

# Maschinelle Sprachverarbeitung

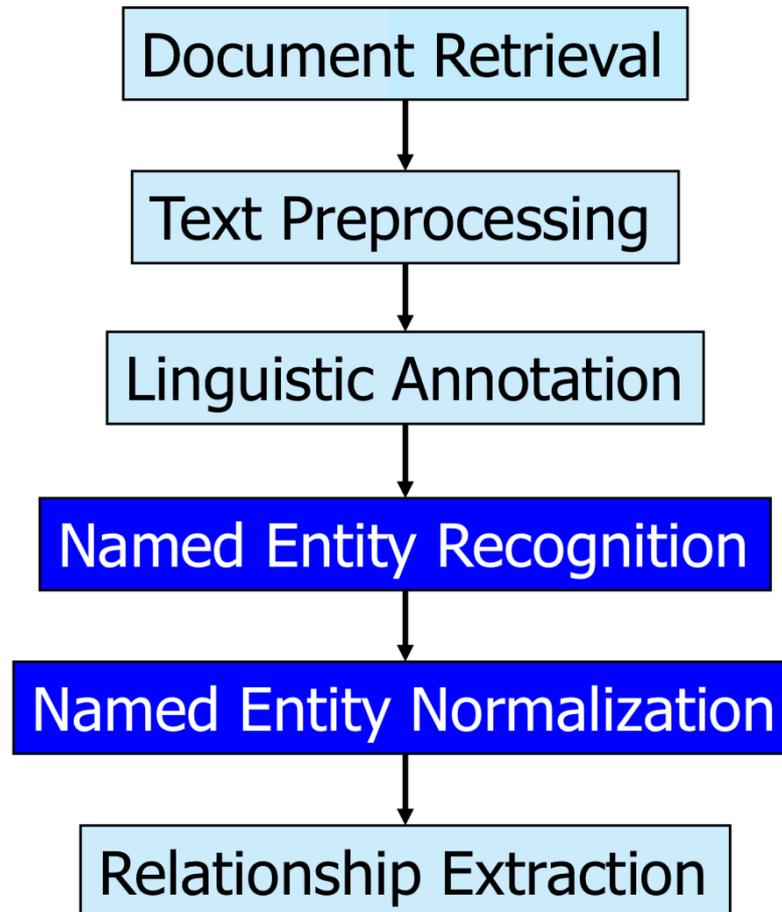
Übung

Aufgabe 4: Regelbasiertes Named Entity Recognition

Mario Sänger

# Informationsextraktionsworkflow

---



# Informationsextraktion

---

- Extraktion von Wissen aus unstrukturierten Texten
  - Ziel: Inhalt der Texte „verstehen“ und Informationen gewinnen
  - Spezialisierung auf eine bestimmte Interessensgebiet oder Domäne
- Beispiel: Erkennung von Unternehmensinformationen

*IBM* hat den deutschen Finanzsicherheitsspezialisten *IRIS Analytics* übernommen, teilt der Konzern mit.



*company\_takeover(„IBM“, „IRIS Analytics“)*

# Named Entity Recognition (NER)

---

- Erkennung von Entitäten ist eine der grundlegenden Problemstellungen in der Informationsextraktion
  - Lokalisierung und Klassifizierung aller Nennungen von Entitäten bzw. Eigennamen im Text
  - Entitäten: Personen, Organisationen, Orte, Datumsangaben, ...

*Barack Obama* und *Angela Merkel* versichern sich zum Abschied ihrer gemeinsamen Werte.

*IBM* hat den deutschen Finanzsicherheitsspezialisten *IRIS Analytics* übernommen, teilt der Konzern mit.

Der Online-Versandhändler *Amazon* hat pünktlich zum Weihnachtsgeschäft sein neuntes Versandzentrum in *Brieselang* eröffnet.

# BioMed Named Entity Recognition

---

- Entitäten: Genen, Proteinen, Mutationen, Zelltypen, Erkrankungen, Medikamenten, ...

*Z-100* is an *arabinomannan* extracted from *Mycobacterium tuberculosis* that has various immunomodulatory activities, such as the induction of *interleukin 12*, *interferon gamma* (*IFN-gamma*) and beta-chemokines. The effects of *Z-100* on *human immunodeficiency virus type 1* (*HIV-1*) replication in *human monocyte-derived macrophages* (*MDMs*) are investigated in this paper. In *MDMs*, *Z-100* markedly suppressed the replication of not only macrophage-tropic (M-tropic) *HIV-1* strain (*HIV-1JR-CSF*), but also *HIV-1* pseudotypes that possessed amphotropic *Moloney murine leukemia virus* or *vesicular stomatitis virus G* envelopes. *Z-100* was found to inhibit *HIV-1* expression, even when added 24 h after infection. In addition, it substantially inhibited the expression of the pNL43lucDeltaenv vector (in which the *env* gene is defective and the *nef* gene is replaced with the *firefly luciferase* gene) when this vector was transfected directly into *MDMs*. These findings suggest that *Z-100* inhibits virus replication, mainly at *HIV-1* transcription. However, *Z-100* also downregulated expression of the cell surface receptors *CD4* and *CCR5* in *MDMs*, suggesting some inhibitory effect on *HIV-1* entry. Further experiments revealed that *Z-100* induced *IFN-beta*

# Herausforderungen

---

- Mehrdeutigkeit von Begriffen
  - Ford (Unternehmen / Person), April (Monat / Person)
- Verwendung von allgemein-gebräuchlichen Wörtern
  - Allianz (Unternehmen), Dickkopf, Spätzle, a (Gene)
- Beachtung der Kontexts sehr wichtig!
- Keine exakte Definition der Entitätsklassen und -grenzen
  - Ist Bart Simpsons ein Person?
- Dynamische Domänen
  - Keine vollständige Auflistung aller Entitäten möglich
  - Beispiel: Stetige Neugründung von Unternehmen

# Ansätze (Skizze)

---

- Dictionary-basierte Verfahren
  - Auflistung aller Entitäten in einem Wörterbuch
  - (Fuzzy-) Matching der Wörterbucheinträge mit dem Text
- Regel-basierte Verfahren
  - Entwicklung von Regeln, welche Indikatoren und Situationen von Entitätsnennungen erfassen
  - Beispiel: [Organisation] übernimmt [Organisation]
- Einsatz von Maschinellem Lernen
  - Klassifiziere jeden Token als (Teil einer) Entität oder nicht
  - Lernen eines Modells basierend auf annotierten Trainingsdaten

# Aufgabenstellung

---

- Erkennung von Genen mit einem Regel-/Dictionary-basierten Ansatz
  - Verwendung von Maschinellem Lernen (SVM, HMM, CRF, ANN, ...) ist nicht erlaubt!
- Wir stellen einen Trainingskorpus mit annotierten Gen-/Proteinnamen
  - Korpus ist im IOB-Format annotiert
  - Alle Multi-Token Gene wurden entfernt (nur B-protein-Tags)

[https://hu.berlin/ue\\_maschsp1718\\_ue4\\_training](https://hu.berlin/ue_maschsp1718_ue4_training)

# Annotationsbeispiel

---

Number	O
of	O
glucocorticoid	B-protein
receptors	O
in	O
lymphocytes	O
and	O
their	O
sensitivity	O
to	O
hormone	O
action	O
.	O
The	O
study	O
demonstrated	O
a	O

# Aufgabenstellung

---

- Bereitstellung eines Gen-Wörterbuchs:
  - Auszug aus Entrez Gene (<https://www.ncbi.nlm.nih.gov/gene>) mit rund 100.000 Genen vom Menschen
  - Alle Namen sind single-token und lower-case
  - Das Wörterbuch kann beliebig erweitert und bearbeitet werden
  - Verwendung von anderen / weiteren Wörterbüchern ist erlaubt  
[https://hu.berlin/ue\\_maschsp1718\\_ue4\\_gene\\_names](https://hu.berlin/ue_maschsp1718_ue4_gene_names)
- Bereitstellung einer Liste mit ~500 englische Stoppwörtern  
[https://hu.berlin/ue\\_maschsp1718\\_ue4\\_stop\\_words](https://hu.berlin/ue_maschsp1718_ue4_stop_words)

# Aufgabenstellung

---

- Implementation eines regelbasierte Verfahren
  - Verwendung der Edit-Distance, N-Gram-Überschneidung, Stemming, reguläre Ausdrücke, eignes Fuzzy-Matching-Verfahren, ...
  - Einsatz von Maschinellern Lernen ist nicht erlaubt!
- Regeln müssen auf Trainingsdaten basieren:
  - OK: Zählen von POS-n-Gram-Pattern und Umwandlung dieser in spezifische Regeln
  - Nicht OK: Zählen von POS-n-Gram-Pattern und Umwandlung dieser in Features

# Aufgabenstellung

---

- Verwendung eines IE-Frameworks möglich
  - Java: LingPipe, GATE, UIMA, OpenNLP, ...
  - Python: NLTK, ...
  - Aber: Keine Klassifikation!
- Bei Unklarheit, ob ein Ansatz erlaubt ist oder nicht:  
Vorher mit mir abstimmen!

# Programmaufruf

---

- Implementation kann in Java, Scala oder Python erfolgen
- Programmablauf:
  - Programm liest (un-annotierte) Texte aus „*input\_file*“
  - Anwendung des regelbasierten Verfahrens auf die Texte
  - Programm schreibt annotierte Version der Text nach „*output\_file*“
- Aufrufsyntax:

```
java -jar uebung4-gruppeX.jar input_file output_file
```

```
python uebung4-gruppeX.py input_file output_file
```

# Ein- und Ausgabedatenformat

---

*input\_file*

Number	0
of	0
glucocorticoid	0
receptors	0
in	0
lymphocytes	0
and	0
their	0
sensitivity	0
to	0
hormone	0
action	0
.	0
The	0
study	0
... .	

*output\_file*

Number	0
of	0
glucocorticoid	B-protein
receptors	0
in	0
lymphocytes	0
and	0
their	0
sensitivity	0
to	0
hormone	0
action	0
.	0
The	0
study	0
... .	

# Evaluation

---

- Wir stellen ein Evaluationskript zur Verfügung
  - Prüft die Wirksamkeit der Regeln mit Hilfe des Skripts!
  - Scala: [https://hu.berlin/ue\\_maschsp1718\\_ue4\\_eval\\_sc](https://hu.berlin/ue_maschsp1718_ue4_eval_sc)
  - Jar-Archiv: [https://hu.berlin/ue\\_maschsp1718\\_ue4\\_eval\\_jar](https://hu.berlin/ue_maschsp1718_ue4_eval_jar)

```
java -jar uebung4-eval.jar goldstandard.iob prediction.iob
```

```
scala uebung4-eval.scala goldstandard.iob prediction.iob
```

- Bereitstellung der Testdaten (ohne Annotationen):  
[https://hu.berlin/ue\\_maschsp1718\\_ue4\\_test](https://hu.berlin/ue_maschsp1718_ue4_test)

# Evaluation - Ausgabe

---

```
// False positives
FP (RR1,3019,3022)
FP (RNAs,3064,3068)
...
// False negatives
FN (carboxyhemoglobin,961664,961681)
FN (transglutaminase,963570,963586)
...
// Result figures
True Positives:      869.0
False Positives:    979.0
False Negatives:    709.0

Precision:           0,470238
Recall:              0,550697
F1 Score:            0,507297
```

# Abgabedetails

---

- Abgabe eines ZIP-Archivs uebung4-gruppeX.zip
  - Ausführbares Programm und dessen Quellcode
  - Python: Auflistung der Dependencies (requirements.txt)
  - Ergebnisse des Verfahrens auf den Trainingsdaten
    - Verwendet hierzu das Evaluationsprogramm
- Testet Euer Programm vor der Abgabe!
  - F1-Score auf Trainingsdaten muss  $\geq 0,35$  sein
- Abgabe bis spätestens 22.01.2018, 23:59 Uhr über:  
[https://hu.berlin/ue\\_maschsp1718\\_ue4](https://hu.berlin/ue_maschsp1718_ue4)

# Wettbewerb

---

- Die Lösung mit dem höchsten F1-Score gewinnt
  - Evaluation erfolgt auf dem Test-Datensatz
  - Verwendung des Evaluationsprogramms
  - Laufzeit des Programms spielt keine Rolle!
- Die besten 3 Teams erhalten 5/3/1 Punkt(e)

---

# Fragen?