System ZENON – Semantic Analysis of Intelligence Reports

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 good candidates for applying techniques from computational linguistics. In this paper, the ZENON system is described, 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 objective of this research is to realize a navigatable Entity-Action-Network. The information about the actions and named entities are identified from each sentence. These representations can be combined and presented in a network. After a short introduction, the information extraction approach is explained. The ZENON system is described in detail. English HUMINT reports from the KFOR deployment form the basis for the development of the experimental ZENON system. These reports are used to build the KFOR text corpus, which is described as well.

Index Terms: information extraction, text mining, semantic analysis, intelligence, HUMINT reports, text corpus

1. Introduction

The *processing of human language* was identified as a critical capability in many future military applications [1]. Especially the *content analysis* of free-form texts is important for any information operation of the Network Centric Warfare (NCW) concept [2, p. 5-15]. The content analysis can be realized through *Information Extraction* (IE) which is a natural language processing technique [3], [4].

We set up the *research project ZENON*[1], in which the information extraction (IE) approach is used for the (partial) content analysis of English HUMINT reports from the KFOR deployment of the Bundeswehr [5], [6], [7], [8], [9], [10], [11]. The overall objective of this research is to create 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. These formal representations can be combined and presented in the navigatable network.

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 [3] 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* [6], 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. 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* [7]. 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.

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

3. ZENON system

Starting with English HUMINT reports (and a list of the city names) from the KFOR deployment of the Bundeswehr we have realized in the ZENON system that is able to do a (partial) content analysis of these reports [8]. 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 content of the reports.

3.1. Toolbox GATE

Since most of the reports are in English, GATE (General Architecture for Text Engineering, [12] 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 processing resources, which are used to realize the natural language processing parts of the ZENON system (e.g., morphological analyzer, part-of-speech (POS) tagger, pre-defined transducer to recognize English verbal phrases,

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

chunk-parsing). The functionality to select, combine and present the extracted information from different sentences and different reports is realized by XSLT (Extensible Stylesheet Language Transformation) filtering and the *Information Extraction Presentation System* (IEPS, [13]).

3.2. ZENON processing chain

In Figure 2 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 result 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 system. In the third sub-component the extracted content of different reports can be combined and selected according to predefined XSLT sheets. The result of the analysis is presented graphically and can be navigated interactively.

3.3. Named entities

An important processing step during the natural language processing is the recognition of the domain- and application-specific named entities. In the ZENON system 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*. An example is shown in Figure 1.

```
Rule: FVGPrePerPasNeg
//Recognizes:       Present Perfect Passive Negative:
//                  e.g. "hasn't been eaten"
//Pattern: (has | have) not been VBN
//Output: VG{adverb, infinitive, neg='yes',
//          tense='PrePer', type='FVG', voice='passive'}
(
        (
                {Token.string == "has"}|
                {Token.string == "have"}
        )
        (NEGATION)
        {Token.string == "been"}
        (ADVS):adverb
        ({Token.category == VBN}):verb
):x ==> {... Java code …}
```

Figure 1: *Verb phrase transducer 'FVGPrePerPasNeg'.*

3.4. Extraction of verb phrases, action types and sentence content

GATE offers various transducers to recognize the English verb groups. We have adapted and extended these transducers to fit our application. In addition to finite and non-finite verbal phrases also modal verb phrases, participles and special composed verb expressions are recognized.

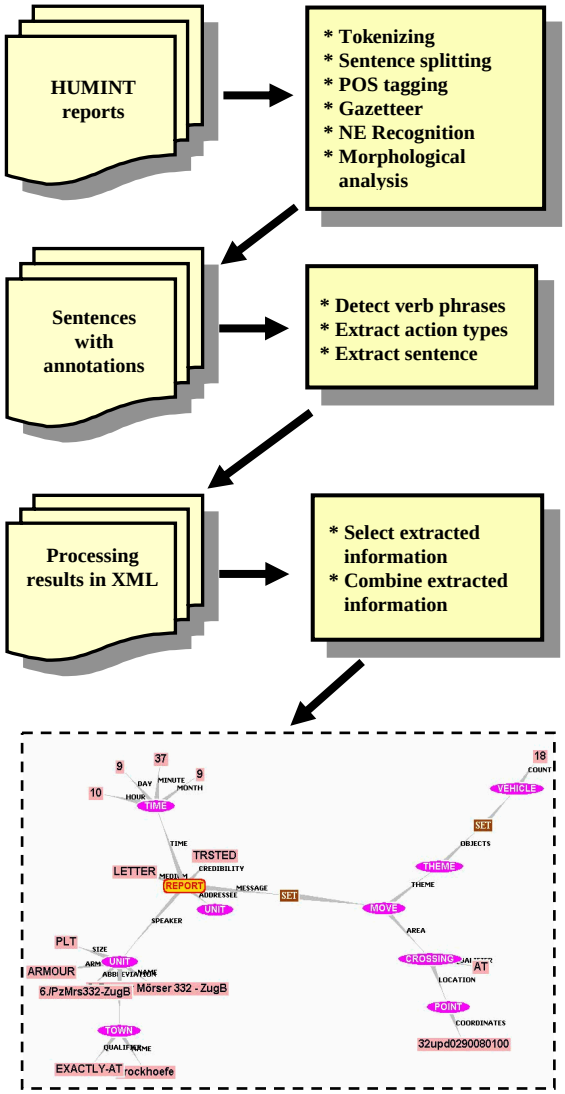

Figure 2: *ZENON processing chain.*

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 [14] 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. These semantic arguments are called the *frame elements* of the frame. Figure 3 shows an example. The core (must exist) frame elements for the frame 'killing' are CAUSE or KILLER and VICTIM. In the example

sentence 'John' fills the role KILLER and 'Martha' fills the role VICTIM.

Associated with each semantic frame are examples with typical syntactic realization of the frame elements. These examples and examples from the KFOR reports form the basis to construct the transducers, which produce the sentence content.

Semantic Frame 'killing':	A KILLER or CAUSE causes the death of the VICTIM.
Core frame elements:	CAUSE, KILLER, VICTIM
Non-core frame elements:	DEGREE, DEPICTIVE, INSTRUMENT, MANNER, MEANS, PLACE, PURPOSE, REASON, RESULT, TIME
Example sentence:	[John $_{KILLER}$] DROWNED [Martha $_{VICTIM}$]

Figure 3: *Semantic frame 'killing'.*

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.

3.5. Meaning space navigation

The natural language processing module of the ZENON system creates for the relevant sentences in each KFOR report a formal representation of the content. This contains information chunks about activities, events, entities, times and places. These basic units are selected and combined (e.g., all information about a specific person) through complex filters which can be defined for each scenario in the ZENON system. The filter functionality is realized through *Extensible Stylesheet Language Transformation* (XSLT, [15]). The result of the transformation is in XML format.

The intelligence analyst must be able to access and explore this *meaning space*. Therefore the meaning space is visualized as a *graphically navigatable Entity-Action-Network* by the sub-component IEPS [13]. This is a graphical software tool (see Figure 4) 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.

4. KFOR Corpus

4,498 military reports (mostly in English) from the KFOR deployment of the German Federal Armed Forces were used for the realization and optimization of the ZENON system. From these reports 800 were manually annotated and form the *KFOR Corpus.*

This corpus is a specialized micro-text corpus [16]. The corpus covers 886,000 tokens and contains the annotations in different *annotation layers* [11]. The following layers are available:

- *Original markups*: In this layer those parts of the message are annotated that are already formatted (e.g. addressee, topic, source).

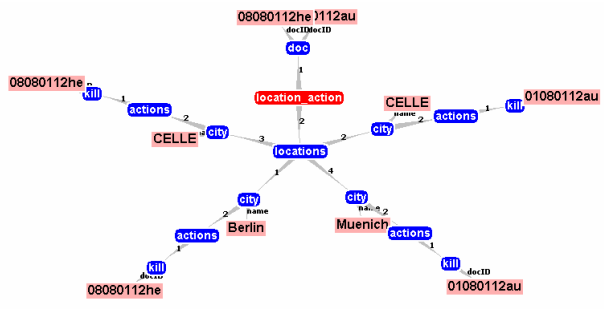

Figure 4: *Location-action meaning space.*

- *Token*: This layer contains the annotations about words, numbers, etc. The part-of-speech information and the lemma are also given.
- *Gazetteer*: In this layer those expressions are annotated that were identified over lists of names (e.g., first names, city names).
- *Sentence*: These annotations refer to sentences and begin and end markers of comments.
- *Named entities*: In this layer the above listed named entities are annotated.
- *Verb group*: The verbal phrases are annotated.
- *Thematic roles*: The syntactic and semantic function of expressions in sentences is annotated [17].

During the creation of the corpus a first version of the annotations was produced automatically. These annotations were then checked manually and corrected. GATE was used for both working-steps.

The corpus contains both syntactic and semantic annotations. The Figure 5 indicates, which annotation layers and annotation types are present, whether they are syntactic or semantic annotation types, and which of the annotation types were manually corrected.

Syntacticl/ semantical	Annotati. layer	Annotation type	Checked manually
syntactical	Original markup	DocID, DTGMeldung, Einsatz, Empfaenger, Hauptthema, Koordinate, Meldung, Meldungstyp, Ort, Quelle, Sachverhalt, Schlagworte, Titel, Unterthema	no
syntactical	Token	Token, SpaceToken	no
semantical	Gazetteer	Lookup	no
syntactical	Sentence	Sentence Comment Split	yes yes no
semantical	NE	City, Company, Coordinates, Colour, CountryAdj, Currency, Date, DocumentID, GeneralOrg, MilDateTime, MilitaryOrg, Number, Percent, Person, PoliticalOrg, Province, Region, River, Time, Title	yes

syntactical	VG	VG	yes
semantical	Thematic Role	ThRo	yes

Figure 5: *Annotation layers and annotation types*.

For each *annotation* the type, the layer, the start- and the end-position and a set of annotation-specific *features* are given. Each feature consists of a name and a value. A feature appears only, if a value is present. In the example

```
City NE xxx yyy {name=BERLIN}
```

the annotation is of type `City`. It belongs to the annotation layer `NE`. The string to which the annotation refers begins in position `xxx` and ends with position `yyy`. The annotation possesses a feature with the name `name` and the value `BERLIN`.

5. Development status

The 1st version of the ZENON system was realized. The system is able to process the action classes KILL, REPORT, KNOW, COMMAND, PROPOSE, EXPLODE and its associated semantic frames. The NE transducers were optimized with the help of the KFOR corpus. The verb group transducers were extended to recognize also modal verb phrases, participles and special composed verb expressions.

For the planned the 2nd version of the ZENON system a HUMINT ontology is under construction. In this new version the information extraction will also be *multilingual*. For this, processing resources to handle the language *Dari* were developed (cf. [18]).

6. Conclusions

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 system was described in detail. The GATE toolbox, the processing chain, and the extraction of named entities, verb phrases, action types and the sentence content was explained. The meaning space navigation was also mentioned. At the end, the KFOR corpus and the development status were presented.

7. References

[1] 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.

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

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

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

[5] 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.

[6] 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.

[7] 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.

[8] Hecking, M. *Domänenspezifische Informations-extraktion 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.

[9] 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.

[10] Hecking, M. *Navigation through the Meaning Space of HUMINT Reports*. In: "Proceedings of the 11th International Command and Control Research and Technolgy Symposium", September 26-28, 2006, Cambridge, UK.

[11] Hecking, M. *Das KFOR-Korpus als Ergebnis semantisch annotierter militärischer Meldungen*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Wachtberg, Germany, FKIE-Bericht Nr. 124, 2006.

[12] 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.

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

[14] *FrameNet*. http://framenet.icsi.berkeley.edu/index.php (24.10.2007).

[15] *XSL*. http://www.w3.org/TR/xslt (24.10.2006).

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

[17] Kremer, C. *Eine Untersuchung von Bewegungsverben im KFOR-Korpus im Vergleich zu FrameNet*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Wachtberg, Germany, FKIE-Bericht Nr. 117, 2006.

[18] Schwerdt, C. *Analyse ausgewählter Verbalgruppen der Sprache Dari zur multilingualen Erweiterung des ZENON-Systems*. Forschungsgesellschaft für Angewandte Naturwissenschaft e.V. (FGAN), Wachtberg, Germany, FKIE-Bericht, 2007.