Richtlinien zum Annotieren von Named Entities

Tom Hombergs
Praktische Informatik VII
FernUniversität Hagen
23.01.2006

Inhaltsverzeichnis

1 Aufgabenstellung

Dieser Guide ist eine Sammlung von Vorschlägen zur Annotation von Eigennamen in einem Korpus. Eigennamen werden hier nur in die Klassen **PER** (Namen von Personen), **LOC** (Namen von Lokationen), **ORG** (Namen von Organisationen) und **MISC** (alle anderen Namen) eingeteilt. Bei der Annotation wurden folgende Annahmen getroffen:

- Eine Phrase kann nur als Eigenname höchstens einer Klasse markiert werden, es gibt keine optionalen oder disjunktiven Annotationen.
- Eigennamen überlappen sich nicht.
- Bei Eigennamen in Eigennamen (“Postbank Bochum”) wird nur der gesamte Eigenname markiert.

2 Richtlinien für alle Klassen

2.1 Adjektivische Attribute

Bei Großschreibung des Adjektivs ist es häufig Teil eines Eigennamens, bei Kleinschreibung vermutlich nicht:

- Freiwillige Feuerwehr
- [Oppenheimer Landstraße]$_{LOC}$
- das Langener Parlament
- [Oberurseler Waldjugend]$_{ORG}$
- allgemeines Krankenhaus
- katholische Pfarrei [Urberach]$_{LOC}$ *aber*: der [Evangelische Pressedienst]$_{ORG}$
- [Müngersdorfer–Stadion]$_{LOC}$
- die [Grünen]$_{ORG}$, aber: die Liberalen, die Sozialen
- [sozialdemokratische Partei Österreichs]$_{ORG}$

2.2 Eingebettete Eigennamen

Konform zu [Chinchor(1997), Abschnitt A.1.3]:

- „[Abenteuer am Mississippi]$_{MISC}$“ (Filmtitel; keine Markierung für „Mississippi“)
- [Postbank Bochum]$_{ORG}$ (keine eigene Markierung für „Postbank“ oder „Bochum“)
- [Max Müller GmbH]$_{ORG}$ (keine Markierung für „Max Müller“)

2.3 Ellipsen und Trunkierung

- [Nord–]$_{LOC}$ und [Südamerika]$_{LOC}$ (widerspricht [Chinchor(1997), Abschnitt 4.1.1], dort wäre es als [Nord– und Südamerika]$_{LOC}$ markiert)
- [Offenbacher–]$_{LOC}$ oder [Nelizystraße]$_{LOC}$

Für diese Art von Eigennamen erweist sich tokenbasiertes Tagging als unpassend, da in obigen Beispielen die Ausdrücke „Offenbacher–“ und „Nord–“ allein als Eigenname interpretiert würden!

2.4 Fremdsprachliches Material

- [Four Seasons Hotel]$_{ORG}$
- [Machinists Union]$_{ORG}$
- „[The Godfather]$_{PER}$“ *aber*: der „[Godfather]$_{PER}$“
- [The White House]$_{ORG}$ *aber*: das [Weiße Haus]$_{ORG}$

2.5 Interpunktion

- die „[Drake]$_{LOC}$“– und die „[Edwards]$_{LOC}$“–Kaserne
- *aber*: in Zukunft wird es eine „[Lex Nieder–Roden]$_{MISC}$“ geben
- die [Brecht / Weill–Chansons]$_{MISC}$
- das [„Grüne Punkt“ – Duale System]$_{MISC}$

In vielen Fällen ist tokenbasiertes Tagging hier unpassend, da Anführungszeichen manchmal Teil eines Token sind, manchmal nicht. Die Behandlung von Anführungszeichen als einzelne Token führt auch zu widersprüchlichen Markierungen (vergleiche Punkte 1 und 4, einmal sind Anführungszeichen Teil des Eigennamens, einmal nicht).

2.6 Komposita

Ein Kompositum wird nur als Eigennamen markiert, wenn es in seiner Gesamtheit ein Name ist, oder sein Kopf ein Name ist.

- UNO–Generalsekretär [Annan]$_{PER}$
- Microsoft–Chef [Ballmer]$_{PER}$
- die Kennedy–Familie
- Clinton–Regierung
- Macintosh–Computer
- Kerberöffnung (Kerb: hessisch für Kirmes)
- Europameister
- [Nobelpreis]$_{MISC}$ für Physik
- [Kreis-SPD]$_{ORG}$, [Kreis-CDU]$_{ORG}$
- [Stufenhecku–Carina]$_{MISC}$ (Carina: Automarke)

2.7 Nummerierung

Werden Nummern zur Identifikation von Entitäten benutzt, so wird die Nummer nur als Teil des Eigennamen markiert, wenn sie ein unabdingbarer Bestandteil ist:

- [1. Weltkrieg]$_{MISC}$
- 24. Miltenberger Rockfestival
- [1. FC Köln]$_{ORG}$
- Ortsbeirat 2, Ortsbezirk 5
- [B257]$_{LOC}$ ([Bundesstraße 257]$_{LOC}$)

2.8 Appositionen

- Stadt [Frankfurt]$_{LOC}$
- Kreis [Offenbach]$_{LOC}$
- Hotel „[Vier Jahreszeiten]$_{ORG}$“
- Kreisverband [Offenbach–Land]$_{LOC}$
- *aber*: [Frankfurt City]$_{LOC}$ (oder: [Frankfurt–City]$_{LOC}$)
- [Schutzgemeinschaft Deutscher Wald]$_{ORG}$ („Deutscher Wald“ hätte ohne den Zusatz „Schutzgemeinschaft“ eine völlig andere Bedeutung)

2.9 „Unauffällige“ Eigennamen

- Bundesliga, Zweite Liga, Kreisliga Nord (keine Markierung)
- die [Naturfreunde]$_{ORG}$

3 Richtlinien für LOC

3.1 Eigennamen der Klasse LOC

- Länder, Städte, Regionen, Gemeinden
- Berge, Flüsse, Seen, Meere
- Straßen, Brücken, Häfen
- Sportplätze, Turnhallen, Treffpunkte (z.B. Veranstaltungsorte, Jugendheime)

3.2 Geographische Eingrenzungen

Analog zur Annotation von Komposita. In den folgenden Fällen ist der Kopf des Kompositums jeweils ein Eigenname.

- [Nordatlantik]$_{LOC}$
- [Westeuropa]$_{LOC}$
- [Ostdeutschland]$_{LOC}$
- [Südosteuropa]$_{LOC}$

3.3 Straßennamen

- [B 380]$_{LOC}$ ([Bundesstraße 380]$_{LOC}$)
- [Schmidtstr.]$_{LOC}$ 150

4 Richtlinien für MISC

4.1 Eigennamen der Klasse MISC

- Währungen, Gesetze
- Titel von Veranstaltungen und Ereignissen
- Titel von Büchern, Zeitschriften, Musikstücken, Aufführungen
- Fahrzeugnamen, Produktnamen

4.2 Historische Eigennamen

- der [Erste Weltkrieg]$_{MISC}$
- die Wende (es könnte jede historische Wende sein)
- die Mauer (es könnte jede Mauer sein)

4.3 Titel von Veranstaltungen, Schriften, Konferenzen ...

Titel werden nur als MISC markiert, wenn der Titel eindeutig der Name für etwas ist. Nicht alles, was in Anführungszeichen steht ist automatisch ein Eigenname (siehe Punkt 3).

- [Frau und Technik – na klar?]$_{MISC}$ (Titel eines Seminars)
- [Aktionstage Umwelt]$_{MISC}$ '92
- *aber*: ... zum Thema „Militär und Konversion im Raum [Hanau]$_{LOC}$"...
- 10. Köpperner Bachfest

4.4 Produktnamen

- [VW Passat]$_{MISC}$, [Ford Escort]$_{MISC}$

5 Richtlinien für ORG

5.1 Eigennamen der Klasse ORG

- Organisationen, Firmen, Institutionen
- Vereine, Gruppen, Zeitschriften, Zeitungen
- Schulen, Hochschulen

5.2 Deutsche Institutionen

- der Bund
- die Bundesregierung
- die Bundesrepublik
- [Amt für Versorgung und Besoldung]$_{ORG}$
- [Bundeswehr]$_{ORG}$
- Arbeitsamt [Wiesbaden]$_{LOC}$
- das Forstamt
- **Grenzfälle:** [Bundesrat]$_{ORG}$, [Bundestag]$_{ORG}$, [Bundesgesundheitsministerium]$_{ORG}$

5.3 Zeitschriften und Zeitungen

In Zeitungskorpora wird für Zeitungen und Zeitschriften hauptsächlich die Lesart als Organisation benutzt und nicht die Lesart als bedrucktes Papier.

- die [Süddeutsche Zeitung]$_{ORG}$ berichtet...
- in einem Interview mit der [Washington Post]$_{ORG}$...
- *aber:* er zerreißt die [Süddeutsche Zeitung]$_{MISC}$

5.4 Keine Markierung

- Ober–Mörler Sozialdemokraten
- Mainzer Universitätsklinik
- Christ– und Sozialdemokraten

5.5 Ortsangaben

Im Allgemeinen gehört der Ort nicht zum Eigennamen, wenn „in“ zwischen den Ort und den Rest des Ausdruckes gesetzt werden kann ohne die Bedeutung zu verändern (z.B. „Altenheim (in) Bischofsheim“ im Gegensatz zu „Sportgemeinschaft (in) Wattenscheid“).

- Altenheim [Bischofsheim]$_{LOC}$
- Flughafen [Frankfurt]$_{LOC}$
- Umlandverband [Frankfurt]$_{LOC}$
- *aber*: [DLRG]$_{ORG}$ [Frankfurt]$_{LOC}$, [Sportgemeinschaft Wattenscheid]$_{ORG}$, [Tanzsportclub Rödermark]$_{ORG}$

5.6 Konstruktionen aus Eigennamen und Nicht–Eigennamen

- Amt des Frankfurter Oberbürgermeisters (keine Markierung)
- [Kirche der Unifikation]$_{ORG}$
- *aber*: [Amt für Umwelt, Tiefbau und Abfallwirtschaft]$_{ORG}$
- die [Junge Union]$_{ORG}$, die [Jungen Liberalen]$_{ORG}$
- **Grenzfall:** Jugendbildungswerke von Stadt und Kreis [Offenbach]$_{LOC}$

5.7 Vereine, Gemeinden

Konform zu [Chinchor(1997), Abschnitt A.1.6]. Komplette Vereinsnamen werden als ORG markiert (TV Wicker). Wird aber nur der Heimatort des Vereins genannt, wird er als LOC markiert (Wicker).

- [TV Wicker]$_{ORG}$ (*aber*: [Wicker]$_{LOC}$ gewann gegen den [TV Gelnhausen]$_{ORG}$ . . .)
- . . . die [Eintracht]$_{ORG}$ hätte den besten Torwart . . .
- Gemeinde [St. Bonifatius]$_{ORG}$, Matthäus–Gemeinde (Lesart als Kirchengemeinde)
- *aber:* Gemeinde [Schwanheim]$_{LOC}$, Gemeinde [Nauheim]$_{LOC}$ (Lesart als regionale Gliederung)

5.8 Von Lokationen abgeleitete (substantivierte) Adjektive

Grundsätzlich werden von Eigennamen abgeleitete Adjektive oder substantivierte Adjektive *nicht* als Eigenname markiert.

- Italiener
- Bochumer
- deutsch
- italienisch
- ein Berliner Bürger

6 Richtlinien für PER

6.1 Eigennamen der Klasse PER

- Vornamen, Familiennamen, zusammengesetzte Namen
- Tiernamen

6.2 Titel

Konform zu [Chinchor(1997), Abschnitt A.1.2]:

- Bundeskanzler [Schröder]$_{PER}$
- UNO–Generalsekretär [Annan]$_{PER}$
- Lord [Baden–Powell]$_{PER}$ of [Gilwell]$_{LOC}$

6.3 Vollständige Namen

- [Max M. Mustermann]$_{PER}$
- [Otto Normalverbraucher]$_{PER}$
- [Johann Friedrich Eosander Freiherr von Göthe]$_{PER}$

6.4 Tiernamen

- Daktari [Thompsons]$_{PER}$ Bernhardiner [George]$_{PER}$
- der Schäferhund [Lassie]$_{PER}$

Literatur

[Chinchor(1997)] Chinchor, Nancy (1997). MUC-7 Named Entity Task Definition Version 3.5. In *Proceedings of the Seventh Message Understanding Conference (MUC7).*