Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

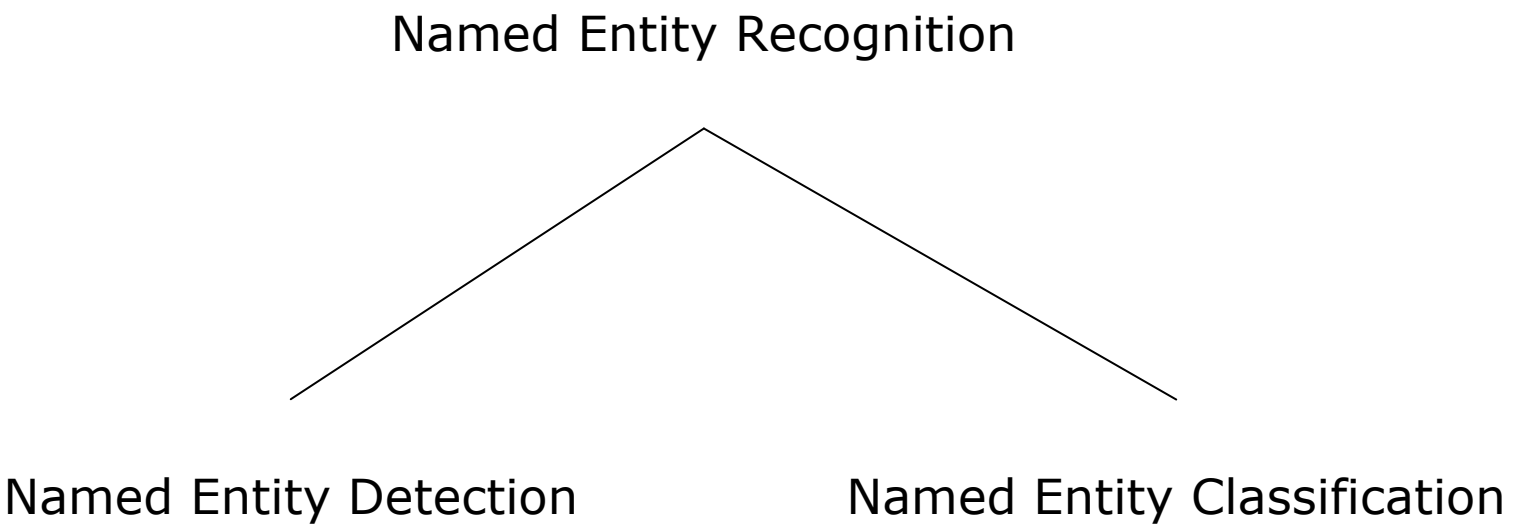
NER – Verfahren II
mit regulären Ausdrücken

von Ralf Gehrer und Fabian Lindenberg

Überblick

- Einführung
- Einteilung der Verfahren
- Merkmale regelbasierter NER-Verfahren
- Erkennung mittels regulärer Ausdrücke
  - Named Entities
    - Interne und externe Evidenz
  - Alterserkennung und Zuordnung am Beispiel
  - Titelerkennung

Einführung


Named Entity Recognition
Named Entity Detection
Named Entity Classification

Einteilung der Verfahren

- listenbasierte Verfahren
- regelbasierte Verfahren
- statistische Verfahren

} hybride Systeme

Merkmale regelbasierter NER-Verfahren

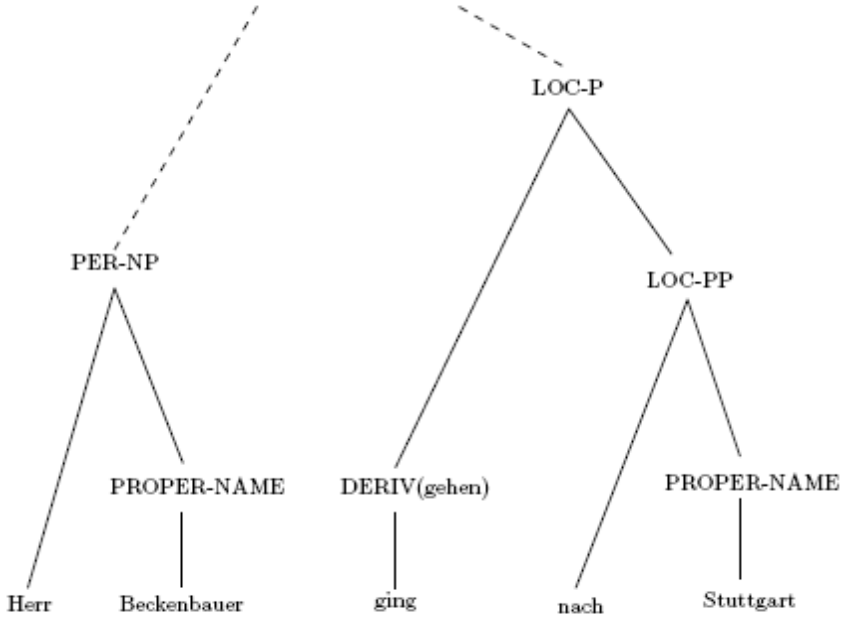
- Einsatz verschiedener Wissensformen
  - syntaktisch
  - morphologisch
  - domänenspezifisch
  - lexikalisch

Merkmale regelbasierter NER-Verfahren

- exakte Kontrolle des Prozesses
- manuelle Erstellung der Regeln
- sprach- und domänenabhängig
  - Expertenwissen und Heuristik
- kostenintensiver Wörterbuchabgleich nicht (zwingend) notwendig
- Kombination erhöht Präzision
  - Abdeckung verschiedener Fälle

Realisierung mittels Grammatiken

- Entwicklung verschiedener Produktionen



Realisierung mit regulären Ausdrücken

- kontextabhängige reguläre Ausdrücke
- gemeinsames Muster verschiedener Entities

- Muster zur Erkennung von gültigen Emailadressen
  - [A-Za-z0-9](([_\.\-]?[a-zA-Z0-9]+)*)@ ([A-Za-z0-9]+)(([\.\-]?[a-zA-Z0-9]+)*)\.([A-Za-z]{2,})

  - scarlett.johansson@prominentpeople.info
  - Mario.Gómez.García@prominentpeople.info

Deutsch vs. Englisch

Jeanne Carmen, 50er-Jahre Pin-Up und spätere B-Movie-Schauspielerin, wünscht sich Scarlett Johansson als Darstellerin ihrer früheren Freundin Marilyn Monroe für die Verfilmung ihrer Biografie *The Wild, Wild Life of Jeanne Carmen.* Ursprünglich soll Christina Aguilera für den Part im Gespräch gewesen sein, dieser kam allerdings ihre Schwangerschaft dazwischen und nun könnte Scarlett bald als Marilyn zu sehen sein. Angeblich ist auch noch Kate Bosworth für den geplanten Film im Gespräch, wir können uns aber selbst mit viel Fantasie nicht vorstellen, dass Hungerhaken Kate - die vom Typ überhaupt nicht den 50er Jahre-Schönheiten entspricht, entweder Monroe oder Carmen selbst spielt.

Marilyn Monroe's one-time best friend Jeanne Carmen wants Scarlett Johansson or Kate Bosworth to play her in a new biopic because mom-to-be Christina Aguilera is too focused on motherhood for the project. Pop star Aguilera was set to make her big-screen debut in the story of the busty blonde pinup, but Carmen can't wait for her to have a child and get back into shape. So she has revealed she wants either Johansson or Bosworth to take the role in the film, based on her son Brandon James' biography *The Wild, Wild Life of Jeanne Carmen*. Carmen says, "Christina was interested in playing the part and is a born pinup girl at heart. She became pregnant, however, and the script is now in the hands of Scarlett Johansson's agent and Kate Bosworth's agent.

Deutsch vs. Englisch

Jeanne Carmen, 50er-Jahre Pin-Up und spätere B-Movie-Schauspielerin, wünscht sich Scarlett Johansson als Darstellerin ihrer früheren Freundin Marilyn Monroe für die Verfilmung ihrer Biografie *The Wild, Wild Life of Jeanne Carmen.* Ursprünglich soll Christina Aguilera für den Part im Gespräch gewesen sein, dieser kam allerdings ihre Schwangerschaft dazwischen und nun könnte Scarlett bald als Marilyn zu sehen sein. Angeblich ist auch noch Kate Bosworth für den geplanten Film im Gespräch, wir können uns aber selbst mit viel Fantasie nicht vorstellen, dass Hungerhaken Kate - die vom Typ überhaupt nicht den 50er Jahre-Schönheiten entspricht, entweder Monroe oder Carmen selbst spielt.

Marilyn Monroe's one-time best friend Jeanne Carmen wants Scarlett Johansson or Kate Bosworth to play her in a new biopic because mom-to-be Christina Aguilera is too focused on motherhood for the project. Pop star Aguilera was set to make her big-screen debut in the story of the busty blonde pinup, but Carmen can't wait for her to have a child and get back into shape. So she has revealed she wants either Johansson or Bosworth to take the role in the film, based on her son Brandon James' biography *The Wild, Wild Life of Jeanne Carmen*. Carmen says, "Christina was interested in playing the part and is a born pinup girl at heart. She became pregnant, however, and the script is now in the hands of Scarlett Johansson's agent and Kate Bosworth's agent.

Deutsch vs. Englisch

Jeanne Carmen, 50er-Jahre Pin-Up und spätere B-Movie-Schauspielerin, wünscht sich Scarlett Johansson als Darstellerin ihrer früheren Freundin Marilyn Monroe für die Verfilmung ihrer Biografie *The Wild, Wild Life of Jeanne Carmen*. Ursprünglich soll Christina Aguilera für den Part im Gespräch gewesen sein, dieser kam allerdings ihre Schwangerschaft dazwischen und nun könnte Scarlett bald als Marilyn zu sehen sein. Angeblich ist auch noch Kate Bosworth für den geplanten Film im Gespräch, wir können uns aber selbst mit viel Fantasie nicht vorstellen, dass Hungerhaken Kate - die vom Typ überhaupt nicht den 50er Jahre-Schönheiten entspricht, entweder Monroe oder Carmen selbst spielt.

Marilyn Monroe's one-time best friend Jeanne Carmen wants Scarlett Johansson or Kate Bosworth to play her in a new biopic because mom-to-be Christina Aguilera is too focused on motherhood for the project. Pop star Aguilera was set to make her big-screen debut in the story of the busty blonde pinup, but Carmen can't wait for her to have a child and get back into shape. So she has revealed she wants either Johansson or Bosworth to take the role in the film, based on her son Brandon James' biography *The Wild, Wild Life of Jeanne Carmen*. Carmen says, "Christina was interested in playing the part and is a born pinup girl at heart. She became pregnant, however, and the script is now in the hands of Scarlett Johansson's agent and Kate Bosworth's agent.

Interne und externe Evidenz

- Hinweise innerhalb des Tokens
  - \+ großgeschrieben
  - \+ besonders hervorgehoben
  - \- beinhaltet Ziffer(n)
  - \- enthält Satzzeichen
  - \+ Großbuchstabe gefolgt von Punkt
  - \- Kleinbuchstaben mit abschließendem Punkt
- Eigenschaften der umliegenden Tokens
  - \+ erfüllen ähnliche Kriterien
  - \+ gleiche Formatierung
  - \+ führende Tokens lauten „Schauspielerin“ / „Herr“ / „Frau“ / bel. Titel
  - \- folgende Tokens lauten „Images“/ „Darsteller“ / „GmbH“ / …

Interne und externe Evidenz

- die Personensuchmaschine www.prominentpeople.info entwickelt werden
- möchte der „Harry Potter“ Darsteller
- amerikanische Schauspielerin **Scarlett Johansson** zu
- jetzt (Foto: Getty Images) ZUM
- laut Johannes B. Kerner soll

Zahlen über Zahlen

23-jährige BLZ 60069676 2,2 sek. 50€
148.000 Einwohner 22:10 Uhr 22.05.1978
D-14482 Potsdam (34) 4,99Euro 5 Stunden
Matrikl Nr. 224339 ISBN:0387954333 1,5% Zinsen
(+49) 331 5509 286 7 Stück Hausnr. 152a 29 Seite 13
1983 35 cm² Dortmund 2:0 Schalke 33 Meter
einunddreißig 1+2=3 §433(1) BGB
Windows 2003
Bestellnr. 1234 5678 30. Geburtstag

Reguläre Ausdrücke: Alterserkennung

- Altersangaben von 0-129 in Klammern: (23)
  \([1-9] | [1-9] [0-9] | 1 [0-2] [0-9] \)
- 0-129-jährig/e/er/es: 54-jährige
  ([1-9] | [1-9][0-9] | 1[0-2][0-9]) – [Jj]ährig(e | er | es | en)?
- Geburtsjahr von 1900-2099: 1984
  (19 | 20)[0-9][0-9]
- ausgeschriebene Zahlen zwischen zwanzig und neunundneunzig
  ((ein | zwei | drei | vier | fünf | sechs | sieben | acht | neun)und)?
  (zwanzig | dreißig | vierzig | fünfzig | sechzig | siebzig | achtzig | neunzig)

Reguläre Ausdrücke: Alterserkennung

- Datum (Achtung: evtl. Geburtsdatum): 11.12.2007
  (0?[1-9] | [1-2][0-9] | 3[0-1])\ .
  (0?[1-9] | 1[0-2])\.
  ((19 | 20)? [0-9] [0-9])

- 1.-129. Geburtstag
  ([1-9]|[1-9][0-9]|1[0-2][0-9])(\.|te|ten|ter) ?Geburtstag

- 1-129 Jahre alt
  ([1-9]|[1-9][0-9]|1[0-2][0-9]) Jahr(e)? alt(e|er)?

Zuordnungsproblem

- [...] Sean Stewart (26) wurde von der Polizei verhaftet, nachdem er ein Pärchen gewaltsam attackiert haben soll. Laut "dailymail.co.uk" hat der 26-jährige Sohn von Rocklegende Rod Steward angeblich eine Frau angegriffen, nachdem [...]

- [...] Prinz Charles wird nicht bei dem Gedenk-Konzert für Prinzessin Diana im Juli dabei sein. "Dafür bin ich zu alt", sagte der britische Thronfolger "contactmusic.com" zufolge. Stattdessen will der 58-Jährige zusammen mit seiner Frau Camilla an dem Gottesdienst für Diana Ende August teilnehmen. [...]

Zuordnungsproblem

- einzelne Jahreszahlen ungenau
- unterschiedliches Datumsformat
- irrelevante Daten von anderen Anlässen

- Bezug zu welcher Person ?
- Abstand nicht immer ausschlaggebend

→ eine Altersangaben notfalls mehreren Personen mit Wahrscheinlichkeit zuordnen

→ endgültige Festlegung in VerifiedPerson anhand aller gefundenen Altersangaben

Titelerkennung

- i.d.R. kein bestimmtes Schema
- Titel vs. Namenszusatz
  - Prof. oder Dr. vs. Sen. oder Jr.

→ listenbasiertes Verfahren

- teilweise Vorauswahl möglich
  - Bsp.: Prof. Dr. med. Max Musterman
    Diplom Ingenieur (FH) Hans Müller

Quellen

- Text Mining - Predictive Methods for Analyzing Unstructured Information, Sholom M. Weiss et. Al. , Springer Verlag
- Named Entity Recognition, Joel Lang, Uni Karlsruhe
- Der Stand der Kunst in der Eigennamen-Erkennung, Jeannette Roth, Universität Zürich
- http://www.alias-i.com/lingpipe/demos/tutorial/ne/read-me.html
- http://www.promiblogger.com und http://www.hollywood.com