11th ICCRTS
COALITION COMMAND AND CONTROL IN THE NETWORKED ERA

Navigation through the Meaning Space of HUMINT Reports

Dr. Matthias Hecking
FGAN/FKIE
Neuenahrer Straße 20
53343 Wachtberg-Werthhoven
Germany

Phone: +49 228 9435 576
Fax: +49 228 9435 685
hecking@fgan.de

Navigation through the Meaning Space of HUMINT Reports

Dr. Matthias Hecking
FGAN/FKIE
Neuenahrer Straße 20
53343 Wachtberg-Werthhoven, Germany
hecking@fgan.de

Abstract

The new deployments of the German Federal Armed Forces cause the necessity to analyze large quantities of Human Intelligence (HUMINT) reports. These reports are characterized by a large topical and linguistic variety. Therefore, they are good candidates for applying natural language processing techniques. In this paper, an approach to realize a graphical navigation through the meaning of HUMINT reports is presented. This meaning space navigation is realized as a graphically navigatable Entity-Action-Network and is part of the ZENON project. In this project an information extraction approach is used for the (partial) content analysis of English HUMINT reports from the KFOR (Kosovo Force) deployment of the Bundeswehr. The information about the actions and named entities are identified from each sentence and the content of the sentences is represented in formal structures. These structures can be combined and presented in the network. After a short introduction, the information extraction as an approach to process natural language and the ZENON project are explained. In the main part of the paper, the navigation through the meaning space of HUMINT reports is described in detail. Different techniques for doing this are presented. The Information Extraction Presentation System is used to realize the Entity-Action-Network. This presentation system is introduced and various examples are given.

1. Introduction

On one hand, the new deployments of the German Federal Armed Forces cause the necessity to analyze large quantities of HUMINT reports. These reports are characterized by a large topical and linguistic variety. For that reason, they are good candidates for applying techniques from computational linguistics to analyze natural languages. On the other hand, the *processing of human language* was identified as a critical capability in many future military applications (cf. [Steeneken, 1996]). Especially the *content analysis* of free-form texts is important for any information operation of the Network Centric Warfare (NCW) concept (s. [NCW, 2001], p. 5-15). We set up the research project ZENON[1], in which an information extraction approach is used for the (partial) content analysis of English HUMINT reports from the KFOR deployment of the Bundeswehr. The overall objective of this research is to realize a *graphically navigatable Entity-Action-Network*. 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 the navigatable network.

[1] according to: Zenon of Citium, 336 BC - 264 BC, philosopher, founder of the Stoicism

The ZENON project is based on the results of the former SOKRATES[2] project. In this project we applied information extraction to the analysis of German free-form battlefield reports (cf. [Casals, 2004a], [Casals, 2004b], [Frey, 2004], [Hecking, 2001], [Hecking, 2002], [Hecking, 2003a], [Hecking, 2003b], [Hecking, 2004a], [Hecking, 2004b], [Hecking, 2004c], [Schade, 2003a], [Schade, 2003b], [Schade, 2006]). The SOKRATES prototype was able to process written battlefield reports (e.g., messages about hostile movements, deployments) in German. The reports were analyzed, represented in feature structures and semantically enhanced with the help of an ontology. With the SOKRATES prototype we showed the general applicability of the *Information Extraction* (IE) technology for military purposes.

This paper is structured as follows. First, a short introduction into the information extraction approach is given. Then, the ZENON project is described. In the main part of the paper, the navigation through the meaning space of HUMINT reports is described in detail. Different techniques for doing this are presented. The Information Extraction Presentation System is used to realize the Entity-Action-Network. This presentation system is introduced and various examples are given.

2. Information Extraction

In the last decades various techniques for processing spoken and written natural languages were developed (e.g. speech recognizer in dictation systems, machine translation, grammar checking). IE is an engineering approach (cf. [Appelt, 1999]) for content analysis of free-form texts based on results of computational linguistics. Each IE system is tailored to a specific domain and task. IE uses a *shallow syntactic approach* (cf. [Hecking, 2003b]), i.e. that only parts of the sentences (so-called ‘chunks’) are processed with finite state automatons or *transducers*.

During the IE relevant information about the Who, What, When, etc. in natural language texts is identified, collected, and normalized (cf. [Pazienza, 1999], [Hecking, 2004a]). The relevant information is described through patterns called *templates*. These domain and task specific templates represent the meaning of the relevant information. During the IE task the templates are filled with the extracted information. One possibility to realize the templates is to use *typed feature structures* (cf. [Hecking, 2004b]). Therefore, IE can be seen as the process of normalizing free-form text into a defined semantic structure.

To realize an IE system, language-specific resources (lexicon, grammar) and appropriated parsing software are necessary.

In order to achieve robust and efficient IE systems, domain knowledge must be integrated and shallow algorithms must be used. The domain knowledge is tightly integrated with the language knowledge, e.g., the name ‘Leopard’ in the lexicon has the categorical information ‘tank’. This association between words and semantic information is domain-specific and has to be change for other applications.

[2] according to: Socrates, 469 BC - 399 BC, philosopher

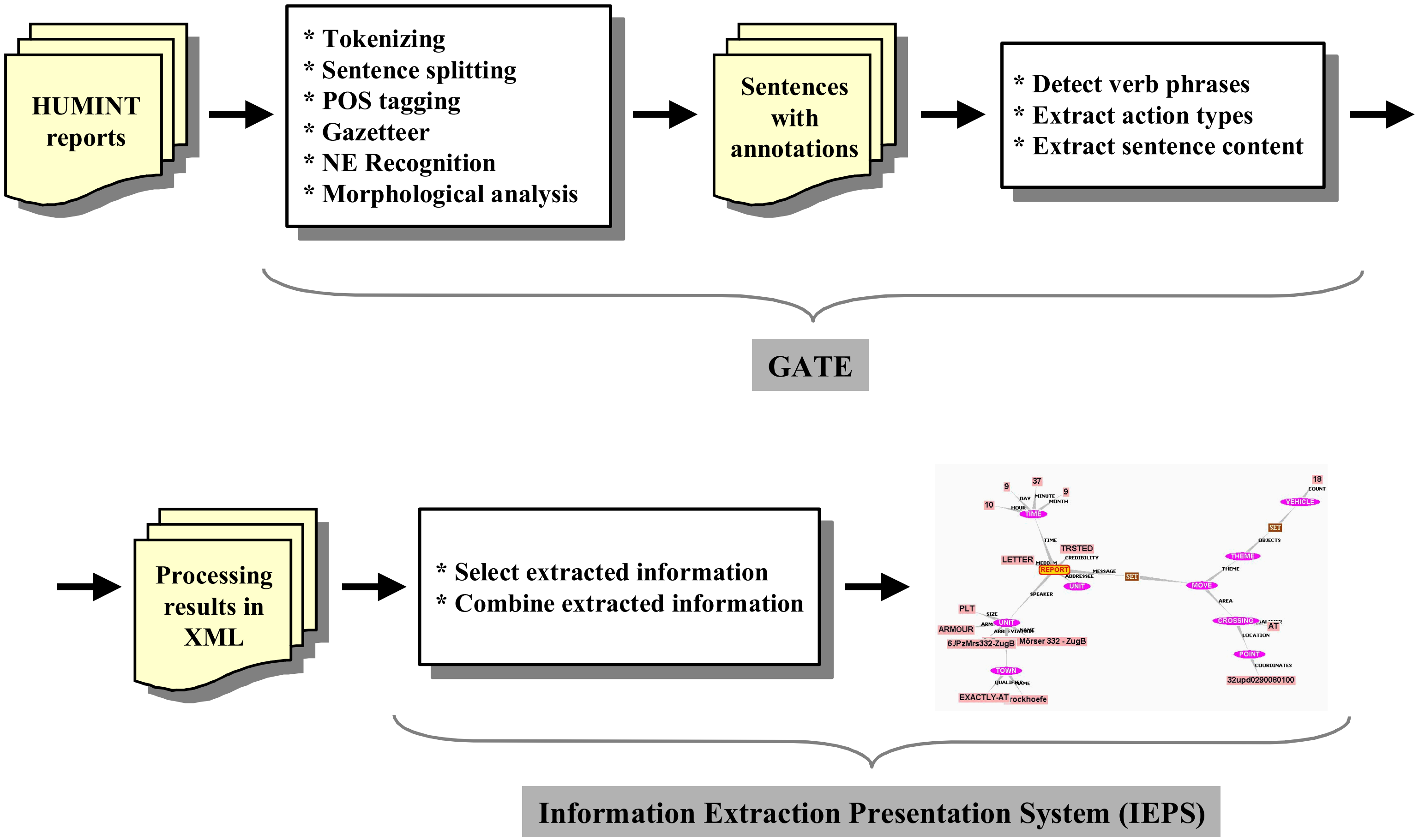

Figure 1: The ZENON processing chain

The IE process itself is divided into sub steps. After tokenizing the text, the sentence boundaries must be identified. Then, the morphological component identifies the word stems, the abbreviation, and detects the syntactic information (e.g., grammar case and gender). After this, the chunk parsing with transducers selects parts of the natural language text that are relevant for the specific information extraction task. The chunks are then used to instantiate the templates, which represent the action/event descriptions. They are the result of the IE process.

The IE is used as the core natural language processing technique in the ZENON project.

3. The ZENON Project

Starting with English HUMINT reports (and a list of the city names) from the KFOR deployment of the German Federal Armed Forces we have realized in our ZENON project an experimental prototype that is able to do a (partial) content analysis of these reports (cf. [Hecking, 2005a], [Hecking, 2006a]). The content of these KFOR reports are from a wide spectrum. Apart from descriptions of conflicts between ethnic groups, tensions between political parties, information about infrastructure problems, etc. there are also reports, which concern individuals or other entities. Statements of the form ´A meets B´, ´A marries C´, ´A shoots B´, etc. contains information about activities/events and involved entities. This information, completed with location and time data, is combined into a graphically navigatable Entity-Action-Network (e.g.; with a person in the center of the network). The intelligence analysts can use this network to navigate through the meaning space of the reports.

Since most of the reports are in English, GATE (General Architecture for Text Engineering, cf. [Cunningham, 2002]) was selected as the used toolbox. GATE is an architecture, a free open source framework (SDK) and graphical development environment for Natural Language Engineering and offers a lot of tools, which are used to realize the natural language processing parts of the ZENON prototype (e.g., morphological analyzer, part-of-speech (POS) tagger, pre-defined transducer to recognize English verbal phrases, chunk-parsing). The functionality to select and combine the extracted information from different sentences and different reports is realized by the *Information Extraction Presentation System* (IEPS, cf. [Casals, 2005]). IEPS is a graphical tool to visualize information extracted from free-form texts.

In Figure 1 the ZENON processing chain is shown. HUMINT reports are fed into the first sub-component. In this component the natural language text is tokenized (i.e., find words, numbers, etc.), the sentence boundaries are detected, the part-of-speech (i.e., whether it's a noun, a verb, etc.) is determined, simple names of cities, regions, military organizations etc. are annotated (through the Gazetteer), named entities (i.e., complex names of e.g. political organizations, person names, etc.) are recognized and a morphological analysis is done. The results of this sub-component are the annotated sentences of the reports. The second sub-component uses these annotations to extract the action type (e.g., 'kill') starting with the verb of the sentence. If the action type is determined the other parts of the sentence (e.g., subject, object, time expressions) are located and formally represented in *typed feature structures*. These structures are coded in XML (Extensible Markup Language) format and represent the output of the natural language part of the ZENON prototype. In the third sub-component (IEPS) the extracted content of different reports can be combined and selected according to predefined XSLT (Extensible Stylesheet Language Transformation) sheets. The result of the analysis can be navigated interactively.

An important processing step during the natural language processing is the recognition of the domain- and application-specific named entities. In the ZENON prototype transducers for the recognition of the following named entities were developed: *City, Company, Coordinates, Country, CountryAdj, Currency, Date, GeneralOrg, MilitaryOrg, Number, Percent, Person, PoliticalOrg, Province, Region, River, Time* and *Title*.

Another important step is the extraction of verb phrases, action types and the sentence content. The ZENON prototype uses various transducers to recognize finite and non-finite verbal phrases, modal verb phrases, participles and special composed verb expressions.

Based on the recognized verb groups, different action types can be detected (e.g., from the infinitive of 'murder', 'kill', 'decapitate', … the action class 'kill'). After detecting the action type the verb phrase and other parts of the sentence must be combined. In the ZENON project we use the *Semantic Frames* from the FrameNet project (cf. [FrameNet]) to realize this combination. Semantic Frames are schematic representations of situation types (eating, killing, spying, classifying, etc.) together with lists of the kinds of participants, objects, and other conceptual roles that are seen as components of such situations.

During the processing, the associated Semantic Frame is inferred from the detected action type. With the identified Semantic Frame the core and non-core frame elements are give. Recognized named entities, POS tagging and expressions from the sentences are used to fill in the frame elements. The filled-up Semantic Frame and other information from the processing of the natural language text represent the result of the first sub-component and are coded as typed feature structures. In Figure 2 an example is given.

For a more detailed description of the above described processing steps see [Hecking, 2006a].

4,498 military reports (mostly in English) from the KFOR deployment of the German Federal Armed Forces were used for the realization of the ZENON prototype. From these reports 800 were manually annotated and form the *KFOR Corpus*[3]. This corpus is a specialized micro text corpus (cf. [McEnery, 2001, p. 191]). The corpus covers 886,000 tokens and contains the annotations in different layers (cf. [Hecking, 2005b]).

4 Meaning Space Navigation

The natural language processing module of the ZENON prototype creates for each sentence in each KFOR report a formal representation of the content. This contains information pieces about activities, events, entities, times and places. These pieces are now put together according to specific analysis requirements (e.g., all information about a specific person). The intelligence analyst must be able to access and explore this *meaning space*. For this,

- the analysis-specific information must be preprocessed (select and combine), and
- the result must be accessible and navigatable.

The result of this is a *graphically navigatable Entity-Action-Network*. The intelligence analyst can use this network for faster access the important information from the used set of reports.

[3] Since the KFOR corpus is classified, it is not freely available.

The question of how to present, access and navigate the meaning space is handled in various disciplines dealing with visualization and navigation (e.g., information visualization, document visualization, thematic navigation, topic maps, topic structure).


```xml
<?xml version="1.0" encoding="UTF-8"?>
<KFOR-report xsi="http://www.w3.org/2001/XMLSchema-instance" …>
  <docID>01080112au</docID>
  <title>Max Mueller, member of the …, killed in a weapon accident in ERTAED.</title>
  <documentContent>...</documentContent>
  <source>...</source>
  <reportType>HUMINT</reportType>
  <deployment>KFOR</deployment>
  <allSentenceContent>
    <list>
      <action>
        <type>kill</type>
        <causeAll>this accident</causeAll>
        <rule>killAction1</rule>
          <sentenceContent>… with MUELLER … were killed by this accident.
          </sentenceContent>
        <verbGroup>
          <infinitive>kill</infinitive>
          …
        </verbGroup>
        <victims>
          …
        </victims>
      </action>
      <action>
        <type>kill</type>
        <causeAll>an explosion incident</causeAll>
        <places>
          <set>
            <city>
              <name>HANNOVER</name>
              <rule>City</rule>
            </city>
          </set>
        </places>
        <rule>killAction2</rule>
          ...
  </allSentenceContent>
  <allCities></allCities>
  <allCoordinates></allCountries>
  <allCountryAdjs></allCountryAdjs>
  <allDates></allDates>
  …
  <allPersons></allPersons>
  <allPersonTitles></allPersonTitles>
  <allPoliticalOrgs></allPoliticalOrgs>
  <allTitles></allTitles>
</KFOR-report>
```


Figure 2: Example result from the natural language part of ZENON (abbreviated**)**

Visualization techniques can be categorized in those representing the data in *hierarchies* and those using graphs or *networks*. Based on this distinction a wide variety of possible visualization technique is available (cf. [Neumann, 2005]):

- *Hierarchies* are often visualized using interactive trees. A typical example for such a tree structure is the file-browser.
- A special form of a hierarchic tree is the *hyperbolic browser* (cf. [Inxight, 2006]).
- In the *Level of Detail (LOD) concept* important objects appear first, less important ones only after zooming into the meaning space. Examples of this technology are graphical front-ends of search engines (e.g., cf. [Kartoo, 2006]).
- A *themescape* is a thematic terrain where the elevation indicates theme strength. Peaks indicate where concentrations of closely related objects appear. Using this map metaphor, spatial navigation and analysis tools can be used to explore the landscape of concepts (e.g., cf. [MicroPatent, 2006]).
- A *treemap* is a nested two-dimensional representation of a multi-level hierarchy. Each data element is represented as a cell. The cell arrangement, size, text labels and color each represent an attribute of that element. The lower hierarchy level is nested into the cell of the ancestor (cf. [Johnson, 1991]).

The optimal technique for the navigation in a meaning space depends on the nature and the volume of the data to be visualized. Various toolsets are available to realize the meaning space navigation, e.g., OpenDX (cf. [OpenDX, 2006]) or TouchGraph (cf. [TouchGraph, 2006]).

The ZENON natural language processing sub-module delivers the basic semantic units (named entities, English verbal phrases, action types) that can be used for the processing and visual navigation. The basic units are selected and combined through filters. These complex filers can be defined for each scenario in the ZENON prototype (e.g., see Figure 3). The filter functionality is realized through *Extensible Stylesheet Language Transformation* (XSLT, cf. [XSL, 2006]). The result of the transformation is also in XML format (e.g., see Figure 4).

```
<xsl:stylesheet version="2.0" xmlns:xsl="http://www.w3.org/1999/ XSL/…">
  <xsl:output method="xml" encoding="UTF-8" indent="yes"/>
  <xsl:template match="/">
    <Zeitpunkte_Aktionen xmlns:xsi="http://www.w3.org/2001/XML...instance">
      <Dokumente>
        <xsl:for-each select="/reports/report">
          <docID>
            <xsl:apply-templates select="document(@filename)/KFOR-…docID"/>
          </docID>
        </xsl:for-each>
      </Dokumente>
      <xsl:call-template name="dates-actions"/>
    </Zeitpunkte_Aktionen>
  </xsl:template>
  <xsl:template name="dates-actions">
    <xsl:for-each select="/reports/report/document(@filename)/KFOR-…/date">
      <xsl:sort select="."/>
      <Zeitpunkt>
        <Wert><xsl:value-of select="concat(year, '-', month, '-', day)"/></Wert>
        <Aktionen>
          <xsl:element name="{../../../type}">
            <docID>
              <xsl:value-of select="//docID"/>
            </docID>
          </xsl:element>
        ...
</xsl:stylesheet>
```

Figure 3: Filter for 'time_action' relation (abbreviated**)**

```
<?xml version="1.0" encoding="UTF-8"?>
<Zeitpunkte_Aktionen xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
 <Dokumente>
   <docID>01080112au</docID>
   …
 </Dokumente>
 <Zeitpunkt>
   <Wert>01-jul-23</Wert>
   <Aktionen>
     <propose>
       <docID>01080112au</docID>
     </propose>
   </Aktionen>
 </Zeitpunkt>
 <Zeitpunkt>
   <Wert>01-jul-30</Wert>
   <Aktionen>
     ...
   </Aktionen>
 </Zeitpunkt>
</Zeitpunkte_Aktionen>
```

Figure 4: Result of applying an XSLT stylesheet (abbreviated**)**

The result of the transformation is visualized by the sub-component IEPS. IEPS is a *technical* possibility to realize navigation through the meaning space. It is a graphical software tool (see Figure 5) for visualizing information typically extracted from free-form texts by a natural language processing system. Additionally, it offers a framework to organize all the files being employed during the processing in user-defined scenarios and to activate the IE process. IEPS represents extracted information by means of interactive graphs. The visual interface is based on TouchGraph (cf. [TouchGraph]).

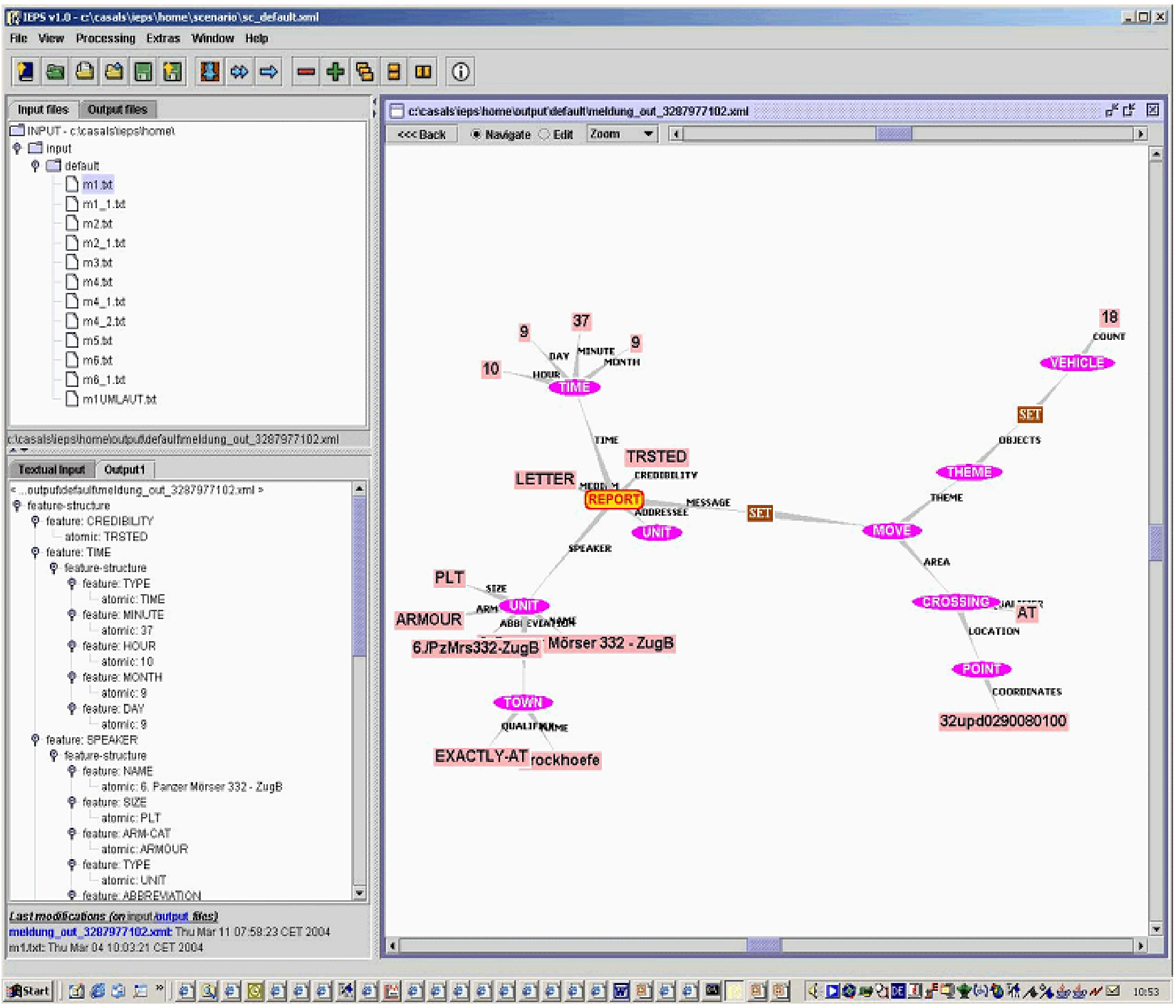

Figure 5: Information Extraction and Processing System (IEPS)

Entity-Action-Networks are a *conceptual* possibility to realize navigation in a meaning space. In the following three examples are shown. Which semantic basic units are contained, how they are combined and what additional information is shown is determined by the used complex filters.

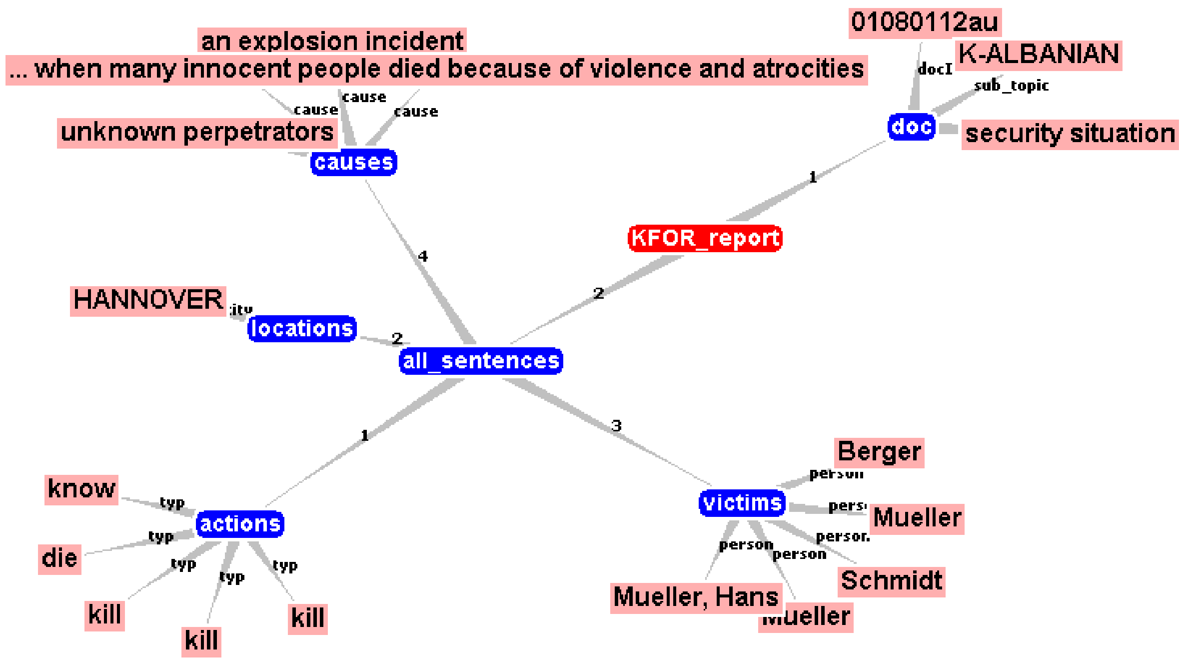

Figure 6: Meaning space of one document

A simple example is given in figure 6. The meaning space of *one* document is shown. The document has the document identification (docID) '01080112au', the main topic is 'security situation' and the sub-topic is 'K-ALBANIAN'. All the actions, victims, locations and causes from all sentences are shown. The used filter for this example doesn't relate actions with victims, locations and causes.

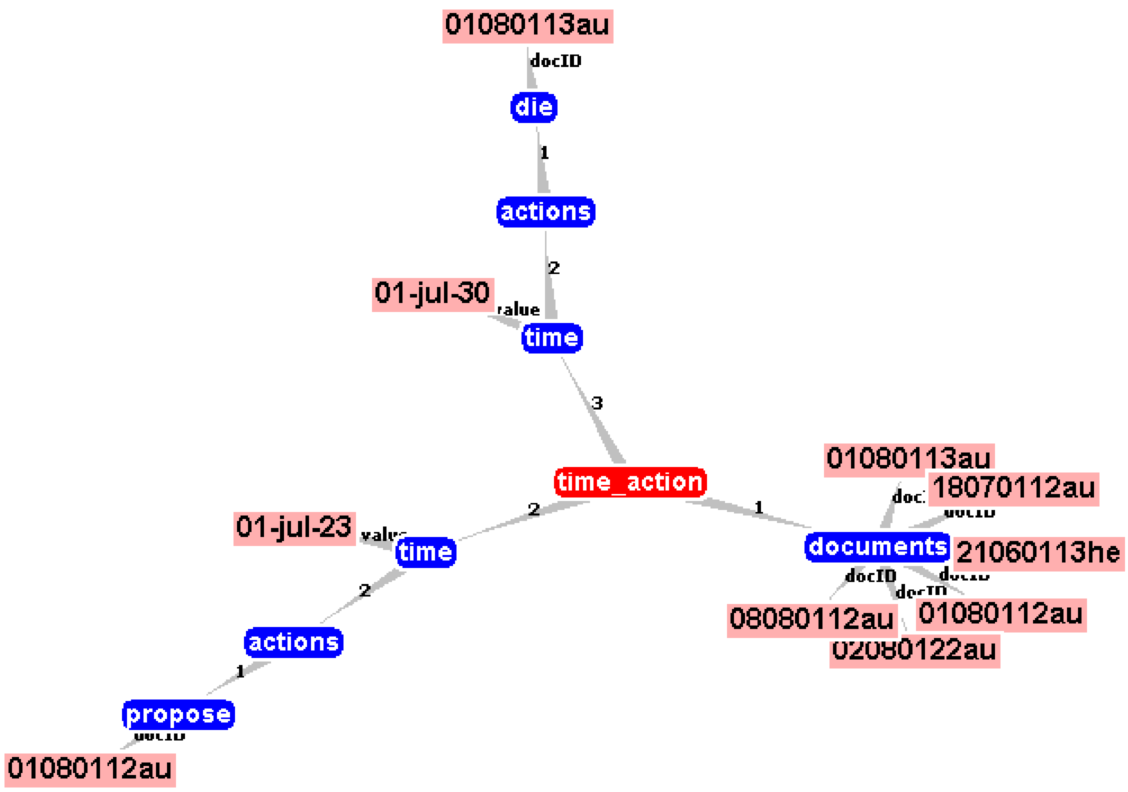

Figure 7: Time-action meaning space

In contrast to Figure 6, the example shown in Figure 7 relates actions and points in time. The filter searched various documents and found two points in time associated with an action. For each found time-action relationship the document ID is given.

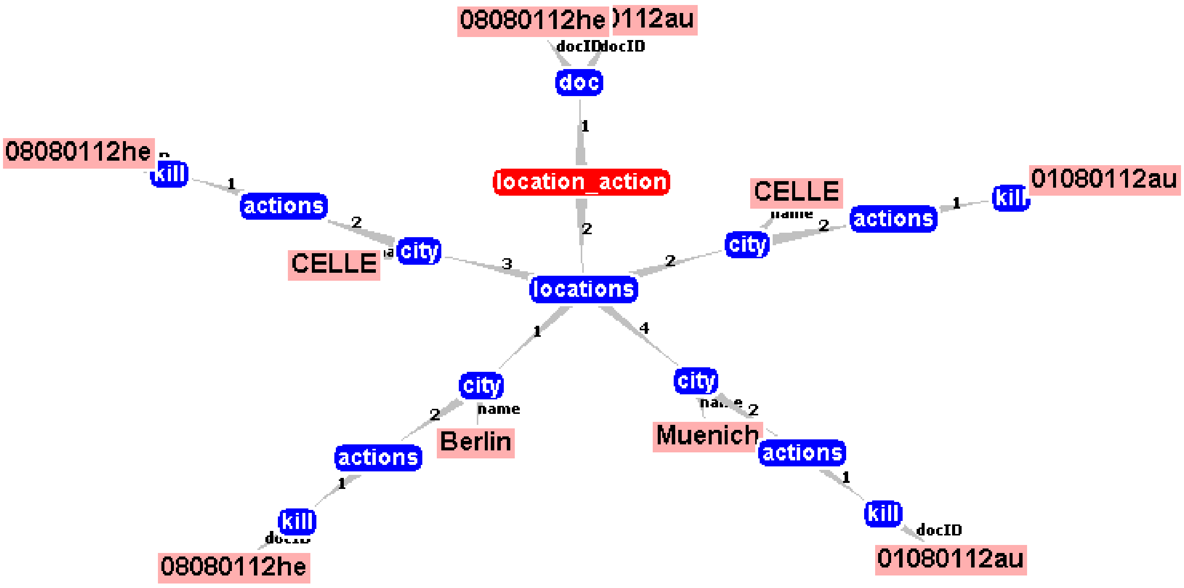

Figure 8: Location-action meaning space

In Figure 8 locations are related with actions. The city CELLE appears twice. The two branches of the graph are not collapsed due to the design of the filter.

4. Conclusion

In this paper, the ZENON project was presented. In this project an information extraction approach is used for the (partial) content analysis of English HUMINT reports from the KFOR deployment of the Bundeswehr. First, a short introduction into the information extraction approach was given. Then, the ZENON project was described. In the main part of the paper, the navigation through the meaning space of HUMINT reports was illustrated in detail. Various examples to realize the navigation were presented.

Work on the ZENON prototype goes on. The language processing capabilities will be extended and the approach to navigate through the meaning space will be evaluated. The main question here is, whether the IEPS approach is appropriated for presenting the results of analyzing a large quantity of documents. Other possible visualization techniques for the navigation through the meaning space were described above.

References

[Appelt, 1999] Appelt, D. & Israel, D. *Introduction to Information Extraction Technology*. Stockholm: IJCAI-99 Tutorial, 1999, http://www.ai.sri.com/~appelt/ie-tutorial/.

[Casals, 2004a] Casals Elvira, X. *Translation from Semantically Enriched Linguistic Representations to SQL Statements for the Project SOKRATES*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), FKIE-Bericht Nr. 76, Wachtberg, Germany, 2004

[Casals, 2004b] Casals Elvira, X. *Processing of Headers for the Information Extraction System of SOKRATES*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Technischer Bericht FKIE/ITF/2004/2, Wachtberg, Germany, 2004.

[Casals, 2005] Casals Elvira X., Hecking, M.. IEPS: *A Framework to Manage and to Visualize Information Extraction Results*. Forschungsgesellschaft für Angewandte Naturwissenschaften e.V. (FGAN), Technischer Bericht FKIE/ITF/2005/2, Wachtberg, Germany, 2005.

[Cunningham, 2002] Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.. *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[FrameNet] http://framenet.icsi.berkeley.edu/index.php (11.1.2006).

[Frey, 2004] Frey, M. L. and Schade, U. *Modular Framework for Military Report Processing*. FKIE-Bericht. Wachtberg: Forschungsgesellschaft für Angewandte Naturwissenschaft e.V., FKIE-Bericht Nr. 73, 2004.

[Hecking, 2001] Hecking, M. *Natural Language Access for C2 Systems*. Paper presented at the RTO IST Symposium on „Information Management Challenges in Achieving Coalition Interoperability“, held in Québec, Canada, 28-30 May 2001, and published in RTO MP-064.

[Hecking, 2002] Hecking, M. *Analysis of Spoken Input to C2 Systems*. In: Proceedings of the 7th International Command and Control Research and Technology Symposium (ICCRTS), Québec City, Canada, 2002.

[Hecking, 2003a] Hecking, M. *Information Extraction from Battlefield Reports*. In: Proceedings of the 8th International Command and Control Research and Technology Symposium (ICCRTS), Washington, DC, U.S.A., 2003.

[Hecking, 2003b] Hecking, M. *Analysis of Free-form Battlefield Reports with Shallow Parsing Techniques*. Paper presented at the RTO IST Symposium on „Military Data and Information Fusion“, held in Prague, Czech Republic, October 20-22, 2003.

[Hecking, 2004a] Hecking, M. *Informationsextraktion aus militärischen Freitextmeldungen*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), FKIE-Bericht Nr. 74, Wachtberg, Germany, 2004.

[Hecking, 2004b] Hecking, M. *How to Represent the Content of Free-form Battlefield Reports*. In: Proc. of the 2004 Command and Control Research and Technology Symposium (CCRTS) "The Power of Information Age Concepts and Technologies", June 15-17, 2004, San Diego, California.

[Hecking, 2004c] Hecking, M. *Improve Interoperability by Formalizing the Natural Language Parts of Military Messages*. In: Proc. of the Information Systems Technology Panel Symposium "Coalition C4ISR Architectures and Information Exchange Capabilities", September 27-29, 2004, The Hague, The Netherlands.

[Hecking, 2005a] Hecking, M. *Domänenspezifische Informationsextraktion am Beispiel militärischer Meldungen*. In: A.B. Cremers, R. Manthey, P. Martini, V. Steinhage (Hrsg.) "INFORMATIK 2005", Band 2, Lecture Notes in Informatics, Volume P-68, Bonn, 2005.

[Hecking, 2005b] Hecking, M. *KFOR-Korpus – Annotierungsvorschrift*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), unpublished, 2005.

[Hecking, 2006a] Hecking, M. *Content Analysis of HUMINT Reports*. In: Proc. of the 2006 Command and Control Research and Technology Symposium (CCRTS) "THE STATE OF THE ART AND THE STATE OF THE PRACTICE", June 20-22, 2006, San Diego, California.

[Inxight, 2006] Inxight Software, Inc., *VizServer - Online Demos*. http://www.inxight.com/products/vizserver (23.3.2006).

[Johnson, 1991] Johnson, B. and Shneiderman, B., *Treemaps: a space-filling approach to the visualisation of hierarchical information structures*, IEEE Proceedings Information Visualization, p. 284-291, 1991.

[Kartoo, 2006] KartOO S.A.. *KartOO - Visual Meta Search Engine*. http://www.kartoo.com (23.3.2006).

[McEnery, 2001] McEnery, T. and Wilson, A. *Corpus Linguistics*. Edinburgh University Press, Edinburgh, 2nd edition, 2001.

[MicroPatent, 2006] MicroPatent. *Aureka*. http://www.micropatent.com/static/advanced.htm (23.3.2006).

[NCW, 2001] Department of Defense. *Network Centric Warfare – Report to Congress*. 27 July 2001.

[Neumann, 2005] Neumann, A.. *Thematic Navigation in Space and Time*. In: Proc. of the 4th Annual Conference on Scalable Vector Graphics, The Netherlands 2005.

[OpenDX, 2006] OpenDX. *Open Data Explorer Visualization Software*, http://www.opendx.org (23.3.2006).

[Pazienza, 1999] Pazienza, M. T. (ed.). *Information Extraction*. Berlin, 1999.

[Schade, 2003a] Schade, U. *Ontologieentwicklung für Heeresanwendungen*. Forschungsgesellschaft für Angewandte Naturwissenschaften e.V. (FGAN), FKIE-Bericht Nr. 57, Wachtberg, Germany, 2003.

[Schade, 2003b] Schade, U. *Towards an Ontology for Army Battle C2 Systems*. In: Proceeding of the 8th ICCRTS, 2003.

[Schade, 2006] Schade, U. *Abschlussbericht SOKRATES - Ontologiegestützte Semantische Anreicherung für Interoperabilität und Lagedarstellung*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), FKIE-Bericht, Wachtberg, Germany, 2006.

[Steeneken, 1996] Steeneken, H. J. M. *Potentials of Speech and Language Technology Systems for Military Use: an Application and Technology Oriented Survey*. NATO, Technical Report, AC/243(Panel 3)TP/21, 1996.

[TouchGraph, 2006] http://www.touchgraph.com (11.1.2006).

[XSL, 2006] http://www.w3.org/Style/XSL (23.3.2006).