Informationsextraktion

Christoph Wiewiorski
Patrick Hommers

Informationsextraktion(IE) - Einführung

- Ziel: Domänenspezifische Informationen aus freiem Text gezielt aufspüren und strukturieren
- Gleichzeitig sollen irrelevante Informationen „überlesen“ werden
- Es soll nicht das gesamte Dokument analysiert werden, sondern nur Textpassagen, die relevante Informationen beinhalten
- Relevantes wird durch domänenspezifische Lexikoneinträge oder Regeln dem System fest vorgegeben

Informationsextraktion(IE) – Einführung II

- Das bei der Extraktion benötigte Wissen muss so detailliert und genau wie möglich festlegen, welche Typen von Information extrahiert werden sollen
- Dies trifft vor allem auf traditionelle IE zu
- Moderne Ansätze, besonders bei domänenoffener IE, versuchen alle relevanten Informationen in einer Textkollektion zu erkennen und zu klassifizieren

Informationsextraktion(IE) – Einführung III

- Beide Ansätze wollen hierbei in freien Texten das Vorkommen (die Erwähnung) von Entitäten und ihre Beziehungen zueinander lokalisieren und in ein strukturiertes Format überführen
- Damit wird eine tiefere Verstehensleistung angestrebt, als volltextbasiertes Information Retrieval

Informationsextraktion(IE) – Einführung IV

Die wichtigsten konkreten Lösungsschritte umfassen:

1. Erkennung von spezifischen Eigennamen (Beispiel folgt)
2. Die Referenzierung von Eigennamen
3. Die Erkennung und Klassifikation von Relationen unterschiedlicher Komplexitäten zwischen den identifizierten Eigennamen

Letztlich ist die zentrale Aufgabe der IR die

Extraktion von semantischen Relationen.

Beispiel: Extraktion von Eigennamen

- Eigennamen sind in der Regel sprachliche Ausdrücke, die auf Individuen von Klassen oder Typen bestimmter Entitäten referieren
- Einfachste Art der Eigennamenerkennung (NER) wäre die vollständige Aufzählung aller Eigennamen eines bestimmten Typs in einer Liste (Gazetteer)
- Begehbarer erster und auch weit verbreiteter Weg, aber nicht ausreichend

Beispiel: Extraktion von Eigennamen II

- Für Eigennamen gibt es in der Regel viele Formulierungsmöglichkeiten
- Prozess der Generierung von Eigennamen ist sehr produktiv, z.B. bei Firmen- oder Produktnamen
- Durch die immer noch stattfindende Expansion von IE-Anwendungen und dem Aufkommen von neuen verwandten Themen hat eine drastische Erweiterung der NE-Kategorien stattgefunden

Beispiel: Extraktion von Eigennamen III

- Heutzutage dominiert die Entwicklung von daten-gesteuerten Verfahren zur NER
- Der Fokus liegt hierbei auf überwachte und semi-überwachte Strategien
- Dies geschieht in Form überwachter Lernverfahren
- Die Idee dahinter ist die automatische Berechnung von optimalen Merkmalskombinationen aus positiven und negativen Beispielen von NE‘s
- Diese liegen in einem großen Textkorpus in der Regel manuell annotiert vor

Beispiel

Dr. Hermann Wirth, bisheriger Leiter der Musikhochschule München, verabschiedete sich heute aus dem Amt. Der 65 jährige tritt seinen wohlverdienten Ruhestand an. Als seine Nachfolgerin wurde Sabine Klinger benannt. Ebenfalls neu besetzt wurde die Stelle des Musikdirektors. Annelie Häfner folgt Christian Meindl nach.

- Wie könnte das Template dazu aussehen?

- **Dr. Hermann Wirth**, bisheriger **Leiter** der **Musikhochschule München**, verabschiedete sich **heute** aus dem Amt. Der 65 jährige tritt seinen wohlverdienten Ruhestand an. Als seine Nachfolgerin wurde **Sabine Klinger** benannt. Ebenfalls neu besetzt wurde die Stelle des Musikdirektors. Annelie Häfner folgt Christian Meindl nach.

PersonOut	Dr. Hermann Wirth
PersonIn	Sabine Klinger
Position	Leiter
Organization	Musikhochschule München
TimeOut	heute
TimeIn	-

Wo liegt die Schwierigkeit bei diesem Text?

Projekt Infex BA

- Ziel: Verfahren der Informationsgewinnung für Geschäftsanwendung zu entwickeln und in Software-Komponenten nutzbar zu machen
- Schwerpunkt liegt auf:
  - Trend- und Stimmungsanalyse auf Basis öffentlicher Foren
  - Analyse von Märkten und Unternehmensberichten

Quelle: http://infexba.uni-paderborn.de/

Projekt Infex BA

- Bisheriger Stand:
  - Korpus für Entwicklung und Evaluierung von *Information Extraction*-Verfahren
- Beispiel:

  Rekordumsatz bei Wittenstein

  Produktion - 28.09.2006

  Der Mechatronik-Spezialist Wittenstein konnte seinen Umsatz 2005/2006 um 14,8 % auf den neuen Rekordwert von 133 Mio Euro steigern. […]

  http://www.produktion.de/news/detail/16620

Projekt Sokrates

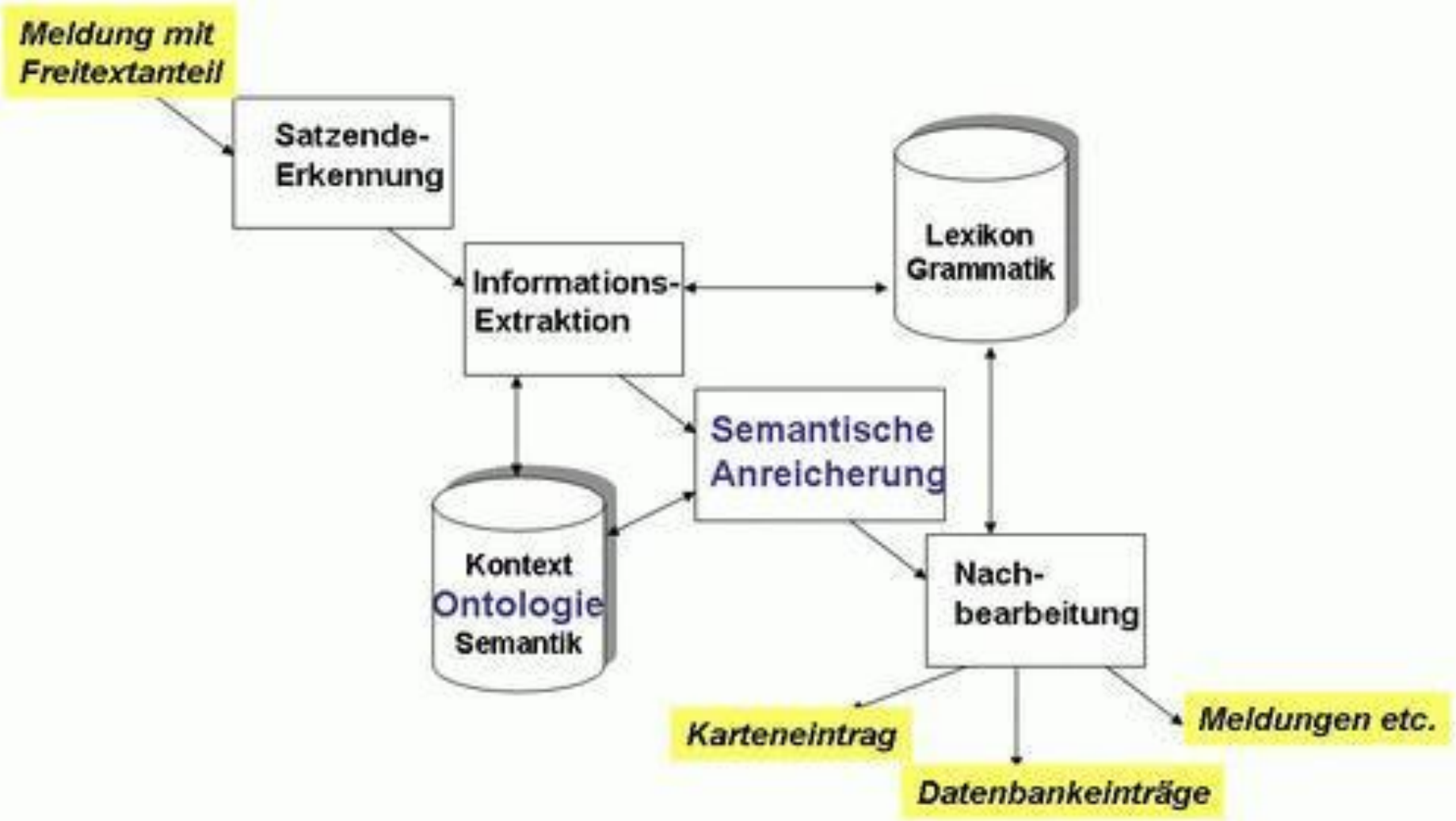
- System zur Analyse und Verarbeitung von sprachlichen und schriftlichen militärischen Meldungen

<u>Dreistufiger Prozess:</u>

1. eingehende Meldung durch Informationsexktraktion analysieren
   - > formale Repräsentation des Meldungsinhalts

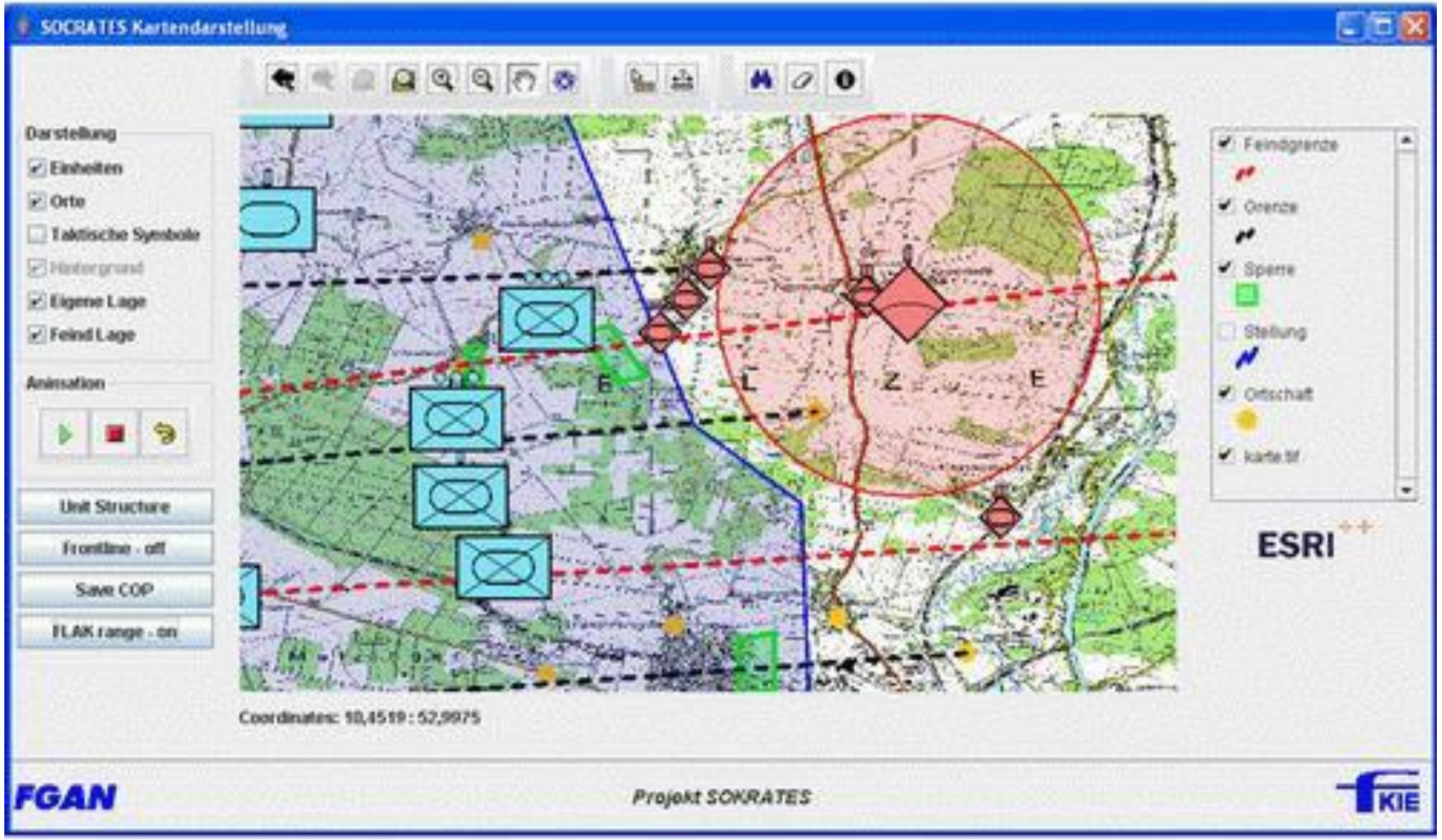
Quelle: http://www.fgan.de/fkie/fkie_c41_f13_de.html

2. Anreicherung der Informationen durch eine semantische Analyse
   - > allgemeines Wissen zu militärischen Operationen (Zeit, Ort, etc.) anwenden
3. Informationen weiterverarbeiten und z.B. in Karten einpflegen



Quelle: http://www.fgan.de/fkie/fkie_c41_f13_de.html

Umsetzung von Sokrates



Quelle: http://www.fgan.de/fkie/fkie_c41_f13_de.html

Quellen

- http://infexba.uni-paderborn.de/
- http://www.fgan.de/fkie/fkie_c41_f13_de.html
- http://www.dfki.de/~neumann/publications/new-ps/ie.pdf
- Computerlinguistik und Sprachtechnologie – Eine Einführung (3. Auflage 2010) Kapitel 5.3.3