Named Entity Recognition

Anne Schiller, XRCE, December 2008

The XIP Grammar for German provides a list dependencies and features for Named Entity Recognition as descibed below.

Named Entity Recognition aims to detect and categorize proper nouns, which can be single nouns like "*China*", or more complex names like *"George Washington"*.

Named entities are output as unary dependencies. Moreover, some features involved for Named Entity Recognition can be output as well, that can give more fine-grained Named Entity Recognition (e.g. distinguish first name and full name of a person).

The different types of named entities that can currently be recognized are the following:

- Locations: *Frankreich, Rhein, Zugspitze, Bodensee*
- Person mames: *George W. Bush, Kofi Annan, Kleopatra*
- Organization and business names: *NATO; Xerox; Müller AG*
- Names of artefacts: *Grundgesetz; Cola; BGB*
- Date and time expressions: *12. August; 7 Uhr 30; 12:45*

Note 1: Proper names are often ambiguous with common words (c.f. English "Bush"), and named entity recognition must resolve such ambiguity. Spelling rules of most languages (with Latin script) indicate proper names with a capital letter at the beginning which may help for disambiguation. In German, however, cot only proper names, but also common nouns start with a capital letter. While this may help to resolve syntactic ambiguities (like adjectives/nouns or verb/noun), it makes recognition of named entities more difficult, if not impossible without a wider context.

Example: (1) *Herr Müller kommt mit Herrn Hund.*
Mr. Miller comes with Mr. Dog.
(2) ... u*nd Müller kommt mit Hund*
... and Miller comes with Dog
... and (the) miller comes with (his) dog.

Note 2: The actual lexicons for organizations and artefacts are far from being exhaustive. They may be extended for general coverage, but it may be more appropriate to use customized lexicons for specific applications (e.g. political news, medical articles, etc.). Especially the problem of ambiguities with common nouns can be improved when the application domain is known.

1 Dependencies for Named Entities

Dependency	Feature	Example
LOC	CONTINENT	Asien
	COUNTRY	Bundesrepublik Deutschland; Frankreich
	STATE	Baden-Württemberg; Arizona
	REGION (includes mountains, islands, etc.)	Böhmen; Zugspitze; Sylt
	TOWN	Berlin; London; New York
	RIVER	Rhein; Amazonas; Themse
	LAKE	Bodensee; Genfer See
	SEA	Nordsee; Mittelmeer
PERSON	FIRSTNAME	Anna; Hans-Otto
	LASTNAME	Goethe; von Bora;
	MAN	Georg Wintermantel; Hans; Herr Hamburg
	WOMAN	Bettina von Arnim; Annemarie
	GROUP	Familie Maier
	DEMONYM = inhabitant of a country, town or region (combined with a LOC feature)	Franzose; Sachse; Stuttgarter
ORG	BUS	Xerox;
	LOC	Europäische Union; Vereinigte Staaten von Amerika
		ZDF; NATO; Greenpeace
ARTEFACT	DOC (= document)	Grundgestetz; StVO
	PROD (=product name)	(der) Daimler; (eine) Cola; (im) Duden
DATE		24. 12. 2008; 1. August
	HOLIDAY	Ostern; Allerheiligen
	YEAR	2009
	MONTH	Mai
	DAY	(am) Montag (dem) 13.
TIME		14.20 Uhr; 16 Uhr 20; 23:17

2 Examples

2.1 Geographic Entities

Hans wohnt in Singen und kommt zum Essen nach Wyk auf Föhr.

LOC_TOWN_REGION(Wyk)
PERSON_MAN_FIRSTNAME(Hans)
LOC_TOWN(Singen)
LOC_REGION(Föhr)

- *"Essen" is not misinterpreted as a city name.*

Die Schweden und die Sachsen reisen gerne nach Böhmen und Bayern.

LOC_REGION(Böhmen)

LOC_STATE(Bayern)

PERSON_COUNTRY_DEMONYM(Schweden)

PERSON_STATE_DEMONYM(Sachsen)

Die Böhmen und die Bayern reisen gerne nach Schweden und Sachsen.

LOC_COUNTRY(Schweden)

LOC_STATE(Sachsen)

PERSON_REGION_DEMONYM(Böhmen)

PERSON_STATE_DEMONYM(Bayern)

- *disambiguation between country names and names of inhabitants (demonym)*

Der Rhein fliesst in den Bodensee und dann zwischen Baden und dem Elsass.

LOC_RIVER(Rhein)
LOC_LAKE(Bodensee)
LOC_TOWN_REGION(Baden)
LOC_REGION(Elsass)

Die Rhone fliesst vom Genfer See ins Mittelmeer.

LOC_RIVER(Rhone)

LOC_LAKE(Genfer See)

LOC_SEA(Mittelmeer)

2.2 Person Names

Herr Hamburg ist vom 1.12. bis Weihnachten in Berlin.

DATE(1.12.)
DATE(Weihnachten)
PERSON_MAN_LASTNAME(Hamburg)
NMOD_TITLE(Herr, Hamburg)
LOC_TOWN_STATE(Berlin)

- *"Berlin" is both a state and a city.*
- *Titles("Herr", "Frau", "Prof.") are not part of person names, but NMOD (noun modifier).*

Familie Reiter hat 2009 in Stuttgart gelebt, sagte Müller.
DATE(2009)
PERSON_GROUP_LASTNAME(Reiter)
NMOD_TITLE(Familie, Reiter)
PERSON(Müller)
LOC_TOWN(Stuttgart)

- *Proper names that are ambiguous with common nouns (e.g. "Reiter" = horseman, "Müller" = miller) are disambiguated as far as possible according to the syntactic or semantic context.*

Annamaria Wintermantel spricht mit Hans Ofenrohr.
PERSON_WOMAN_NAME(Annamaria Wintermantel)
PERSON_MAN_NAME(Hans Ofenrohr)

- *This is another example for ambiguity between common nouns and proper names.*

Kaiser Karl V. besuchte an Ostern 1234 Friedrich den Kahlen.
DATE(Ostern 1234)
PERSON_MAN_FIRSTNAME(Karl V.)
PERSON_MAN(Friedrich den Kahlen)

Ein Franzose und eine Chinesin sind in der englischen Botschaft in Frankfurt am Main.
LOC_TOWN(Frankfurt)
LOC_RIVER(Main)
PERSON_COUNTRY_DEMONYM(Franzose)
PERSON_COUNTRY_DEMONYM(Chinesin)

2.3 Organisation names

Die neugegründete Sachsen AG fusioniert mit der Maeyer und Müller GmbH.
ORG_BUS(Sachsen AG)
ORG_BUS(Maeyer und Müller GmbH)

Die Europäische Union unterstüzt die NATO.

ORG_LOC(Europäische Union)

ORG(NATO)

Note: Other subtypes of organizations such as political, sports, etc. are not available yet.

2.4 Names of Artefacts

Volkswagen baut einen neuen VW.

ORG_BUS(Volkswagen)

ARTEFACT_PROD(VW)

Note: The actual version of the grammar does not identify product names that correspond to common nouns, such as "Käfer" (="bug" or "(VW) Beetle"), or "Tempo" (="speed" or name of a paper towel).

Viele Leute kennen das Grundgesetz und die StVO (Straßenverkehrsordnung).

ARTEFACT_DOC(Grundgesetz)

ARTEFACT_DOC(StVO)

ARTEFACT_DOC(Straßenverkehrsordnung)

2.5 Date and Time Expressions

Einsendeschluss ist der 31.12.2008 oder der 15. 1. 2009.

DATE(31.12.2008)

DATE(15. 1. 2009)

Er kommt im Mai oder an Weihnachten oder erst 2010.

DATE_HOLIDAY(Weihnachten)

DATE_YEAR(2010)

DATE_MONTH(Mai)

Am Montag, dem 13., kommt er wieder.

DATE_DAY(Montag)

DATE_DAY(13.)

Er kommt am 17. Mai um 17.15 Uhr.

DATE(17. Mai)

TIME(17.15 Uhr)

Der Film läuft von 17 Uhr 30 bis 19.00 Uhr.

TIME(17 Uhr 30)

TIME(19.00 Uhr)