Universität zu Köln
Computerlinguistik
SoSe 2020
Dr. Jürgen Hermes

Information Extraction

Ben Bigalke und Alexander Lürwer

Gliederung

- Definition / Erklärung
- Geschichte
- Methoden
- Bewertung
- Allgemein
- Regelbasierte Methoden
- Listenbasierte Methoden
- Natural Language Processing
- Allgemein Architektur
- Einfache Schritte
- part-of-speech (POS) tagging
- Named Entity Recognition (NER)
- (Noun-phrase) Chunking
- Klassifizieren

Definition / Erklärung

Information Extraction	≠	Information Retrieval
Extraktion innerhalb eines Dokumentes		Dokumentsuche innherhalb einer Datenbank

- Herausfiltern von Fakten aus unstrukturiertem Text in strukturierte Daten nach vorher festgelegten Kriterien/Spezifikationen
- identifiziert u. Klassifiziert „named entities“, Beziehungen zwischen diesen (relations) und Ereignisse
- Ausgabe bspw. In Datenbankform, zur weiteren Verarbeitung geeignet
- entweder spezielles Erkenenn der Information oder entfernen der nicht gesuchten Elemente

Geschichte

- entwickelt und geformt durch drei Serien von Evaluationskonferenzen:
  - MUC (Message Understanding Conference) 1988 – 1998
    - Ausfüllen eines aufgabenspezifischen Templates aus Flottenkommunikation, Artikeln

```
22.1    VISUAL SIGHTING OF PERISCOPE
        FOLLOWED BY ATTACK WITH ASROC AND TORPEDOS.
22.2    SUBMARINE WENT SINKER.
22.3    LOOSEFOOT 722/723 CONTINUE SEARCH.
22.4    FOUR BUOY ROAD PLACED BETWEEN CONSTELLATION AND DATUM.
```

```
MESSAGE ID

EVENT: HIGHEST LEVEL OF ACTION  DETECT, TRACK, TARGET,
                                HARASS, ATTACK, OTHER
FORCE INITIATING EVENT:         FRIENDLY, HOSTILE, NO DATA
CATEGORY(S) OF EVENT AGENT(S):  AIR, SURF, SUB, NO DATA
CATEGORY(S) OF EVENT OBJECT(S): AIR, SURF, SUB, LAND, NO DATA
ID(S) OF 0-TH LEVEL AGENT(S):
ID(S) OF 0-TH LEVEL OBJECT(S):
INSTRUMENT(S) OF 0-TH AGENT(S):
LOC OF OBJECT(S) AT EVENT TIME:
TIME(S) OF EVENT:
RESULT(S) OF EVENT:             1. RESPONSE BY OPPOSING FORCE
                                2. HOLDING CONTACT, LOST CONTACT
                                3. CONTINUING TO TRACK,
                                   STOPPED TRACKING
                                4. HOLDING TARGET, LOST TARGET
                                5. (NO) DAMAGE OR LOSS TO AGENT,
                                   (NO) DAMAGE OR LOSS TO OBJECT -
                                6. else, NO DATA
```

Geschichte

- ACE (Automatic Content Extraction) 1999 – 2008
  - mehr Beziehungen und Ereignisse bzgl. der named entities

    (supervised methods)
- TAC – KBP (Text Analysis Conference – Knowledge Base Population) 2009 – heute
  - größere Datenmenge zu untersuchen, einheitliche Datenbank ist das Ziel
  - Verknüpfung zwischen tausenden entities und Beziehungen
  - zu ausgewählten entities sollen Fragen beantwortet werden

    (semi-supervised methods)

Geschichte

- Organisiert durch US-Regierung (bspw. DARPA)
- deshalb anfänglich hauptsächlich US-Teilnehmer
- mögliche Regierungsverträge sind Anreiz
- Ablauf: Textkorpora werden zur Verfügung gestellt, Ergebnis-Template wird
- gestellt, Teilnehmer haben 1-6 Monate Zeit, am Ende Evaluation

Methoden
Bewertung der Ausgabe / Evaluation

$$\text{Precision} = \frac{\text{Richtig ausgefüllte Stellen in Antwort}}{\text{Anzahl von ausgefüllten Stellen in Antwort}}$$

- Wahrscheinlichkeit, mit der ein relevantes Ergebnis gefunden wird

$$\text{Recall} = \frac{\text{Richtig ausgefüllte Stellen in Antwort}}{\text{Anzahl von ausgefüllten Stellen im Schlüssel}}$$

- Wahrscheinlichkeit, mit der ein gefundenes Ergebnis relevant ist

$$\text{F-Maß} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

- Kombination von Genauigkeit und Trefferquote als Vergleichsmittel

Methoden
Allgemein

- 2 Arten: Regelbasiert und NLP (Natural Language Processing)
- Erste Ansätze waren regelbasiert
- Bei steigenden Anforderungen (s. Konferenzen) wurden neue, komplexere Verfahren entwickelt
- Allerdings werden für viele Methoden innerhalb der Architektur eine Mischung aus regelbasierten und NLP-Verfahren genutzt

Methoden
Regelbasierte Methoden

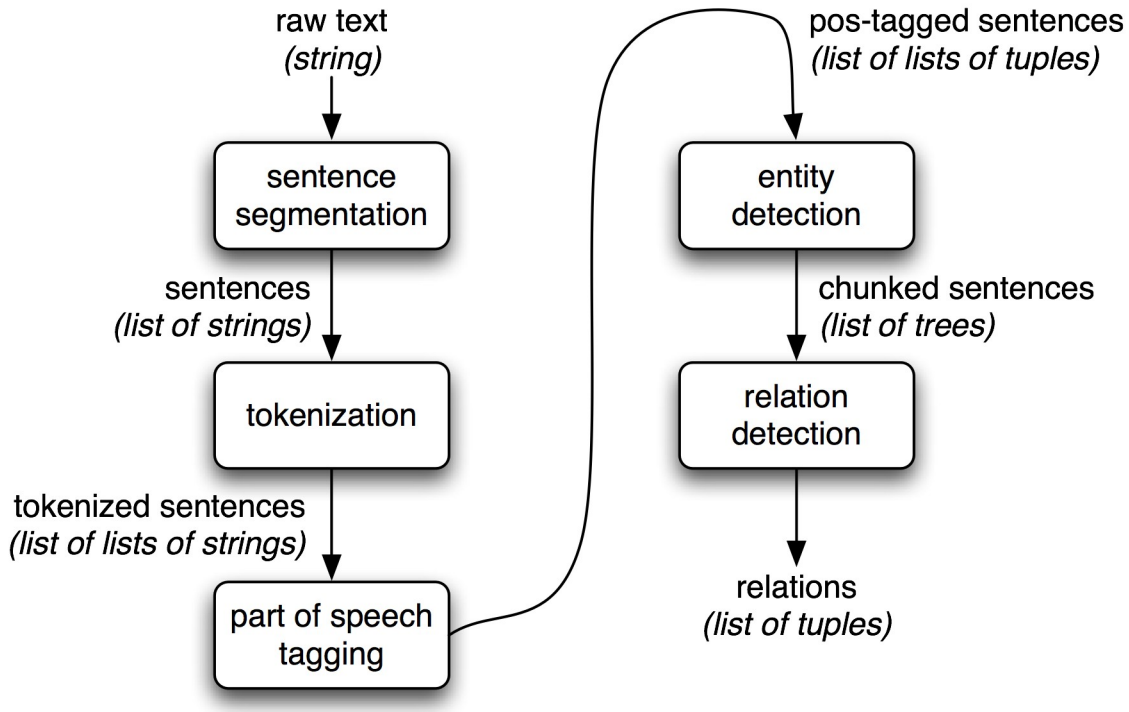
- Syntaktische Regeln beschreiben Stringeigenschaften
- bspw. Regular Expression (regex)
- Informationen, die der Regel entsprechen werden nach Vergleich extrahiert
- Beispiel „[0-9]+:[0-9]+“ für Uhrzeit im Format 12:41
- Vorteile:
  - Schnell
  - Einfach zu erstellen
  - Regeln können automatisch erzeugt werden (bspw. WHISK)
- Nachteile:
  - Striktheit der Regeln (Balance zwischen precision vs. Recall)
  - Keine Beziehungen zwischen entities

Methoden
Listenbasierte Methoden

- ein Wortliste von entities wird erstellt
- mittels Vergleich wird die entsprechende entity extrahiert
- Beispiel: Länderliste, Personenliste etc.

Natural Language Processing (NLP) Allgemeine Architektur

- Beispielhaft:



Methoden
Einfache Schritte

- Sentence Segmentation
  - Unterteilung in einzelne Sätze
- Tokenization
- (Stammformreduktion)
- (Assignment of semantic features)
  - Man is
    - [+human],
    - [+male],
  - Woman is
    - [+human],
    - [−male],

Methoden
part-of-speech (POS) tagging

- Supervised:
  - Hidden Markov Model
  - Regelbasiert
- Unsupervised:
  - Stochastisches tagging
  - Neural networks
- Für jede Sprache ein Tagset
  - Deutsch: Stuttgart-Tübingen-Tagset (STTS)
- Beispiel:
  - „Petra/NE liest/VVFIN einen/ART langen/ADJA Roman/NN“

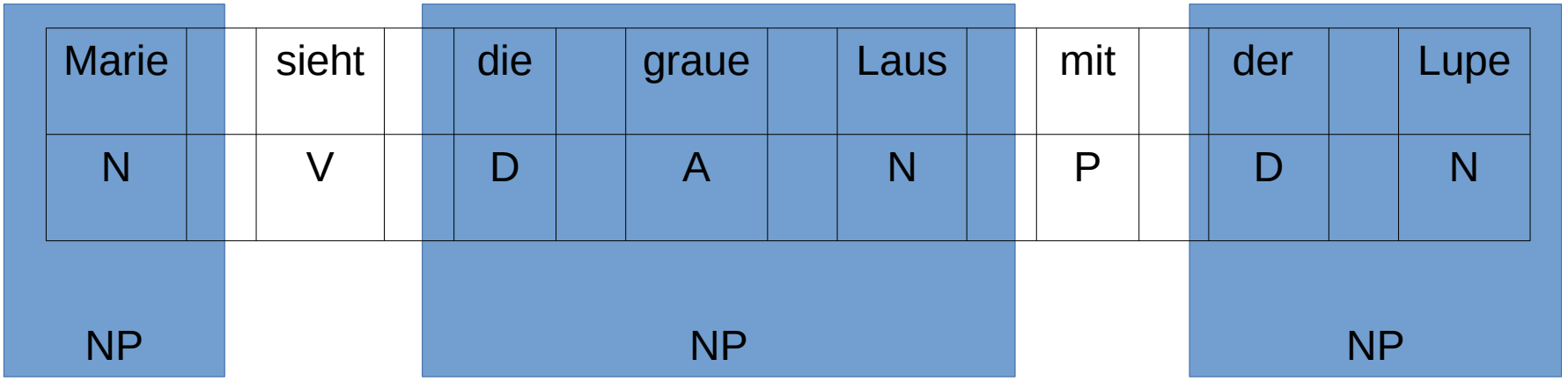
Methoden
Named Entity Recognition (NER)

- Unformatierter Text wird mit Eigennamen Tags versehen, basierend auf vorher festgelegte Kategorien (z.b. Personen, Orte, Datum, Medizin, Chemikalien etc.)
- Problem: Doppelbedeutung : Person Ford <--> Firma Ford
- Setzt POS-tagging vorraus
- Typische Vorgehensweise:
  - Chunking → Klassifizieren

Methoden
(Noun-phrase) Chunking

- Auch partial oder shallow parsing genannt
- Meist regelbasiert

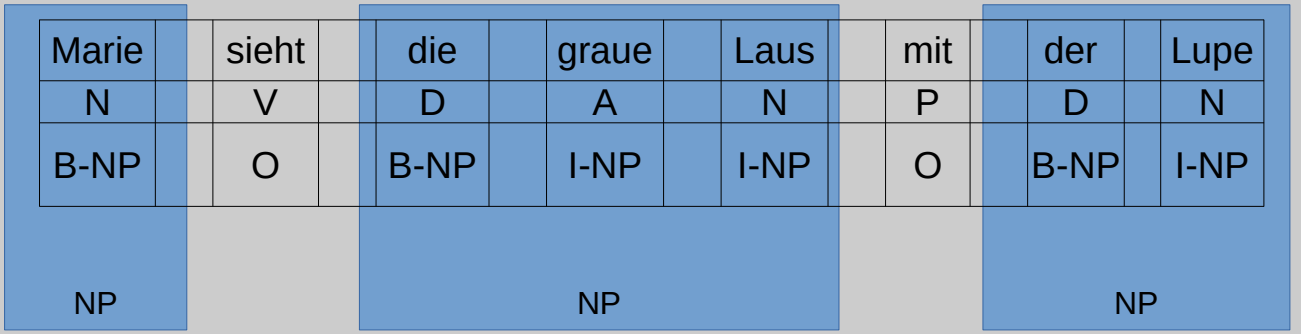
- Noun-phrase Chunking:
- Man bestimmt ein Tag Pattern
- Regelbasierter ChunkParser findet Nomen-Phrasen

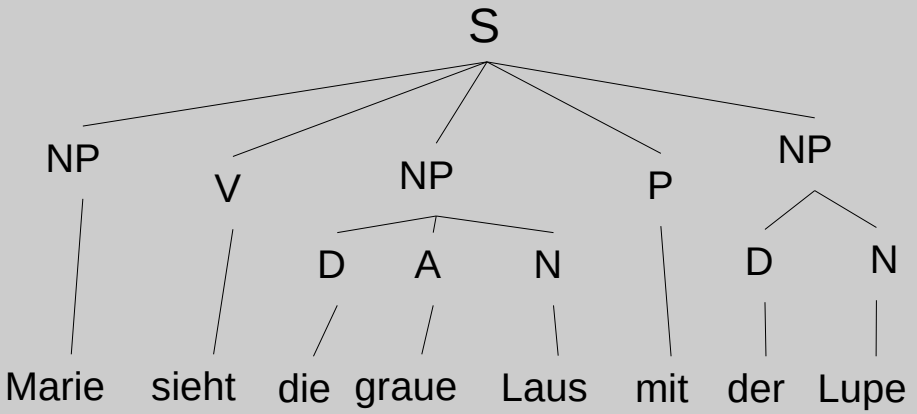


Methoden
(Noun-phrase) Chunking + IOB

- Typisches Aussehen der Chunks mit Inside-outside-Beginning

Wort	POS	IOB
Marie	NN	B-NP
sieht	V	O
die	DT	B-NP
graue	A	I-NP
Laus	NN	I-NP
mit	P	O
der	DT	B-NP
Lupe	NN	I-NP
.	.	O





Methoden
Klassifizieren

- vorweg: Auch hier handelt es sich in der praktischen Anwendung häufig um eine Mischung aus all diesen Herangehensweisen.
- Regelbasiert:
  - Sichere Textstrukturen
  - Beispiele:
    - Auf „Dr. „ folgt eine Person
    - Absender eines Briefes, etc.
    - Artikel: Name des Autors
    - Konkret: „[DD]-[MM]-[YYYY]“ für Datum im Format 23-06-2020

Methoden Klassifizieren

- stochastisch/statistisch:
  - Mathematische Grundlagen, die anhand von Wahrscheinlichkeiten bestimmen ob eine Phrase einer gegebenen Klasse entspricht
  - machine learning wird genutzt, um diese Algorithmen anhand von Trainingskorpora zu trainieren (supervised vs. unsupervised)
  - Genutzte Modelle:
    - Robust Risk Minimization Classifier
    - Maximum Entropy Markov Model Classifier
    - Transformation-Based Learning Classifier
    - Hidden Markov Model Classifier

Methoden Klassifizieren

- gazetten/lexikal/listenbasiert:
  - Datenbanken, die auf vorproduzierten Listen basieren
  - Beispiele:
    - Ortslisten
    - Telefonbuch
    - Firmenlisten
    - Personenlisten
    - Von Hand erstellt
  - Oft wird auch eine online-Datenbank benutzt, bspw. Wikipedia
  - Werden häufig genutzt, um o.g. statistische/stochastische Algorhythmen anhand von Trainingkorpora anzulernen

Methoden
Relation Detection / Extraction

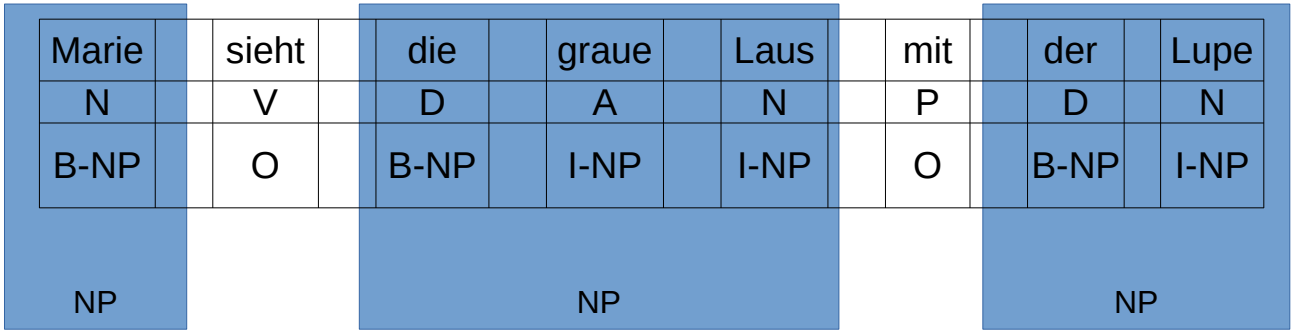
- Weiterverarbeitung durch spezifische Beziehungen zwischen Nes
- Kann entweder regelbasiert oder durch machine learning passieren
- Regelbasierte Systeme versuchen durch pattern matching mit dem text zwischen 2 named entities Beziehungen herzustellen
- Machine learning lernt diese pattern anhand eines Trainings-Korpus automatisch
- Beispiel:
  - Search Person Tier pattern “sieht”
  - Result : “Marie”,”Laus”,”sieht”

Methoden
Zusammenfassendes Beispiel

- Sentence Segmentation:
  - At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Marie sieht die graue Laus mit der Lupe. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua.
- Tokenization:
  - |Marie| |sieht| |die| |graue| |Laus| |mit| |der| |Lupe||.|
- Part-of-speech tagging:
  - Marie/N sieht/V die/D graue/A Laus/N mit/P der/D Lupe/N.

Methoden
Zusammenfassendes Beispiel

- Named Entity Recognition
  - Noun-phrase Chunking u. IOB-Tags



- Klassifizierung:
  - Marie (Person) sieht die graue Laus (Tier) mit der Lupe (Gegenstand)

Methoden
Zusammenfassendes Beispiel

- Relation Extraction / Detection
  - Daten: Marie sieht die graue Laus mit der Lupe.
  - Suche: Search Person Tier pattern “sieht”
  - Ergebnis: “Marie”,”Laus”,”sieht”

Literatur

- Grishman, Ralph (2019),*Twenty-five years of information extraction*,

  https://www.cambridge.org/core/services/aop-cambridge-core/content/view/0E5BB0D6AE906BB3C25037E2D74CA8F3/S1351324919000512a.pdf/twentyfive_years_of_information_extraction.pdf
- Sarawagi, Sunita (2008), *Information Extraction*,

  https://www.cis.uni-muenchen.de/~fraser/information_extraction_2018_lecture/sarawagi.pdf
- Bird, Steven; Klein, Ewan; Loper, Edward (2018), *Natural Language Processing with Python : 7. Extracting Information from Text*
  https://www.nltk.org/book/ch07.html
- 18.04.2013, *A Survey of Stochastic and Gazetteer Based Approaches for Named Entity Recognition - Part 2*

  http://www.datacommunitydc.org/blog/2013/04/a-survey-of-stochastic-and-gazetteer-based-approaches-for-named-entity-recognition-part-2/