# By the way: what are named/funky entities?

Irene Cramer

IGK Doctoral Colloquium – Saarland University

January 2006

# Intro

My research question:

Explore semi-supervised learning methods to train a classifier that annotates classical named and funky entities.

# Intro (cont.)

- Today I'll focus on two aspects of my work:
  - definitions and
  - characteristics (◊ features) of named and funky entities respectively

# Intro (cont.)

I don't want to reinvent the wheel, but

- MUC/CoNLL definitions of named entities are "sort of fuzzy",
- naturally, there are no definitions of funky entities at all …

# Outline

- Intro
- Entities
  - Named Entities (◊ Proper Nouns)
  - Funky Entities
  - Construction of an inter-annotator agreement test
- Feature sets for semi-supervised learning

# Entities – named entities

"The Named Entity task consists of three subtasks (entity names, temporal expressions, number expressions). The expressions to be annotated are "unique identifiers" of entities (organizations, persons, locations), times (dates, times), and quantities (monetary values, percentages)."

MUC-6 guidelines

# Entities – named entities (cont.)

In 1989, U.S. Fish and Wildlife Service devoured 175 to 180 million dollars.

In <TIMEX TYPE="DATE">1986</TIMEX>, <ENAMEX TYPE="ORGANIZATION">U.S. Fish and Wildlife Service</ENAMEX> devoured <NUMEX TYPE="MONEY">175</NUMEX> to <NUMEX TYPE="MONEY">180 million dollars</NUMEX>.

# Entities – named entities (cont.)

"Named entities are phrases that contain the names of persons, organizations, locations, times and quantities."

CoNLL 2002 guidelines

# Entities – named entities (cont.)

```
Wolff B-PER
, O
currently O
a O
journalist O
in O
Argentina B-LOC
, O
played O
with O
Del B-PER
Bosque I-PER
in O
the O
final O
years O
of O
the O
seventies O
in O
Real B-ORG
Madrid I-ORG
. O
```

O: no entity; B-PER: beginning of a person name;
I-PER: within a person name etc.

# Entities – named entities (cont.)

Named entities (more or less) equal proper nouns.

# Entities – proper nouns (cont.)

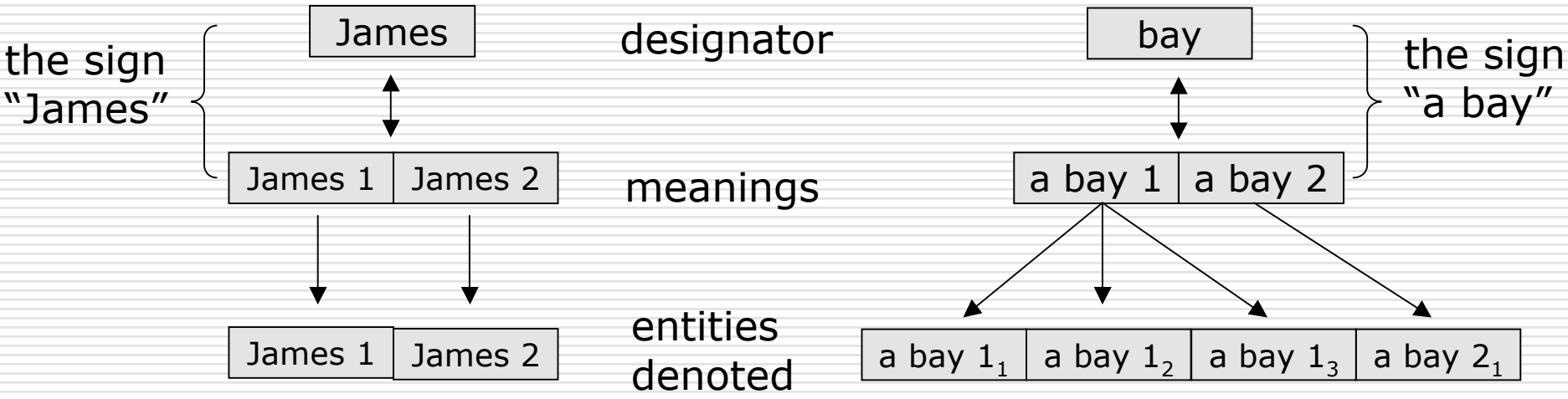
- Proper nouns:
  - are linguistic more precisely lexical signs (◊ semiotic signs);
  - concerning linguistical properties, they resemble substantives (nouns);
  - (in most languages proper nouns are capitalized.)

# Entities – proper nouns (cont.)

- Naive assumption: proper nouns refer to just "*one thing*" in the world...but
  - Jane, Chris, etc. certainly refer to more than one person...
  - the current president of France certainly refers to *one* person but is not a proper noun all the same

# Entities – proper nouns (cont.)

- Naive assumption: proper nouns refer to just "*one thing*" in the world...




adapted from a graphic in Wimmer (1973): Der Eigenname im Deutschen

# Entities – proper nouns (cont.)

There is consensus among lexicographers that proper nouns aren't part of dictionaries

◊ encyclopaedic knowledge!

# Entities – proper nouns (cont.)

- Most frequent proper nouns are:
  - person names (real and fictive ones)
  - toponyms, hydronyms (cities, countries, rivers, etc.) and
  - bodies (companies, organisations, etc.).

  MUC and CoNLL
- In addition there are names of:
  - creatures (domestic animals),
  - products (Big Mac, Cleanex, etc.)
  - (political, historical ...) events, awards,
  - ...

  ???

# Entities – funky entities

- (Currently my) funky entities
  - are those proper names that MUC/CoNLL do not cover;
  - equal lexical signs (= words, word groups ≠ phrases);
  - refer to entities in the real or a fictional world.

note: this doesn't violate classical named entity definitions

# Entities – funky entities (cont.)

- develop guidelines: how to annotate funky entities.

- use these guidelines to conduct an inter-annotator agreement test:

  ϑ subjects most welcome ϑ

  annotation of the funky entities syntactically will follow CoNLL guidelines

# Entities – inter-annotator agreement test

- Crucial questions:
  - Do “normal people” (and linguists) have an intuition about proper nouns?
  - How do they identify proper nouns?
  - Is it possible to classify proper nouns according to (more or less) unambiguous classes of entities?
  - What is the accuracy of that labelling?

# Entities – inter-annotator agreement test

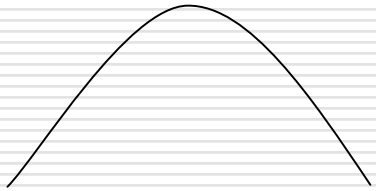
- Subjects will annotate small corpus:

- 3 test phases:
  - entity boundaries
  - entity labelling
  - both

derive final annotation guidelines

# Entities – inter-annotator agreement test

- entity boundaries



sum counts of $w_n$ classifications: if $w_n$ classified as entity add 1, else 0

| $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ | $w_6$ | … | $w_n$ |
|---|---|---|---|---|---|---|---|
| x | x | n | n | x | x | … | x |

  - explore "borderland" of the entities
  - derive new features?

# Entities – inter-annotator agreement test

- entity labelling
  - guidelines include:
    - specification of the classes (taxonomy),
    - “boundary description”
    - examples
  - result analysis ◊ standard deviation / accuracy

| | per | org | … | award |
|---|---|---|---|---|
| entity 1 | 2% | 87% | … | … |
| entity 2 | 89% | 5% | … | … |
| … | … | … | … | … |
| entity n | … | 14% | … | 78% |

# Entities – funky entities (cont.)

- After the inter-annotator agreement test: start to annotate a test corpus with these guidelines

  ◊ gold standard for my experiments!

# Outline

- Intro
- Entities
  - Named Entities (◊ Proper Nouns)
  - Funky Entities
  - Construction of an inter-annotator agreement test
- Feature sets for semi-supervised learning

# Feature sets

- Why extensively analyze features…
- …isn't that contradicting my semi-supervised approach?

à to keep the approach as flexible as possible, I consider all features

à however, at the moment, I am glad about every bit of information I could possibly get

# Feature sets

| feature | class | complexity |
|---|---|---|
| char n-grams | n-gram level | ? |
| suffix & prefix | word level | + |
| special char, cap. | word level | + |
| POS | synt. level | ++ |
| POS of context | synt. level | ++ |
| words in context | sem. level | + |
| trigger words in context | sem. level | ++ |
| one meaning per context | text level | + |
| word in gazetteer or lexicon? | look up | ++ |
| partial matches | look up | + |

# Feature sets:
# Correlation Entity – cap/alpha

| | | | | | |
|---|---|---|---|---|---|
| | | | | | |
| | | | .9 | 5 | 4 |
| | | | 9 | 0.5 | * 3 |
| | | | * | * | 0.45 |

*. 9

# Feature sets:
# Correlation Entity – POS

| | | | | |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

| | | | | |
|---|---|---|---|---|
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |
| | | | | |

# Feature sets:
# Correlation Entity – POS in window

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |
| | | | | | | |

# Feature sets:
# Correlation Entity – POS in window

| | | | | | |
|---|---|---|---|---|---|
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| | | [illegible] | .1 | [illegible] | [illegible] |
| | [illegible] | [illegible] | * | [illegible] | [illegible] |
| | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| | [illegible] | [illegible] | * | [illegible] | [illegible] |
| | [illegible] | [illegible] | 6 | [illegible] | [illegible] |

*: [illegible]

| | | | | | [illegible] |
|---|---|---|---|---|---|
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| | | [illegible] | . | [illegible] | [illegible] |
| | [illegible] | [illegible] | * | [illegible] | * |
| | | [illegible] | [illegible] | [illegible] | [illegible] |
| | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
| | | [illegible] | [illegible] | [illegible] | [illegible] |

*: [illegible]

# Feature sets:
# Correlation Entity – POS in window

| | | | | | |
|---|---|---|---|---|---|
| | | | | | |
| | | | 0.90 | 0.90 | 0.90 |
| | | | * | .72 | .52 |
| | | | 0.95 | 0.52 | . |

*

*

| | | | | | |
|---|---|---|---|---|---|
| | | | | | |
| | | | 0.86 | 0.46 | *0.00 |
| | | | *0.48 | 0.61 | *0.08 |
| | | | 0.66 | 1.4 | 0.66 |

*

# Feature sets:
# Correlation Entity- Feature sets

| feature | correl. coeff. | sig. |
|---|---|---|
| suffix & prefix | -- | -- |
| character combi | -- | -- |
| special char, cap. | -0.295/-0.064 | highly |
| POS | 0.089/-0.41 | highly |
| POS of context | ≤ \|0.1\| | sometimes |
| words in context | -- | -- |
| trigger words in context | -- | -- |
| one meaning per context | -- | -- |
| word in gazetteer or lexicon? | -- | -- |
| partial matches | -- | -- |

# Feature sets

- test char n-grams ◊ mutual information?
- derive new feature sets form inter-annotator agreement test ◊ what do humans do to recognize entities?

# Thank you!