Named Entities

TM the Great Unread

DTL|Digital Arts Initiative
Interacting Minds Centre|Aarhus University

July 28, 2016

Named Entity Recognition

As usual, it is all about adding structure to unstructured data $\Rightarrow$ we want to identify *who*, *where* and *when*

Named entities are unique identifiers for entities (a proper noun serving as a name for someone or something)

Given concrete semantics, the task is to locate elements in text that fit those semantics

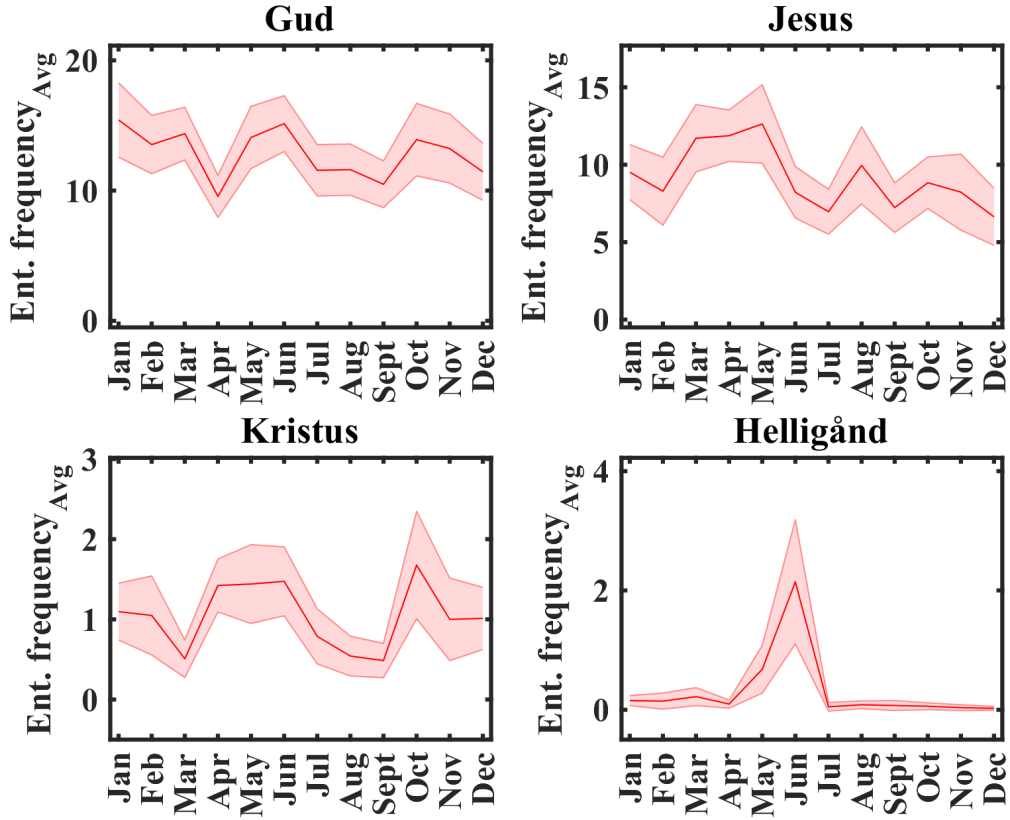
Detection of proper names and classification into categories: person, localization, organization

Recognition of other data/time, percentage, money, email addresses ...

Domain-specific entities (e.g., bioNLP)

```
> person.v
 [1] "David"          "Solomon"        "Mary"           "Joseph"
 [5] "Emmanuel"       "JESUS."         "Rachel"         "Herod"
 [9] "John"           "Jordan"         "Simon"          "Peter"
[13] "James"          "Lord"           "Moses"          "Isaac"
[17] "Matthew"        "Thomas"         "Art"            "Behold"
[21] "Satan"          "Philip"         "John Baptist"   "O"
[25] "Simon Peter"    "Simon Barjona"  "Jesus"          "Elias"
[29] "Remove"         "Of"             "Thou"           "Friend"
[33] "Grant"          "Be"             "Christs"        "Surely"
[37] "Eli"            "Truly"          "Mary Magdalene" "He"
```

NER in the humanities



Entities

It is not always clear what a named entity is ($\sim$ not a solved problem)

- Proper noun (lack of inflexions or determinants, lack of lexical meaning, use of capital letters)
- Rigid designator (identify same object or individual)
- Unique identifier/referent
- Purpose and domain of application

Implementation

NER is not just matching strings with a pre-defined list.

Rely on a language model which is dependent on the language and entity type it was trained for (models are trained on annotated corpora).

In *R* there is currently no good way of doing this, but we use Maxent name finder from the Apache OpenNLP library and models from datacube.wu.au.at (limited to DA, DE, EN, ES, IT, NL, PT, SV).

Pipeline: import text $\Rightarrow$ make annotators for tokenization $\Rightarrow$ make NER annotators ($\Rightarrow$ do analysis)

Issue and applications

Easy? Boundaries between categories are not impermeable:

> "France$_{\text{organization}}$ did not win the UEFA Euro 2016"
> "The UEFA Euro 2016 took place in France$_{\text{localization}}$"

NER is central in information extraction and is often used for pre-processing in classification tasks. Often used for semantic annotation, ontology learning, and opinion mining

In humanities and social sciences it is often combined with network analysis.

Extract list of characters and compare spatial and temporal distribution (of a plot).

BibleNLP