Semantic Analysis for NLP-based Applications

Johannes Leveling

former affiliation:
Intelligent Information and Communication Systems (IICS)
University of Hagen (FernUniversität in Hagen)
58084 Hagen, Germany
Outline

Introduction

The MultiNet Paradigm

Applications based on Semantic NLP
  NLI-Z39.50
  IRSAW
  DeLite
  GIRSA-WP

Conclusions
Deep semantic natural language processing

Knowledge and meaning representation MultiNet (concept-oriented) (Hel06)

Supported by large semantically oriented computational lexicon

Important requirements for meaning representation:

- Homogeneity: representation of lexical knowledge, general background knowledge (world knowledge), dialogue context, and meaning of sentences and texts with the same means
- Universality: independent of domain or language
- Cognitive adequate: concept-centered
- Interoperability: applicable to theoretic research of automatic NLP and in modules of applied AI systems
MultiNet: Meaning Representation of Text

MultiNet (Multilayered Extended Semantic Networks) characteristics:

- **concepts**: lexicalized and non-lexicalized, e.g. “c134”, “New_York.0”, “play.1.1”, “play.1.2”, “play.2.1”

- **semantic relations/functions**, e.g. AGT (agent), OBJ (neutral object), DUR (duration), ORNT (orientation), *IN (location-generating function)

- **layer features**, e.g. FACT (facticity of a concept), REFER (determination of reference), QUANT (quantificational content)

- **semantic sorts**, e.g. $d$ (discrete object), $ta$ (temporal abstractum)
**MultiNet: Selected Semantic Relations**

<table>
<thead>
<tr>
<th>Relation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASSOC</td>
<td>association</td>
</tr>
<tr>
<td>ATTCH</td>
<td>attachment of object to object</td>
</tr>
<tr>
<td>CHPA</td>
<td>change of sorts (property $\rightarrow$ abstract object)</td>
</tr>
<tr>
<td>EXP</td>
<td>experiencer</td>
</tr>
<tr>
<td>MCONT</td>
<td>an informational process or object</td>
</tr>
<tr>
<td>OBJ</td>
<td>neutral object</td>
</tr>
<tr>
<td>PRED</td>
<td>predicative concept specifying a plurality</td>
</tr>
<tr>
<td>PROP</td>
<td>property relationship</td>
</tr>
<tr>
<td>PARS</td>
<td>meronymy</td>
</tr>
<tr>
<td>SCAR</td>
<td>carrier of a state</td>
</tr>
<tr>
<td>SSPE</td>
<td>state specifier</td>
</tr>
<tr>
<td>SUB</td>
<td>conceptual subordination for objects</td>
</tr>
<tr>
<td>SUBS</td>
<td>conceptual subordination for situations</td>
</tr>
<tr>
<td>SYNO</td>
<td>synonymy</td>
</tr>
<tr>
<td>TEMP</td>
<td>temporal restriction for a situation</td>
</tr>
<tr>
<td>*ALTN1</td>
<td>an introduction of alternatives</td>
</tr>
</tbody>
</table>
MultiNet: Tools and Resources

- WOCADI (Word Class Controlled Disambiguating Parser): Syntactic-semantic parser (Har03)
- HaGenLex (Hagen German Lexicon): Large semantic computational lexicon (HHO03)
- LiaPlus (Lexicon in action): Workbench for the computer lexicographer (Oss04)
WOCADI: Semantic Analysis

- WOCADI parser produces semantic network representation from (German) texts, including
  - resolution of anaphoric references (e.g. Peter = he),
  - analysis of idioms, support verb constructions (e.g. kick the bucket = lose one’s life = die),
  - structural and semantic decomposition of compound nouns and adjectives (e.g. swimming pool vs. Schwimmbecken),
  - identification of metonymy (lexicon support via meaning facets),
  - analysis of deictic expressions (e.g. temporal: yesterday)

- Applied to large corpora,
  e.g. CLEF-NEWS newspaper corpus (275,000 articles) and German Wikipedia (2006: 500,000 articles, 12 million sentences; 2009: 20 million sentences)

- Coverage: full semantic network for 54% of sentences, partial semantic network (chunks) for 34%
In which year did Charles de Gaulle die?
In welchem Jahr starb Charles de Gaulle?
Finde Dokumente, die über psychische Probleme oder Stress von Prüfungskandidaten oder Prüflingen berichten. (GIRT topic 116)
“Find documents reporting on mental problems or stress of exam candidates or examinees.” (GIRT topic 116)
HaGenLex: The Computational Lexicon

- HaGenLex is a semantically oriented (German) lexical resource
- Consists of multiple lexicons:
  - full morpho-syntactic and semantic information (30,000 entries),
  - additional flat lexicon (50,000 entries),
  - name lexicons (350,000 entries in 50 classes)
  - compound lexicon (about 500,000 entries; structure and semantics),
HaGenLex:
Sample Concepts

- **essen.1.1** (eat):
  (Der Student) (ißt) (eine Schokolade).
  (The student) (eats) (a bar of chocolate).

- **essen.1.2** (eat [one’s fill]):
  (Der Student) (ißt) sich (satt).
  (The student) (eats) his (fill).

- **essen.2.1** (food):
  Das Kind hat kein Essen bekommen.
  The child did not get any food.

- **essen.2.2** (diner):
  Das Essen am Abend dauerte 2 Stunden.
  The diner in the evening lasted 2 hours.

- **fressen.1.1** (eat):
  (Der Hund) (frißt) (einen Knochen).
  (The dog) (eats) (a bone).

- **fressen.1.2** (be crazy about sth.):
  (Die Großmutter) (frißt) (einen Narren) (an den Blumen).
  (Grandmother) (is crazy about) (flowers).
HaGenLex: Excerpt from Entry essen.1.1 (eat)
HaGenLex: Excerpt from Entry *fressen.1.1* (eat)
Outline

Introduction

The MultiNet Paradigm

Applications based on Semantic NLP
  NLI-Z39.50
  IRSAW
  DeLite
  GIRSA-WP

Conclusions
NLI-Z39.50: Beyond Descriptor Search

Natural language interface for the Z39.50 protocol (Lev06)

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

<table>
<thead>
<tr>
<th>Materialart</th>
<th>b (Bücher)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>jackson, peter</td>
</tr>
<tr>
<td>Verlag</td>
<td>springer oder addison-wesley</td>
</tr>
<tr>
<td>Erscheinungsjahr</td>
<td>nach oder in 1996</td>
</tr>
</tbody>
</table>

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:** XVI, 526 S
   - **ISBN:** 0-201-17578-9
   - **Ausgabe:** 2. ed., reprinted
   - **Schlagwort:** International computer science series
   - **Verfügbarkeit:** GBV / UB Rostock <28> / 128/BB1! ST 302 J13(2) / 2000.16138

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

Introduction

The MultiNet Paradigm

Applications based on Semantic NLP
  NLI-Z39.50
  IRSAW
  DeLite
  GIRSA-WP

Conclusions
IRSAW (Information Retrieval based on a Semantically Annotated Web) (GHL07)

- **Question answering system** using a combination of answer candidate streams
- Also includes a web service for the automatic semantic annotation of (web) documents (RDF/S, OWL)
- Document collections: Wikipedia, CLEF-NEWS, etc.
Apply WOCADI parser (for German) to produce semantic network representation of documents and questions (MultiNet)

→ Allows a full semantic interpretation on which logical inferences are based (state-of-the-art: mostly statistical methods or shallow semantics)
▶ Produce multiple streams of answer candidates with different techniques (ranging from pattern matching to deep semantic analysis)
▶ Combine data streams containing answer candidates
→ Different methods to produce answer streams increase recall and robustness
▶ Logically validate answers
→ Select validated answers from streams of answer candidates to increase precision
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 linguistic phenomena in questions and documents (e.g. idioms, metonymy, and temporal and spatial aspects)
IRSAW: Processing Phases

- 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)
- Employ different methods to produce data streams containing answer candidates, including
  - InSicht (MultiNet-based QA)
  - QAP (Question Answering by Pattern matching), and
  - MIRA (Modified Information Retrieval Approach)
- 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
( (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") )
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.*  
(SDA.951109.0236)

Answer: 1970 (deictic temporal expression resolved; document written in 1995)
QAP: Question Answering by Pattern Matching

- Training phase: generate patterns by processing known question-answer pairs
- Retrieve 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 explicit logical inferences possible
QAP: Pattern Matching Example

- **NL Question:** “Where was Galileo Galilei born?”
- **IR query:** `Galileo_Galilei´/1.0, born/0.7`
- **Text passage:** “Galileo was born in Pisa, in the Tuscany region of Italy on February 15, 1564.”
- **Tagged and tokenized text passage:**
  
  `NAME “was” LWORD appo “Pisa” $comma appo art “Tuscany” “region” art “Italy” $colon`

- **Pattern:** `?words1* NAME ?w0 LWORD appo ?answer+ $comma appo art ?w1 ?words2*`
- **?answer+ = “Pisa”**
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

- Apply a special tagger for answer classes (LOC, PER, ORG etc.)
- Retrieve text passages containing keywords from question
- Use tagger on answer candidate sentence and select most frequent word sequence
  + Highly recall-oriented
  - Low precision, works only for a small class of questions (factoid questions)
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
IRSAW 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
- IRSAW at QA@CLEF 2008: two additional answer streams (FACT, SHASE) → more robustness by diversity of answer candidate producers
Results for answer validation of answer candidates for 600 questions (InSicht:I, MIRA:M, QAP:Q; c=correct, i=inexact, w=wrong) (GHL07)

<table>
<thead>
<tr>
<th>QA streams</th>
<th>c</th>
<th>i</th>
<th>w</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRSAW: I</td>
<td>199.4</td>
<td>10.9</td>
<td>15.7</td>
</tr>
<tr>
<td>IRSAW: I+M+Q</td>
<td>244.4</td>
<td>16.9</td>
<td>255.7</td>
</tr>
<tr>
<td>IRSAW: I+M+Q (Optimum)</td>
<td>290.0</td>
<td>15.0</td>
<td>215.0</td>
</tr>
</tbody>
</table>
Outline

Introduction

The MultiNet Paradigm

Applications based on Semantic NLP
- NLI-Z39.50
- IRSAW
- DeLite
- GIRSA-WP

Conclusions
Text readability checker DeLite (vL07), developed in the BenToWeb project (developing tools and guidelines for accessibility of web sites)

Classic readability scores for text are based on shallow measures, i.e. average sentence length and average word length (e.g. Flesh reading ease score)

DeLite incorporates text analysis of text on different linguistic levels:
- morphological, lexical, syntactic, semantic, discourse level

→ Definition of readability indicators
→ Annotation of text sections (document, sentence, phrase, word) with indicator values (e.g. number of possible anaphoric reference candidates)
→ Computation of global readability score
→ Identification of text passages which are difficult to read
DeLite - A linguistic checker for text readability


Readability indicators:

- Understandability index (Aristad): 34.17
- Number of words: 108
- Number of syllables: 236
- Number of sentences: 6
- Average sentence length: 18
- Ratio of abstract concepts: 0.29
- Ratio of derived nouns: 0.3
- Type-token ratio (lemmata): 0.61
- Type-token ratio (word forms): 0.69

XML report R1 (text structure)
XML report R2 (indicators)
XML report R3 (scores and weights)
Er ist f"ur das Untersuchungsgebiet nachzuweisen.

Das Untersuchungsgebiet
Outline

Introduction

The MultiNet Paradigm

Applications based on Semantic NLP
  NLI-Z39.50
  IRSAW
  DeLite
  GIRSA-WP

Conclusions
GIRSA-WP is a Geographic Information Retrieval (GIR) system combining methods from question answering (QA) and information retrieval (IR) (HL09).

InSicht question answering system
(participated at QA@CLEF 2004–2008)

+ GIRSA geographic information retrieval system
(participated at GeoCLEF 2006–2008)

= GIRSA-WP combination of methods
(participated at GikiP 2008, GikiCLEF 2009)
“What capitals of Dutch provinces received their town privileges before the fourteenth century?”

→ “Name capitals of Dutch provinces.”

→ “Name Dutch provinces.”

= Zeeland (support from article 1530: Besonders betroffen ist die an der Scheldemündung liegende niederländische Provinz Zeeland.)

→ “Name capitals of Zeeland.”

= Middelburg (support from article Miniatuur Walcheren: ... in Middelburg, der Hauptstadt von Seeland (Niederlande).)

= Middelburg (answer to revised question can be taken without change)

→ “Did Middelburg receive its town privileges before”

“the fourteenth century?”

= Ja./“Yes.” (support from article Middelburg: 1217 wurden Middelburg durch Graf Willem I. ... die Stadtrechte verliehen.)

= Middelburg (support: three sentences, from three articles, see above)
Conclusion

- Applications based on semantic networks (MultiNet) have been successful in evaluations in completely different domains, using the **same means for meaning representation** (no need to train a model)
  → Interoperability is a plus
- User interactions allow for **queries** or questions.
  → Most methods (e.g. part-of-speech tagging, language detection, machine translation, parsing) are not optimized for queries!
- Statistical NLP or shallow (syntax-based) NLP often is not enough for complex applications and deep semantic analysis often does not provide enough coverage
  → **Combination** of (many) different approaches results in better performance
Selected References (1/2)


