Reinert, F., Waldschmitt, P., Leuchter, S. & Schönbein, R. (2007). Informationsextraktion durch Verwendung computerlinguistischer Verfahren in Texten mit Makrostruktur. In: R. Koschke, O. Herzog, K.-H. Rödiger & M. Ronthaler (Hrsg.), *INFORMATIK 2007. Informatik trifft Logistik. Beiträge der 37. Jahrestagung der Gesellschaft für Informatik e.V. (GI), 24.-27. September 2007 in Bremen* (Band 1, S. 190-194). Bonn: Gesellschaft für Informatik (LNI; P-109).
http://www.safety-critical.de/doc/informatik2007RecceOnto.pdf

Informationsextraktion durch Verwendung computerlinguistischer Verfahren in Texten mit Makrostruktur

Frank Reinert, Patrick Waldschmitt, Sandro Leuchter & Rainer Schönbein

Abt. Interoperabilität und Assistenzsysteme
Fraunhofer Institut Informations- und Datenverarbeitung (IITB)
Fraunhoferstraße 1
76131 Karlsruhe
vorname.nachname@iitb.fraunhofer.de

Abstract: In diesem Beitrag wird ein System vorgestellt, mit dem semi-strukturierte militärische Aufklärungsmeldungen analysiert werden können, um ontologiegestützt ein automatisches Szenenmodell aufzustellen. Im Rahmen der hier vorgestellten Studie wurde untersucht, in wie weit dazu das Open Source *Natural Language Processing*-Framework GATE eingesetzt werden kann. Ein GATE-basierter Prototyp wurde mit J2EE als Web-Service bereitgestellt und mit einer web-basierte Oberfläche zur Interaktion mit dem System und zur Visualisierung extrahierter Modelle versehen.

1 Einführung

Der Aufwand für die Entwicklung komplexer IT-Anwendungen liegt neben dem Entwurf und der Implementierung von Algorithmen oft zum größeren Teil in der Erschließung und Bereitstellung von Informationen in Form strukturierter Datensammlungen. Die automatische und semiautomatische Informationsgewinnung aus offen zugänglichen Quellen (OSINT, *open source intelligence*) ist deshalb ein wichtiges Hilfsmittel. Zurzeit werden Verfahren der Computerlinguistik, des Information Retrievals und des maschinellen Lernens verwendet, um elektronisch vorliegende Textdokumente zu verarbeiten. Je nach Art der vorliegenden Dokumente (Format, Inhalt, Sprache, Informationsdichte) und der benötigten extrahierten Information sind unterschiedliche Verfahren nützlich. Im Rahmen der militärischen Prozesse der Nachrichtengewinnung und Aufklärung müssen aus Textmeldungen führungsrelevante Informationen extrahiert, geprüft und zu einem aktuellen Lagebild verdichtet werden. Eine Abbildung der extrahierten Inforationen auf ein Domänenmodell (Ontologie) ermöglicht die anwendungsübergreifende Nutzung der gewonnenen Sachverhalte.

2 System

Im Rahmen der hier vorgestellten Studie wurde untersucht, in wie weit das Open Source *Natural Language Processing*-Framework GATE (*General Architecture for Text Engineering* [GATE]) für den Vorgang der Extraktion führungsrelevanter Informationen aus textbasierten Meldungen eingesetzt werden kann. Als Framework bietet GATE wieder verwendbare *Language-Engineering* Komponenten und vorgefertigte Softwarebausteine. Diese können im Rahmen der spezifischen Anforderungen für spezielle Anwendungen angepasst bzw. erweitert werden und zu einer „Pipeline“ aus einzelnen Textprozessierungsschritten integriert werden. Eine graphische Entwicklungsumgebung unterstützt diesen Prozess. Eine Java-API ermöglicht die Verwendung der Komponenten in eigenen Anwendungen. Bei der Untersuchung wurde schwerpunktmäßig die regelbasierte Verarbeitung von GATE betrachtet. Die Untersuchung stützt sich auf einen Satz (ca. 50) konkreter militärischer Aufklärungsmeldungen (Format nach [STANAG 3377]). Dieser Typ militärischer Meldung weist eine Semi-Strukturierung (Makrostruktur) mit strukturierten Inhalten und Freitextanteilen (festgelegt durch [STANAG 3596]) auf.

Für die Meldungen wurde analysiert, welche Prozessierungskomponenten in GATE geeignet sind. Daraus wurde exemplarisch eine spezifische Verarbeitungspipeline (s. Tabelle 1) abgeleitet und implementiert. Die entwickelte Pipeline liefert als Ergebnis für jede Meldung (ausgewählte) Entitäten und deren Beziehungen, die dann in einem nachgeschalteten Schritt in ein Netz aus Instanzen von Konzepten und Relationen einer Ontologie überführt werden. Der Schwerpunkt der Meldungsprozessierung liegt hierbei auf der Ableitung von Entitäten und Relationen zwischen diesen, z.B. Aktionen als Verbindung zwischen Subjekt und Objekt, räumlichen Anordnungsbeziehungen usw. Zu der Ableitung der Zusammenhänge werden zusätzlich die strukturellen Informationen über die Meldung genutzt. Die durch den Gesamtprozess gewonnene domänenspezifische formale Repräsentation kann dann in entsprechenden Anwendungen weiterverwendet werden, z.B. zusammen mit weiteren Informationen in einer automatischen Lagebildgenerierung.

Die Analyse der strukturierten und der Freitextanteile der Meldung dient ausschließlich dem Zweck, Instanzen für eine existierende Ontologie aus einem Satz von Meldungen abzuleiten. Die Instanzierung erfolgt nur im Kontext einer Meldung. Das Auflösen von Referenzierungen einzelner Entitäten (*Coreferencing*) über mehrere Meldungen hinweg wird nicht betrachtet. Für diese Art militärischer Meldungen ist ein Wortschatz zur Beschreibung der Anwendungsdomäne aufklärungsrelevanter Objekte vorgegeben. Dazu wurde auf eine im militärischen Aufklärungswesen eingeführte Kategoriedatenbank (konform zu [STANAG 3596]) mit objektspezifischen Angaben zu Status, Typ, Aktivität etc. mit ca. 20.000 definierten Vokabeln zurückgegriffen. Die Meldungen werden unabhängig von der Ontologie mit einem separaten operationellen Werkzeug erstellt, das auch diese Kategoriedatenbank nutzt. Die verwendete Ontologie (Ziel-Ontologie) besitzt eine geringe taxonomische Tiefe der Konzepthierarchien. Damit ist eine Flexibilität für die Abbildung auf dedizierte Domänen-/Anwendungsontologien gegeben. In Abbildung 1 ist ein Ausschnitt der Ontologie mit zwei Beispielinstanzen dargestellt. Die Ontologie

ist in RDFS formalisiert. Die Mächtigkeit von RDFS ist für diese Anwendung ausreichend. Es soll kein *Reasoning* betrieben werden.

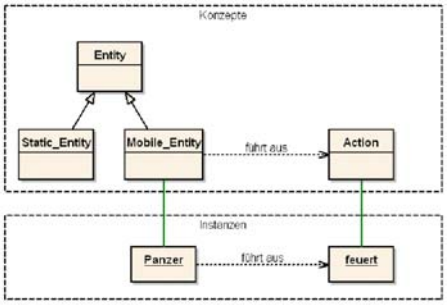

Abbildung 1: Ausschnitt der Ziel-Ontologie

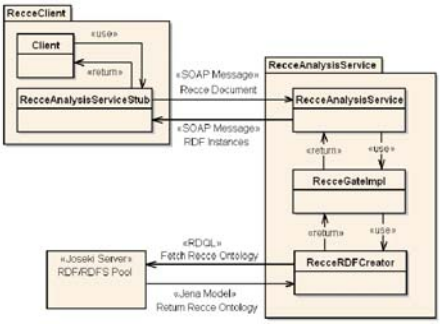

Abbildung 2: Client- und Server-Architektur des Prototypen

Im Rahmen des hier beschriebenen Systems wird die Ontologie für zwei Aufgaben genutzt: (1) als Ziel-Ontologie für die Generierung von Instanzen, (2) zur Konsistenzprüfung der durch die Textanalyse mit GATE erhaltenen Entitäten und ihrer Abbildung auf die Ontologie.

Der Prototyp ist als verteilte Anwendung konzipiert. Er beinhaltet drei Komponenten. Die eigentliche Informationsextraktion und Generierung der Instanzen ist als Web-Service mittels J2EE umgesetzt. Damit ist die Möglichkeit der Integration in bestehenden SOAs gegeben. Eine Benutzungsschnittstelle zur Interaktion mit der Pipeline und zur Ergebnisdarstellung (Darstellung des Instanzen-Netzes) wurde als Webanwendung entwickelt. In Abbildung 2 ist der prinzipielle Aufbau des Prototypen dargestellt. Die zentrale Auswertelogik ist so implementiert, dass die Pipeline der Textprozessierungsschritte offline mittels GATE entwickelt und getestet werden kann. Anschließend wird die Konfigurationsdatei in dem Web-Service *deployed*. Als Web-Service Framework wird Axis2 eingesetzt. Die Kommunikation zwischen dem Client und dem Server erfolgt mittels SOAP in einem LAN. Die dritte Komponente des Prototypen bildet ein Joseki-Server. Er dient als *Repository* für die Ontologie. Der Web-Service greift mittels RDQL-Anfragen auf Modell und Ontologie zu.

Zur Analyse einer Meldung wird diese mittels des Clients an den Server gesendet. Der Text bildet den Inhalt der SOAP-Nachricht. Der Web-Service extrahiert diesen Text und übergibt ihn GATE zur Textanalyse und zur nachfolgenden Erzeugung der Instanzen. Als Ergebnis wird eine SOAP-Nachricht an den Client zurückgesendet. Die erzeugten Instanzen können als Graph angezeigt oder mit weiteren Anwendungen verarbeitet werden. Für die Textanalyse mussten einige GATE Komponenten angepasst werden. Vor allem die Struktur der Meldungen bedingt Änderungen am Standard GATE *Sentence-Splitter*. Für spezielle Entitäten (z.B. für die Erkennung spezieller Koordinatenangaben) musste der Regelsatz des *JAPE-Transducers* anwendungsspezifisch erweitert werden. Auf das Vokabular der Meldungen angepasste *Gazetteerlisten* wurden aus der Kategoriedatenbank abgeleitet. Die Möglichkeit der

zweistufigen Annotation der von GATE verwendeten Gazetteerlisten wurde genutzt, um die Transformation von Entitäten in Ontologieinstanzen zu unterstützen. Dieser Ansatz ist speziell auf die Ziel-Ontologie zugeschnitten. Die Annotationen liefern der nachgeschalteten Anwendungslogik zur Generierung der Instanzen Hinweise darauf, welche Konzepte der Ontologie zu instanzieren sind. In Abbildung 3 sind einige typische Annotationen ausgewählter Regeln der textuellen Analyse einer Meldung abgebildet. Die aus dieser Analyse gewonnenen Metadaten werden zu der Generierung von Instanzen genutzt.

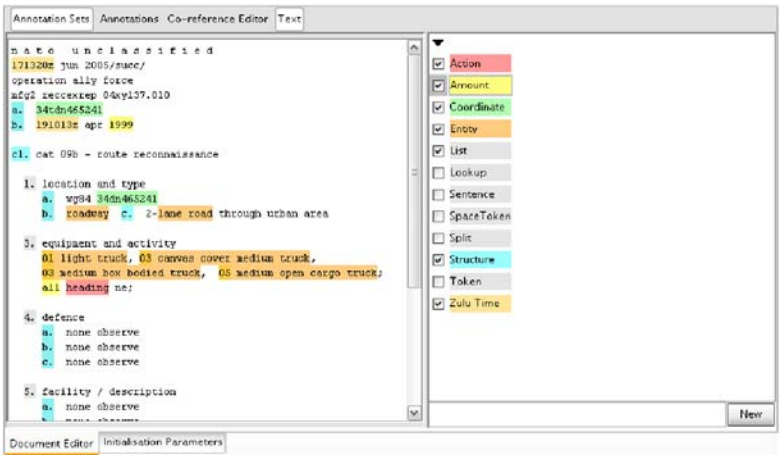

Abbildung 3: Annotationen in GATE

3 Andere Systeme

Hecking [Hec05, Hec06] beschreibt ein anderes System zur computerlinguistischen Analyse von militärischen Freitextmeldungen mit GATE mit dem Ziel, englischsprachige *HUMINT Reports* in graphisch navigierbare Entity-Action-Netzwerke zu überführen. Bei dem hier vorgestellten Systemprototypen geht es hingegen um die Analyse englischsprachiger *RECCEXREP Reports* [STANAG 3377] mit dem Ziel Instanzen einer Ontologie zu generieren und das resultierende semantische Netz graphisch darzustellen. Eine Gegenüberstellung der Arbeitsweise dieser beiden Systeme ist in Tabelle 1 dargestellt.

4 Ausblick

Das vorgestellte System ist als Prototyp realisiert und kann in SOAs eingebunden werden, um den beschriebenen Dienst zur Verfügung zu stellen. Damit ist ein wesentlicher Schritt zur teilautomatisierten Generierung einer interoperablen Lagedarstellung erreicht.

Gegenwärtig werden die Gazetteerlisten anhand des Inhalts der Kategoriendatenbank erweitert, um einen breiten Bereich des verwendeten Vokabulars abzudecken.

Die Überführung der Entitäten in die Instanzen der Ontologie erfolgt bei dem vorgestellten Systemprototypen noch in einer speziell auf die Zielontologie zugeschnittenen Implementierung. Dieser Prozess ist der Informationsextraktion nachgeschaltet. Ein generischerer Ansatz auf der Basis einer deklarativen Beschreibung der Abbildungsvorschrift mittels Regeln in Verbindung mit der Ontologie wird zukünftig entwickelt.

Tabelle 1: Vergleich [Hec06] mit dem dargestellten Prototypen

	[Hec06]	**Dargestellter Prototyp**
GATE-Pipeline	English Tokenizer Sentence Splitter PoS-Tagger Gazetteer NE-Transducer Morphological Analyser	English Tokenizer Sentence Splitter PoS-Tagger Morphological Analyser Flexible Gazetteer Gazetteer NE-Transducer
Verarbeitung	- Erfassen von Verb Phrases - Extraktion von Aktionstypen - Extraktion des Satzinhalts / der Semantik (durch Verwendung von Semantic Frames / FrameNet)	großteils lineare Verarbeitung Annotationen zur - Extraktion von Individuen und Relationen anhand Ontologie - Anreicherung während der Verarbeitung
Präsentation	- *Information Extraction and Processing System* [CH05] - Darstellung mit TouchGraph - Filterung der Darstellung durch Zusammenfassen von Ergebnissen unter Verwendung von XSLT	- Client für Webservice - Darstellung basiert auf Prefuse - Speichern, Öffnen und ungefilterte Darstellung von RDF-Dateien

Literaturverzeichnis

[CH05] Casal E. X., Hecking M.: IEPS: A Framework to Manage and to Visualize Information Extraction Results.

[GATE] http://gate.ac.uk/ (letzter Zugriff: 29.06.2007).

[Hec06] Hecking, M.: Content Analysis of HUMINT Reports. In: Proc. of the 2006 Command and Control Research and Technology Symposium, June 20-22, 2006, San Diego, California.

[Hec05] Hecking, M.: Domänenspezifische Informationsextraktion am Beispiel militärischer Meldungen. In: (Cremers, A.B.; Manthey, R.; Martini, P.; Steinhage, V., Hrsg.): INFORMATIK 2005, Band 2. GI (LNI; P-68), Bonn, 2005, S. 109-113.

[STANAG 3377] STANAG 3377 AR (EDITION 6) – Air Reconnaissance Intelligence Report Forms, 2002.

[STANAG 3596] STANAG 3596 AR (EDITION 5) – Air Reconnaissance Requesting and Target Reporting Guide, 2003.