Named Entity Recognition (NER)

Katharina Stein

Inhalt

- Named Entity Recognition
  - Was ist Named Entity Recognition?
  - Bedeutung für Natural Language Processing
  - Herausforderungen
  - Entity Resolution
- Tool für Named Entity Recognition: Stanford Named Entity Recognizer
  - Funktionsweise
  - Verwendung
- Übung

Was ist Named Entity Recognition?

- Teilaufgabe der Informationsextraktion aus Texten

- NER ist auch bekannt unter den Namen:
  - entity identification
  - entity extraction

- Named Entity:
  - Alles, worauf sich ein Eigenname bezieht
  - Im weiteren Sinne z.B. auch Zeiten, numerische Daten

The wedding of Prince William, Duke of Cambridge, and Catherine Middleton took place on 29 April 2011 at Westminster Abbey in London, United Kingdom.

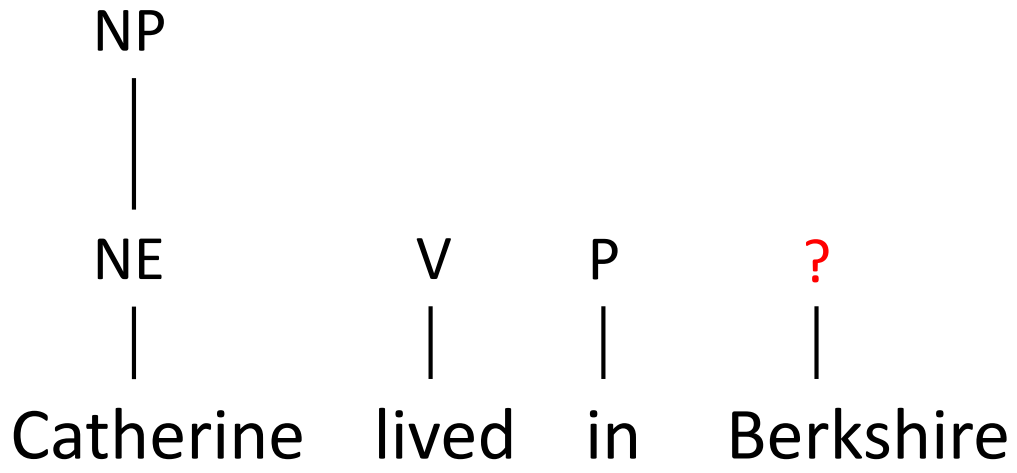
The groom, Prince William, Duke of Cambridge, is the eldest son of Charles, Prince of Wales and Diana, Princess of Wales, and second, after his father, in line to succeed his paternal grandmother, Queen Elizabeth II.

The wedding of Prince William, Duke of Cambridge, and Catherine Middleton took place on 29 April 2011 at Westminster Abbey in London, United Kingdom.

The groom, Prince William, Duke of Cambridge, is the eldest son of Charles, Prince of Wales and Diana, Princess of Wales, and second, after his father, in line to succeed his paternal grandmother, Queen Elizabeth II.

Parsing

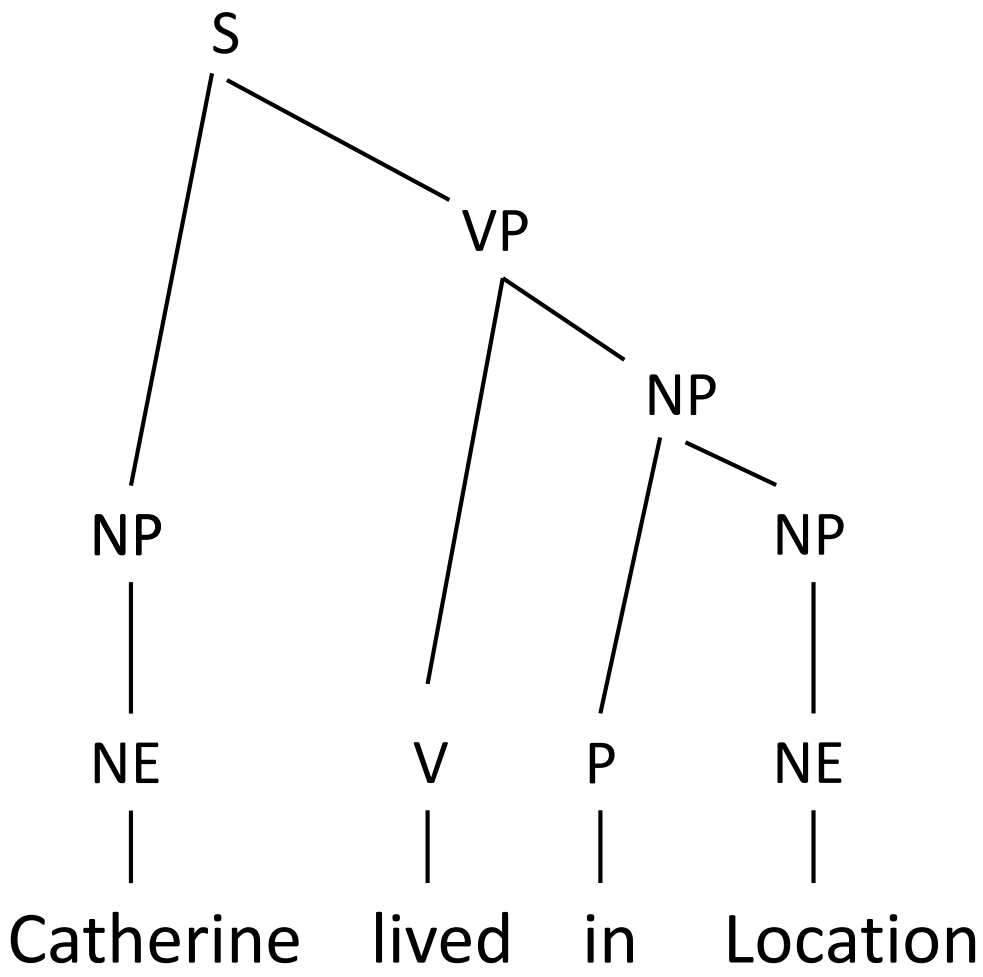
- Catherine lived in Berkshire.
- Grammatik:
  - S → NP VP
  - NP → NE
  - VP →V PP
  - PP → P NP
- Lexikon oder Tagging:
  - Catherine: NE
  - lived: V
  - in: P



Parsing

- NER:

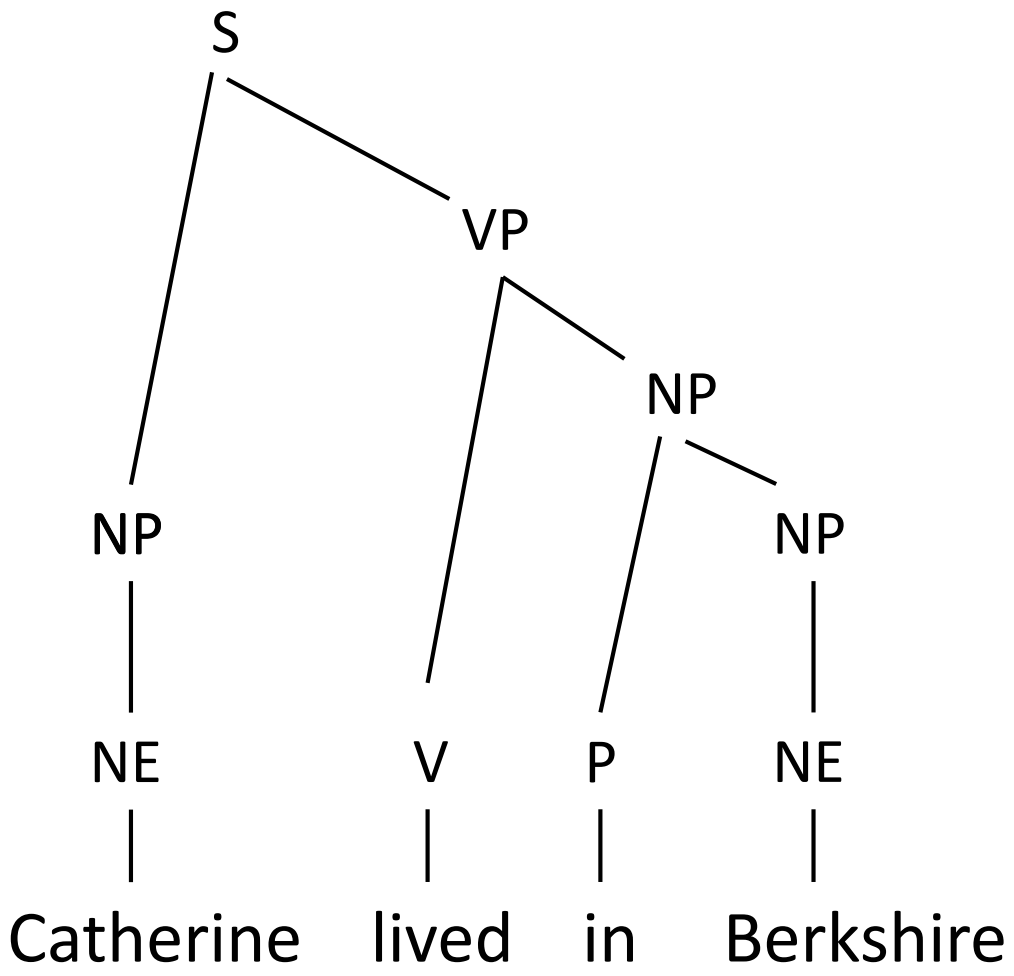
  Catherine lived in Berkshire.
- Catherine lived in Location.
- Grammatik:
  - S → NP VP
  - NP → NE
  - VP →V PP
  - PP → P NP
- Lexikon oder Tagging:
  - NE → Catherine
  - V → lived
  - P → in
  - NE → Location



Parsing

- NER:

  Catherine lived in Berkshire.
- Catherine lived in Location.
- Grammatik:
  - S → NP VP
  - NP → NE
  - VP →V PP
  - PP → P NP
- Lexikon oder Tagging:
  - NE → Catherine
  - V → lived
  - P → in
  - NE → Location



Bedeutung für Natural Language Processing

- NER meist einer der ersten Schritte für Informationsextraktion

- Wichtig für weitere Verarbeitung des Textes
  - Maschinelle Übersetzung
  - POS-Tagging
  - Semantic Role Labeling
  - Semantic Relations

Herausforderung Ambiguitäten (1/4)

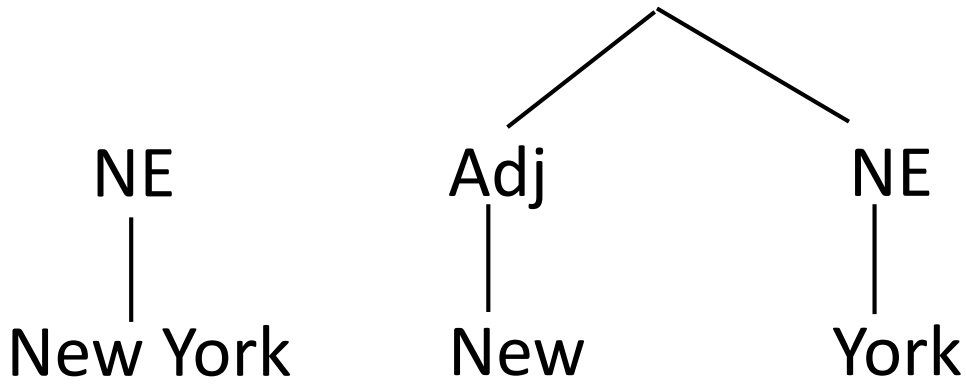
Bedeutungstragende Namen:

- Stein hält heute ihren Vortrag.
- Brown meets his friends.

Herausforderung Ambiguitäten (2/4)

Bedeutungstragende Namen:

- New York



- Los Angeles International Airport

Herausforderung Ambiguitäten (3/4)

- Er besuchte JFK: Person

- Er flog ab JFK: Flughafen bei New York

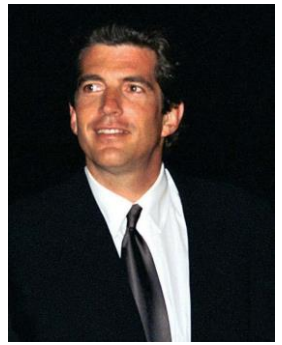
Herausforderung Ambiguitäten (4/4)

- Er besuchte JFK: Person

- Er flog ab JFK: Flughafen bei New York

- Er besuchte JFK: Person

- Er besuchte JFK: Person



Entity Resolution (ER)

- NER nicht immer ausreichend
- zusätzlich zu Klasse auch genaue Entität benötigt
- ER auch bezeichnet als Named Entity Linking
- Weist der Entität einen Link zu einem Datenbankeintrag zu

- Viele Touristen besuchen Westminster Abbey in London.
  - NER: London → LOCATION
  - ER: London → https://de.wikipedia.org/wiki/London
    London → https://de.wikipedia.org/wiki/London_(Ontario)

Tool für NER:
Stanford Named Entity Recognizer

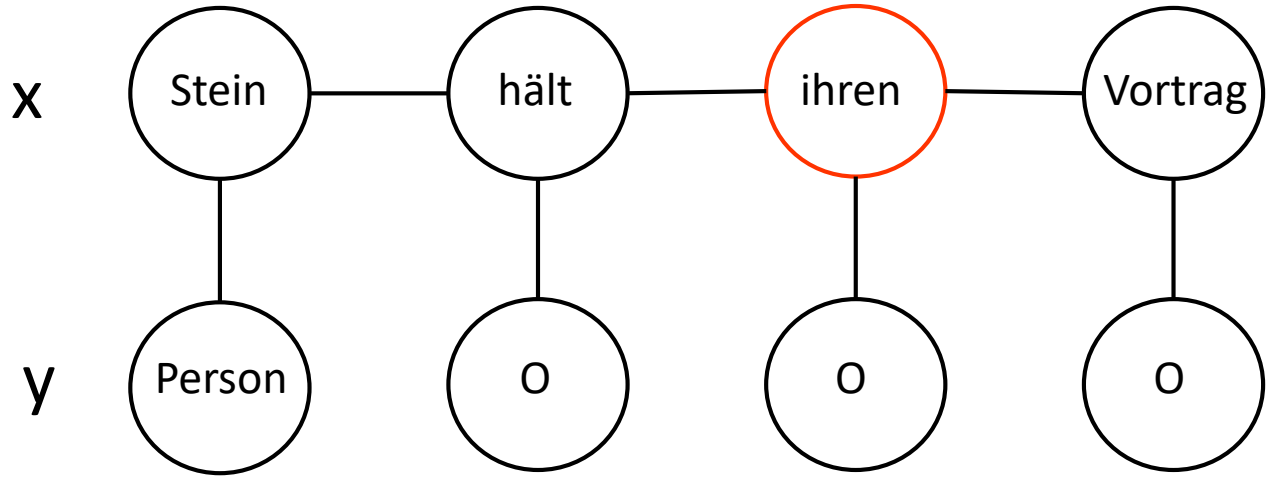
Stanford NER

- Stanford NLP Tools: Verschiedene Natural Language Processing Software zum Lösen computerlinguistischer Probleme
- Java Implementierung
- Training auf annotierten Daten, überwachtes Verfahren:

John	PERSON
Kerry	PERSON
will	O
fly	O
to	O
Paris	LOCATION
this	O
weekend	O
.	O

Conditional Random Field Classifiers

- Sequence-to-sequence problem
- argmax P(y|x)
  y



Features

- Wörter: aktuelles, vorheriges und folgendes Wort, Wörter in festgelegtem Fenster
- Rechtschreibung:
  - Groß- und Kleinschreibung: Hans → Xxxx, eBay → xXxx
  - Buchstaben und Zahlen: MG3 → XX#
- Präfixe und Suffixe
- Distributionelle Ähnlichkeit

Featurefunktionen

- Featurefunktionen definieren
- Gewichte aus Trainingsdaten lernen
- Vorteil CRF:
  - Viele Kombinationsmöglichkeiten
  - Keine Unabhängigkeitsannahme

Gazette features

- Stanford NER verwendet keine
- Können beim Training den annotierten Daten hinzugefügt werden

- Gazetteer: eigentlich ein Ortslexikon, das alle Orte oder Regionen einer Gemeinde auflistet

- Gazette file:

  Location Berkshire
  Location Algermissen

Klassifizierer

- 3 Klassen:
  - PERSON, ORGANIZATION, LOCATION
  - Trainingsdaten: CoNLL, MUC6, MUC7, ACE, OntoNotes, Wikipedia
  - Robust bezüglich amerikanischem und britischem Englisch
- 4 Klassen:
  - PERSON, ORGANIZATION, LOCATION, MISC
  - Trainingsdaten: CoNLL 2003 Shared Task training data
- 7 Klassen:
  - PERSON, ORGANIZATION, LOCATION, DATE, MONEY, PERCENT, TIME
  - Trainingsdaten: MUC

NER für Englisch

- Zuerst in den Ordner wechseln, in dem sich der Stanford NER befindet

Starten unter Windows

- java -mx600m -cp "*;lib/*" edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier [Klassifizierer] -textFile [inputfile]

Starten unter Linux / MacOSX

- java -mx600m -cp "*:lib/*" edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier [Klassifizierer] -textFile [inputfile]

NER für Deutsch

- java -cp "*" edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier edu/stanford/nlp/models/ner/german.conll.hgc_175m_600.crf.ser.gz -tokenizerOptions latexQuotes=false -textFile [inputfile]

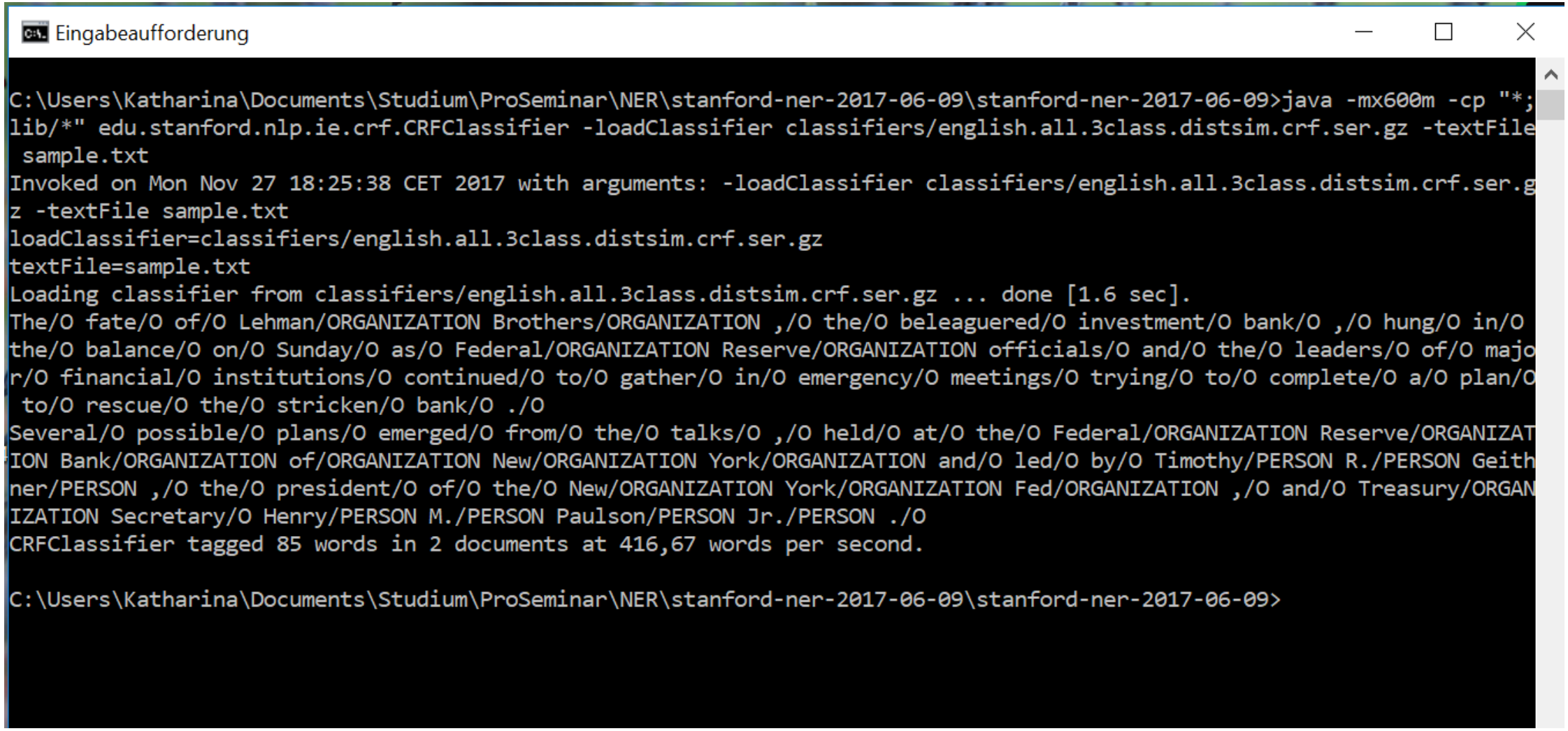
Argumente

- Klassifizierer (nur für Englisch):
  - classifiers/english.all.3class.distsim.crf.ser.gz
  - classifiers/english.conll.4class.distsim.crf.ser.gz
  - classifiers/english.muc.7class.distsim.crf.ser.gz

- Ausgabedatei:
  - java -mx600m -cp "*;lib/*" edu.stanford.nlp.ie.crf.CRFClassifier -loadClassifier [Klassifizierer] –outputFormat [option] -textFile [inputfile] > [outputfile]

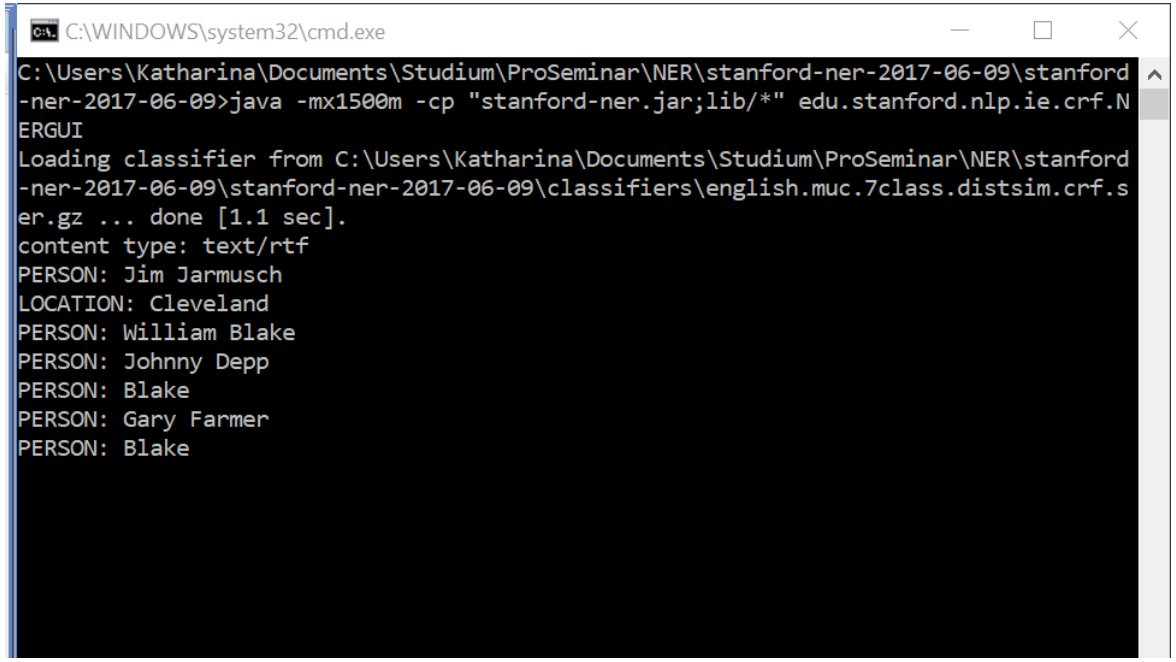
  - OutputFormat: slashTags (default), tsv, tabbedEntities, xml, inlineXML

Ausgabe für sample.txt



GUI (für Englisch)

- Windows: ner-gui.bat
- Linux / MacOSX: ner-gui.sh



Übung

Übung

- Stanford NER über Kommandozeile starten
  - Englisch: sample.txt, sample-w-time.txt
  - Eigene Textdatei
  - Ausgabe auf der Kommandozeile und in Datei
  - Klassifizierer wechseln
  - OutputFormat wechseln

Übung

- Verschiedene Sätze und Texte ausprobieren
  - Wo sind die Problemstellen?
  - Was wird nicht erkannt oder falsch klassifiziert?
  - Was klappt gut?
  - Sätze mit Namen mit Bedeutung: Wann wird Name als Named Entity erkannt, wann nicht? Welche Rolle spielt der Kontext?
  - Ambige oder ungewöhnliche Eigennamen
  - Den gleichen Satz in unterschiedlichen Sprachen ausprobieren
- http://corenlp.run/ oder GUI

Vielen Dank für eure Aufmerksamkeit