007.
- Sven Hartrumpf and Johannes Leveling. University of Hagen at QA@CLEF 2006: Interpretation and normalization of temporal expressions. In Alessandro Nardi, Carol Peters, and José Luis Vicedo, editors, *Results of the CLEF 2006 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2006 Workshop*, Alicante, Spain, 2006.
- Hermann Helbig. *Knowledge Representation and the Semantics of Natural Language*. Springer, Berlin, 2006.

