RUHR UNIVERSITÄT BOCHUM

RUB

CLARIN-D

Forschungsinfrastruktur für Sprachressourcen in den Geistes- und Sozialwissenschaften

HUMBOLDT-UNIVERSITÄT ZU BERLIN

F-AG 7: Angewandte Sprachwissenschaft, Computerlinguistik

Kurationsprojekt 2

Linguistische Annotation von Nichtstandardvarietäten — Guidelines und „Best Practices"

Guidelines NER

Version 1.5

Stand: 20.09.13

Marc Reznicek

Marc.Reznicek@hu-berlin.de

Kafka@WebAnno

Guidelines für die Named Entity Recognition. Sie bauen auf den Guidelines in den STTS-Guidelines (Schiller et al. 1999), dem Stylebook for Tübingen Treebank (Telljohann et al. 2012) und der MUC-6 Named Entity Task Definition (Grishman 1995) auf.

Inhalt

Einführung: Named Entity Recognition

Unter der **Named Entity Recognition (NER)** versteht man die Aufgabe, **Eigennamen (named entities)** in Texten zu erkennen. Technisch gesehen sind hierzu zwei Schritte notwendig. Zuerst müssen in einem laufenden Text die Token gefunden werden, die zu einem Eigennamen **gehören (Named Entity Detection: NED)**, danach können diese Eigennamen semantischen Kategorien zugeordnet werden **(Named Entity Classification).**

Prototypisch ist dabei der Unterschied zwischen **Eigennamen** und **Appellativa** der, dass letztere eine Gattung oder eine Klasse beschreiben, während erstere einzelne Individuen oder Sammlungen von Individuen unabhängig von gemeinsamen Eigenschaften bezeichnen (Gallmann 2009:149).

Die vorliegenden Guidelines sollen es Annotatoren ermöglichen, Eigennamen in Texte aus Standard und Nichtstandard-Varietäten konsistent zu annotieren.

In diesen Guidelines werden die beiden Aufgaben der NED und NEC nicht unterschieden, da die Konzentration auf Beispiele in diesem Dokument, die Trennung künstlich erzeugen müsste und nicht zu erwarten ist, dass die Resultate sich dadurch verbessern würden.

In Anlehnung an die oben genannten Guidelines für Zeitungssprache werden in NoSta-D **vier semantische Hauptklassen** unterschieden (**Personen, Organisationen, Orte und Andere**). Diese werden teilweise um spezifische Verwendungen erweitert (virtuelle Orte), Eigennamen, die Teile größerer Einheiten sind, werden als solche markiert (NEpart), oder Appellativa, die von Eigennamen abgeleitet sind, werden gesondert behandelt (NEderiv).

Wie finde ich eine NE?

Schritt 1:

- Nur **volle Nominalphrasen** können **NEs** sein. Pronomen und alle anderen Phrasen können ignoriert werden.

Schritt 2:

- Namen sind im Prinzip Bezeichnungen für einzigartige Einheiten, die nicht über gemeinsame Eigenschaften beschrieben werden.

Beispiel:

*[**Der Struppi**] folgt [**seinem Herrchen**].*

- Hier gibt es zwei Nominalphrasen als Kandidaten für einen Eigennamen (NE).
- "Der Struppi" bezeichnet eine einzige Einheit. Es kann auch mehrere Struppis geben, aber diese haben an sich keine gemeinsamen Eigenschaften, bis auf den gemeinsamen Namen, daher handelt es sich um einen Eigennamen.
- "seinem Herrchen" bezeichnet zwar (typischerweise) auch nur eine einzige Person allerdings können wir diese nur über die Eigenschaft identifizieren, dass sie ein Herrchen ist und dass dies für Struppi zutrifft. Struppi könnte auch mehrere Herrchen haben, die alle die Eigenschaften teilen, die ein Struppi-Herrchen beinhaltet (z.B. darf Struppi streicheln, muss ihn ausführen und füttern etc.)

Schritt 3:

- Determinierer sind keine Teile des Namens.

*Der [**Struppi**]***NE*** *folgt seinem Herrchen.*

Schritt 4:

- Eigennamen können mehr als ein Token beinhalten.

Beispiel:

Viele Personennamen (PER für *person*):

[**Peter Müller**]**PER**

Filmtitel (OTH für *other*):

[**Knockin' on Heavens Door**]**OTH**

Schritt 5:

- Eigennamen können auch ineinander verschachtelt sein.

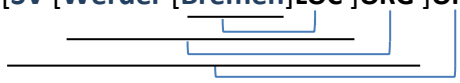
Beispiel:

Personennamen in Filmtiteln:

[[**Shakespeare**]**PER in Love**]**OTH**

Orte (LOC für *location*) in Vereinsnamen (ORG für *organisation*):

[**SV** [**Werder** [**Bremen**]**LOC**]**ORG**]**ORG**



Schritt 6:

- **Titel**, **Anreden** und **Besitzer** gehören **NICHT** zu einem komplexen Eigennamen. Besitzer können natürlich selber Eigennamen sein**.**

Beispiel:

Referenz auf Musiktitel:

[**Vivaldis**]**PER** [**Vier Jahreszeiten**]**OTH**

Referenz auf Personen:

Landesvorsitzende Frau [**Ute Wedemeier**]**PER**

Schritt 7:

- Eigennamen treten auch als **Teil eines komplexen Tokens** auf. Hier wird für das gesamte Token annotiert, dass es einen Eigennamen enthält.

Beispiel:

mit Firmen Assoziiertes:

[**DAEWOO-Millionen**]**ORG**part

mit bestimmten Personen verbundene Erfindungen/Arbeiten:

[**Hartz-Reformen**]**PER**part

[**Ottomotor**]**PER**part

ABER:

- Wenn auch das Gesamttoken einen Eigennamen darstellt, dann wird nur dieser annotiert.

Beispiel:

Stiftungen:

[**Böll-Stiftung**]**ORG**

Schritt 8:

- Kann in einem Kontext nicht entschieden werden, ob eine NP sich **als Eigennamen oder Appellativ** verhält, wird es nicht als NE markiert.

Beispiel:

Ortsnamen vs. -beschreibungen:

> …und zogen mit ihren großen Transparenten gestern vom [**Steintor**] über den [**Ostertorsteinweg**]**LOC** zum [**Marktplatz**].

Zu welcher semantischen Klasse gehört ein Eigenname?

- Wenn der Namenskandidat in der Liste unter der Klasse "**keine NE**" aufgeführt wird, dann handelt es sich nicht um eine NE im Sinne dieser Guidelines.
- Wenn der Eigenname in eine der Klassen in der Liste **NoSta-D-TagSet** gehört, dann annotiere die zugehörige Klasse.
- In Zweifelsfällen hilft auch die Liste der **Einzelfälle** weiter.
- Wenn der Eigennamen in **KEINE** der vorhandenen Klassen passt, dann markieren ihn mit OTH und notiere Dir bitte das Beispiel und schicke uns eine E-Mail an: marc.reznicek@staff.hu-berlin.de. So können wir die Guidelines sukzessive verbessern.

Wie finde ich Ableitungen von NEs?

Schritt 1:

- Eigennamen, die durch morphologische Derivation in andere Wortarten überführt wurden, werden als solche markiert.

Beispiel:

Ortsadjektive:

> die [**Bremer**]**LOC**deriv Staatsanwaltschaft

Personenadjektive:

> die [**Merkelsche**]**PER**deriv Begeisterung für Europa

Zu welcher semantischen Klasse gehört eine Ableitung?

- Die Klasse setzt sich aus dem Tag der Klasse zusammen, in die der ursprüngliche Eigennamen gehört und dem Marker für die Ableitung "deriv".

Beispiel:

Ortsadjektive:

> [**Bremen**] → **LOC**
>
> die [**Bremer**]**LOC**deriv Staatsanwaltschaft

Faustregel zur Unterscheidung einer Klassenbezeichnung und eines Namens

Faustregel zur Unterscheidung einer Klassenbezeichnung und eines Namens:

- Die Elemente der fraglichen Einheit verbinden die gleichen Eigenschaften → Klasse →keine NE
  - Christen glauben an Christus → Christ glaubt an Christus → keine NE
- Die Elemente der fraglichen Einheit verbindet nur der Name oder Element ist Einheit bezeichnet ein spezifisches Individuum → Name → NE
  - Barock bezeichnet spezifische Epoche

ABER: die [Deutschen]**LOCderiv**

NoSta-D-TagSet

Semantische Klasse	Semantische Subklasse	Beispiele	Ausnahmen
PER - Person	Person	*Hans Winkler*	
	Zuname	*(Familie) Feuerstein*	
	Tiernamen	*(Schweinchen) Babe*	
PERderiv Personenableitungen		*Merkelsches Kabinett*	
PERpart Personennamen in Token		*Scherf-Angriff*	
ORG - Organisation	Organisationen	*Nato, EU*	
	Unternehmen	*Microsoft, Bertelsmann*	
	Betreiber	*Lotto 6 aus 49*	
	Institute	*Institut für chinesische Medizin*	
	Museen	*Pergamonmuseum*	
	Zeitungen	*Süddeutsche Zeitung, Der Spiegel*	
	Clubs	*VfB Stuttgart*	
	Theater, Kinos	*Metropol-Theater, CinemaxX*	
	Universitäten	*Freie Universität*	
	Rundfunksender	*Arte, Radio Bremen*	
	Restaurants und Hotels	*Sassella, Adlon*	
	Militäreinheiten	*Blauhelme*	
	Modelabels	*Chanel*	
	Sportereignisse	*Olympische Spiele, Wimbledon*	
	Bands	*Beatles, Die Fantastischen Vier*	~~Parlament~~
	Institution	*Bundestag*	
	Bibliotheken	Amerika Gedenkbibliothek	
	Parteien	*SPD, CDU*	
ORGderiv Organisations-ableitungen		*dieses Appelsche Schulbuchprojekt*	
ORGpart Organisationsnamen in Token		*Google-Hauptquartier*	

LOC - Ort	Bezirke	*Schöneberg*	
	Sehenswürdigkeiten, Kirchen	*Brandenburger Tor,Johanniskirche*	
	Planeten	*Mars*	
	Landschafts-bezeichnungen	*Königsheide*	
	Straßen, Plätze	*Söogestraße, Alexanderplatz*	
	Berge, Seen, Flüsse	*Alpen, Viktoriasee, Spree*	
	Kontinente	*Europa, Asien*	
	Länder, Staaten	*Frankreich, Hessen, Assyrien*	
	Städte	*Berlin, Babylon*	
LOCderiv - Ortsableitungen	Wettbewerbe	*Deutsche Meisterschaft*	Wenn spezifisch, dann ORG.
LOCpart - Ortsnamen in Token		*Deutschlandlied*	
OTH - Andere	Betriebssysteme	*DOS*	
	Buch-, Filmtitel etc.	*Faust, Schlaflos in Seattle*	
	Mottos, Slogans	*Zwischen Himmel und Erde*	
	Kriege	*Zweiter Weltkrieg*	
	Politische Aktionen	*7. Bremer Protesttag gegen Diskriminierung*	
	Projektnamen	*Agenda 21*	
	Währungen	*Euro*	
	Marktindex	*Dow Jones, Dax*	
	Reihennummerierungen	*SZ-Magazin 41/07*	
	Sprachen	*Englisch, Deutsch*	
	Buchtitel mittels Autor	*Helbig et al. ([Helbig]*PER *et al.)*OTH	
	Epochen	*Barock, Romantik (auch Neubildungen: „Neuzeit“))*	
	Webseiten	*www.ebay.de, google, www*	
	Sprachen	*Hochdeutsch, Englisch*	
OTHderiv – Ableitung von OTH	modifizierte Sprachadjektive	***hochdeutsche** Verben*	
OTHpart– OTH in Token		***Euroskeptiker***	
VLOC - virtueller Raum	Chatraum	*Katzenkörbchen, Erdbeerworld*	
Keine NE	Datums- und Zeitangaben	*Montag, April, Feiertage*	
	Religionen	*Christentum, muslimisch*	~~Götternamen~~
	Tiernamen	*Gepard, Schlange*	
	Bezeichner/Fachwörter	*Phosphat, Geodäten, Ikonen*	~~Produkt-namen: Aspirin~~
	Himmelsrichtungen	*südlich, Norden*	
	Titel/Anrede	*Frau, König*	
	Dynastien und Geschlechter	*Habsburger, Wittelsbacher*	
	Politische Strömungen	*Kommunismus, Sozialismus*	

Regeln

Regel	Beispiele	NE
Klassen werden unabhängig von der semantischen Rolle im Kontext vergeben. ABER: Grammatische Hinweise entscheiden.	Nils Petersen geht … **zu** [Bremen]**ORG** **nach** [Bremen]**LOC** Die [Wolfsburger]**LOCderiv** entwickeln Spitzentechnik. (eigentlich VW in Wolfsburg)	✔
Ableitungen NEderiv werden nur dann annotiert, wenn sie den Stamm mit einer NE teilen.	die [decartessche]**PERderiv** Philosophie	✔
	die anglikanische Kirche	✖

Formen

Regel	Beispiele	NE
abgetrennt Kompositionsglieder	*in [West-]**LOC** und ganz besonders in [Ost-Berlin]**LOC*** *[Adenauer-]**ORG** und [Böll-Stiftung]**ORG***	✔
Ortsteile	*[West-Afrika]**LOC*** *[Nord-Berlin]LOC*	✔
Adelstitel	*Herr [von [**Hohenzollern**]**LOC**]PER*	✔
Abkürzungen	*Amis, Sowjets*	✖

Einzelfälle

Begriff	Semantische Klasse	Semantische Subklasse	Kommentar
Bundesliga	**ORG**	Organisationen	
Creditreform-Mittelstandsindex	**ORGpart**	Unternehmen	
Darmstadtium	**ORG/LOC**	Veranstalter / Veranstaltungsort	kontextabhängig
Bibel	**OTH**	Buchtitel	
Hotel Bellevue	**ORG**	Hotels	
Milchstraße	**LOC**	Himmelskörper	*wie Planeten*
Evangelium	**keine NE**	Bezeichnung	
Gott		Bezeichnung	
Polizei & Feuerwehr		Gruppen	
Indianer		Bezeichnung	einzelne Stämme : ORG
ISBN		Bezeichnung	

NoSta-D: Chat

Personennamen:

Personennamen (PER) in Chat sind häufig nur über den Bezug auf die Alias der erwähnten Sprecher zu erkennen. In Abbildung 1 können die Tokens „Lantonie“, „lantonieeeeee“, „LANTOOO“, „Lantööööö“ und „lanto“ über die Normalisierung (siehe Guidelines Vorverarbeitung) erkannt und klassifiziert werden.

#	Sprecher	Beitrag
25	system	Lantonie betritt den Raum.
26	Lantonie	:)
27	quaki	lantonieeeeee
28	Lantonie	Hallo. :)
29	zora	LANTOOO :)))
34	marc30	Lantööööö :o)
35	TomcatMJ	hi lanto
40	Faryen-Angle	hi Lantonie

Abbildung 1: Personennamen in Chat
http://www.chatkorpus.tu-dortmund.de/files/releasehtml/html-korpus/unicum_21-02-2003_1.html

Bibliografie

- **Schiller, Anne; Teufel, Simone; Stöckert, Christine; Thielen, Christine (1999):** Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical Report. University of Stuttgart; University of Tübingen. BibTex
- **Telljohann, Heike; Hinrichs, Erhard W.; Kübler, Sandra; Zinsmeister, Heike; Beck (2012):** Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Seminar für Sprachwissenschaft. Universität Tübingen, Germany. http://www.sfs.uni-tuebingen.de/resources/tuebadz-stylebook-1201.pdf. BibTex
- **MUC-6 Named Entity Task Definition**: http://cs.nyu.edu/faculty/grishman/NEtask20.book_1.html