Navigation through the Meaning Space of HUMI NT Reports

Dr. Matthias Hecking
Forschungsgesellschaft für Angewandte Naturwissenschaften e.V. (FGAN)
Forschungsinstitut für Kommunikation, Informationsverarbeitung und Ergonomie (FKIE)
Abtl. Informationstechnik und Führungssysteme (ITF)
Neuenahrer Straße 20
53343 Wachtberg-Werthhoven
hecking@fgan.de
Content

1. Introduction
2. Information Extraction
3. Project ZENON
4. Meaning Space Navigation
5. Summary
1 | Introduction - I

General problem: There are a lot of natural language texts (military reports, emails, web pages, scientific reports, documents, ...) which can't be evaluated due to missing specialists.

Which technical possibilities exist of automating the content extraction?

⇒ Practical approach: Information Extraction (IE)
Specific problem: content extraction of HUMINT reports

**ZENON project:** The overall objective is to realize an experimental system for (partial) content extraction of HUMINT reports from the KFOR deployment of the Bundeswehr and to realize a possibility to evaluate the formal representation of the content.

For the realization of the IE module approx. 4000 English HUMINT reports are available.
For the realization the toolbox GATE is used.

The ZENON prototype will be integrated into the "DB Einsatz!"

For the evaluation of the ZENON IE the KFOR text corpus was developed.
Information extraction (IE) is the task of identifying, collecting and normalizing information from natural language text.

Relevant information about the Who, What, When, etc. is looked for.

The information of interest is described through domain-specific lexicon rules and patterns called templates.

During the IE task these templates are filled with the collected information.

The templates are domain and task specific, i.e. for each new task and domain they must be newly created.
„... in TUZLA:
10.20 a.m.
18 vehicles
- 8 with attached ZIS–3 and 1 with attached T 12
march
at road crossing (CQ 072368) south of MILESKIJ (CQ 0737) to the north.“
KIE Forschungsinstitut für Kommunikation, Informationsverarbeitung und Ergonomie

Dr. M. Hecking

2| IE - III

addressee ...

message

objects

vehicles

area

location

move

report

... DB

Analysis

who, when, where

Entity-Activity-Net

FGAN Forschungsinstitut für Kommunikation, Informationsverarbeitung und Ergonomie
Technical basis of the information extraction:

- **extensive linguistic knowledge**
  - general and application-specific lexica
  - general and/or application-specific phrase grammars
  - general (and/or application-specific) clauses and sentence grammars
- **cascaded transducers**, i.e. finite state automaton that reads from the input and writes to the output
- only application-relevant parts of the texts are analyzed through transducers (shallow parsing techniques)
The experimental ZENON project:

- The **Information Extraction (IE)** technology is used for the content extraction.
- The information about the **actions** and **named entities** are identified from each sentence and the content of the sentences are formally represented in **typed feature structures**.
- These structures can be combined and presented in a graphically navigatable **Entity-Action-Network**.
3 | ZENON Project - II

Dr. M. Hecking

HUMINT reports → * Tokenizing
* Sentence splitting
* POS tagging
* Gazetteer
* NE Recognition
* Morphological analysis

Sentences with annotations → * Detect verb phrases
* Extract action types
* Extract sentence content

Processing results in XML → * Select extracted information
* Combine extracted information

GATE

Information Extraction Presentation System (IEPS)
GATE:

- "is one of the most widely used human language processing systems in the world."
- "comprises an architecture, framework (or SDK) and graphical development environment ..."
- "... has been under construction in Sheffield since 1995."
- "The system has been used for many language processing projects; in particular for Information Extraction in many languages."
- "GATE is funded by the Engineering and Physical Sciences Research Council (EPSRC), the EU and commercial users."
- http://gate.ac.uk/
3 | ZENON Project: Extraction of Named Entities

- Chunk parsing of named entities (NE): *City, Company, Coordinates, Country, CountryAdj, Currency, Date, GeneralOrg, MilitaryOrg, Number, Percent, Person, PoliticalOrg, Province, Region, River, Time and Title*

- Example:

  ```
  Rule: PersonName1
  ( ( ({PersonTitle}):title ({Token.string == "."})? )
  ( ( {Lookup.majorType == person_first}:firstName )?
  ( {Token.category == NNP, Token.orth == allCaps}:lastName
  ( {Token.string == ":"} {Token.category == NNP})? )
  ):person
  -->
  {...}
  ```
Determine action types:

- extraction of **verb phrases** (modal verb phrases, participles, special composed verb expressions)
- mapping from recognized verb groups to **action types** (e.g., from the infinitive of 'murder', 'kill', 'decapitate', … to action type 'kill').
Sentence analysis:

- The basic structures for the semantic sentence analysis are given by the FrameNet-Project.
- Semantic roles are specified for the lexical units (frames, verbs).
- Example: Frame Killing
  - Def.: A KILLER or CAUSE causes the death of the VICTIM.
  - Roles: CAUSE, KILLER, VICTIM, DEGREE, INSTRUMENT, ...
  - Sentence: [John Mueller and four other persons VICTIM] were killed in [an explosion incident CAUSE] in [GOSTIVAR area PLACE].
  - Formal representation …
4 Meaning Space Navigation - I

Visualization techniques:

- hierarchies vs. graphs/networks
- hierarchies: file-browser
- hyperbolic browser
- level of detail (LOD) concept
- themescape
- treemap
Explore the meaning space:

- select and combine the analysis-specific information (filter)
  - realized with: XSLT
  - in XML format
- access and navigate the Entity-Action-Network
  - based on: TouchGraph
- faster access to a set of HUMINT reports
Information Extraction Presentation System (IEPS):

- Graphical visualization of information typically extracted from free-form texts
- **Scenario:** Name, sets of input and output files (XML coded feature structures), filter.
- **Filter:** Description of the transformation with XSLT
- a *technical* possibility to realize navigation through the meaning space
4| Meaning Space Navigation - IV

Dr. M. Hecking
Entity-Action-Networks:
- a *conceptual* possibility to realize navigation in a meaning space
- complex filters determine:
  - contained semantic basic units
  - how they are combined
  - contained additional information
4 | Meaning Space Navigation - VI

Meaning space of one document:

an explosion incident
... when many innocent people died because of violence and atrocities

unknown perpetrators

KFOR_report

HANNOVER

K-ALBANIAN

01080112au

security situation

Berger

Mueller

Mueller, Hans

Schmidt

all_sentences

locations

actions

kill

die

know

causes

FGAN

Forschungsinstitut für Kommunikation, Informationsverarbeitung und Ergonomie
4 | Meaning Space Navigation - VII

Time-action meaning space:
Location-action meaning space:
5| Summary

- The field of information extraction (IE) was explained.
- Our research project ZENON was presented.
- In the main part of the talk, the navigation through the meaning space of HUMINT reports was illustrated in detail.

- Future work on the ZENON prototype goes on.
- **Main question:** Is the IEPS approach appropriated for presenting the results of analyzing a large quantity of documents?