



europaena
newspapers



Development of Named Entities Recognition for French Newspapers

Journée d'information
Europeana Newspapers

27/11/2014

BnF / Paris, France

Clemens Neudecker, State Library Berlin

@cneudecker

What is „Named Entity Recognition“?

- Named Entity Recognition (NER) is a sub-task of Information Extraction and is typically understood as being part of the area of Computational Linguistics / Natural Language Processing.
- The main aim of NER is the automatic extraction and classification of knowledge or information from semantically unstructured text.
- NER is still subject to academic research (cf. Google & MSR Competition) – practical use in the cultural heritage digitisation sector remains a rare case.

Asked differently: What is a „Named Entity“?

- PERSON:
 - Names of persons and families, but also names of fictional persons („Albert Einstein“, „Präsident der USA“, „Micky Maus“)
- ORGANISATION:
 - Names of companies, governmental or non-governmental organisations („IBM“, „The Beatles“, „Labour Party“)
- PLACE:
 - Cities, Provinces, Counties, geographical areas, asf. („Paris“, „Haute-Pyrénées“, „Alpes“)

NER (I)

1. Detection/Classification of person names, places and organisations in a running text (includes POS)

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell-Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:

LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

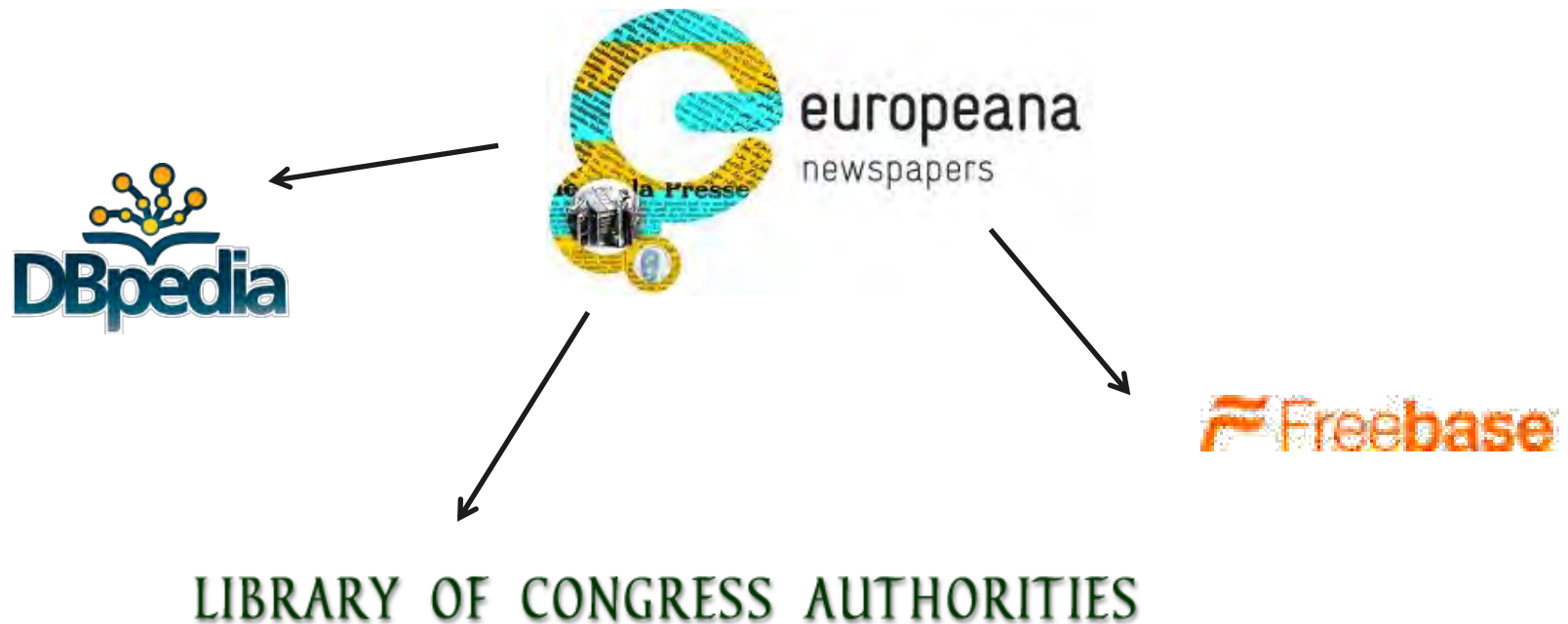
NER (II)

2. Disambiguation of terms (Example “Jordan”) through contextual information



NER (III)

3. Linking to authority files and online databases (Linked Data)



Supported languages in ENP

3 Languages:

- German
- Dutch
- French

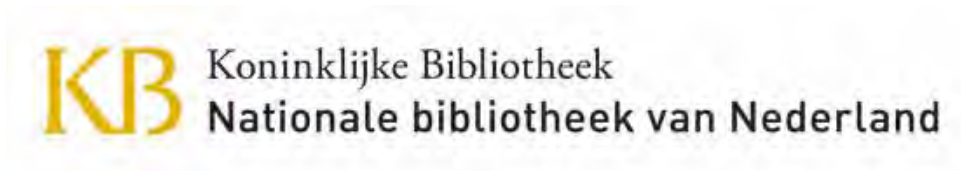


Approaches

- Machine learning vs. rule-based
- Advantages of machine-learning systems:
 - No need for specific linguistic expertise
 - Processing of large amounts of material
- Advantages of rule-based systems:
 - Can be tuned to very high accuracy for particular texts
 - Adaptation to local grammar and specific text style

Software

- Open Source ML software developed by Stanford University, adapted and extended for Europeana Newspapers by the KB National Library of the Netherlands



- Software is available as open source from Github for download and testing:
<https://github.com/KBNLresearch/europeanp-ner>

Training

- Training the NER systems with the help of manually annotated corpora („gold corpus“) and gazetteers



Europeana Newspapers Named Entities Attestation Tool

Select your language:



Telegrafisch Berigt. PARIJS, Maandag 7 Februarij. De Keizer heeft vergadering in persoon geopend, en bij die gelegenheid eene troc volgende voorkomt: Ondanks Frankrijks bloei bestaan er geruchten brengen. Dit is na zoo veel om wentelingen hoe wel ook te bet zonder grond en allezins verrassend, en be wijst den twijfel aan politiek is altijd geweest de orde in Europa te bewaren en aan alliantie met Engeland te bevestigen en ten opzigte der mogendh jegens heit^pl PER (key: p) uden met hunne welwillendhei voor eenige j, ORG (key: o) de vrede", om te toonen da Door Engeland LOC (key: l) nannen werd deze gezindheid alle we derva NOT KNOWN (key: n) bezorgde ons vrede in het C ons. Daarenteg MISC (key: m) bedwezen, was het kabinet va

- Publication of annotated data from ENP as open data

Encoding

- Results of NER are stored in a library specific format: ALTO (Analyzed Layout and Text Object)
- Versions > 2.1 of ALTO specifically allow to use NER „Tags“

```
<String STYLEREFS="ID7" HEIGHT="132.0" WIDTH="570.0" HPOS="5937.0"  
VPOS="3279.0" CONTENT="Reynolds" WC="0.95238096" TAGREFS="Tag5"></String>  
<String STYLEREFS="ID7" HEIGHT="102.0" WIDTH="540.0" HPOS="18438.0"  
VPOS="22008.0" CONTENT="Baltimore" WC="0.82539684,, TAGREFS="Tag10"></String>  
...  
<Tags>  
  <NamedEntityTag ID="Tag5" TYPE="Person" LABEL="Reynolds"/>  
  <NamedEntityTag ID="Tag6" TYPE="Place" LABEL="Baltimore"/>  
</Tags>
```

Problems and challenges

- OCR errors reduce the accuracy of the classification and slow down the overall processing time for recognition due to high noise.
 - Historical spelling variation for person names and place names in particular.
 - In many cases the historical spelling variants can not be found in online knowledge bases.
- Specific adaptation of the software via external modules

Initial results: Dutch

	Persons	Places	Organisations
Precision	0.940	0.950	0.942
Recall	0.588	0.760	0.559
F-measure	0.689	0.838	0.671



Thank you for your attention!
Merci de votre attention!

[@eurnews](https://twitter.com/eurnews)

<http://www.europeana-newspapers.eu>

<http://www.theeuropeanlibrary.org/tel4/newspapers>

<http://www.europeana.eu/>