12th ICCRTS
"Adapting C2 to the 21st Century"

June 19-21, 2007
Naval War College
Newport, RI
U.S.A.

Paper's Title: The KFOR Text Corpus

Topic: C2 Technologies and Systems; C2 Concepts, Theory, and Policy

Author
Name: Dr. Matthias Hecking
Organization: FGAN/FKIE
Address: Neuenahrer Straße 20
53343 Wachtberg-Werthhoven
Germany
Phone: +49 228 9435 576
Fax: +49 228 9435 685
E-Mail: hecking@fgan.de

The KFOR Text Corpus

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 HUMINT reports. The realized ZENON system uses an information extraction approach for the (partial) content analysis of English HUMINT reports from the KFOR deployment of the Bundeswehr. More then 4,000 military reports from this deployment were used as a starting point for the realization of the ZENON prototype. From these reports 800 were manually annotated and form the *KFOR Text Corpus*. This corpus is a specialized micro text corpus, which contains the syntactic and semantic annotations in different layers. In this paper, the KFOR Text Corpus and its use in the evaluation and the improvement of the ZENON system are presented. After a short introduction, an explanation is given, why corpora are needed for the evaluation of natural language processing systems. In the main part of the paper, the KFOR Text Corpus and its use for the evaluation of the ZENON system is described in detail. First, the different annotation layers and annotation types are presented. The corpus structure is also explained. Then, the use of the corpus to evaluate and improve the ZENON system is shown. Various examples are given.

Outline of the Paper

1. Introduction

- *Processing of human language* is identified as a critical capability in many future military applications (cf. [1])
- C*ontent extraction* from free-form texts is important for any information operation of the NCW concept (cf. [6], p. 5-15).
- *Information extraction* (IE) as a natural language processing technique is an engineering approach (cf. [4], [7])
- In the ZENON project a partial content extraction was realized for *KFOR HUMINT reports* (cf. [12], [11], [5], [2], [3], [7], [8])
- The *KFOR Text Corpus* was realized to *evaluate* and *improve* the information extraction components (cf. [10])

2. Corpora for Evaluation

- different types of corpora: audio, text, multi-media (cf. [9])
- different uses of corpora
- why we need the KFOR Text Corpus

3. The KFOR Text Corpus

- KFOR Corpus: specialized text micro-corpus with semantic annotations (cf. [9])

- for what purposes the corpus is used

3.1 Annotation Layers and Annotation Types

- definition and structure of the realized annotations
- syntactic and semantic annotations
- the realized annotation types: City, Company, Coordinates, Country, CountryAdj, Currency, Date, GeneralOrg, MilitaryOrg, Number, Percent, Person, PoliticalOrg, Province, Region, River, Time, Title, VerbGroup, ThematicRole
- Annotation rules and problems

3.2 The Corpus Structure

- the structure
- how the corpus can be used: various formats, format conversion software, TIGERSearch
- how the corpus was build: automatic pre-annotation, manual checking of the annotation

4. Evaluation of the ZENON System

- Measures: Precision, Recall, F-Measure
- How the KFOR Text Corpus is used to quantatively evaluate the ZENON system
- Results of the evaluation
- How to use this information to improve the ZENON system

5. Conclusion

- work done
- next steps

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] 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.

[3] 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.

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

[5] 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.

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

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.