



Large-scale refinement of digital historical newspapers with named entity recognition



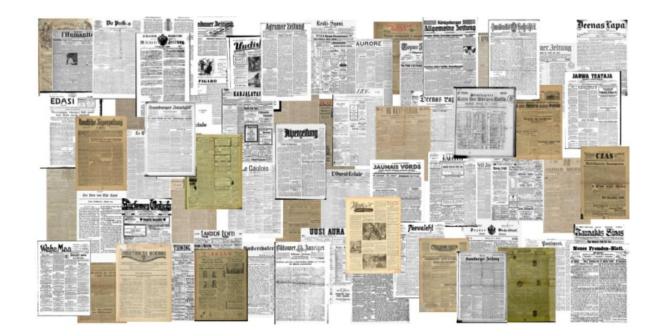
IFLA Newspaper Pre-Conference

14 August 2014, Geneva

Clemens Neudecker, SBB, @cneudecker

Overview

- Background
- NER Introduction
- Approach
- Challenges
- Scalability
- First results
- Outlook

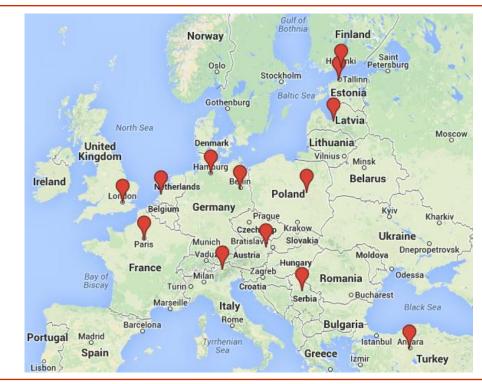






Background

- <u>Europeana Newspapers</u>
 EU Best Practice Network
- **10** million newspaper pages with full-text from 12 libraries
- 36 million newspaper pages with metadata for <u>Europeana</u>







Named entity recognition (I)

Detect names of persons, places, organisations

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





Named entity recognition (II)

2. Disambiguate entities









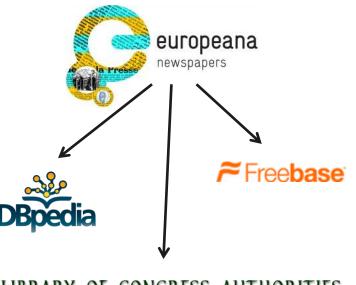
This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community



http://ec.europa.eu/ict_psp

Named entity recognition (III)

3. Link to online resources



LIBRARY OF CONGRESS AUTHORITIES





Approach (I)

Tackle content in Dutch, German, French (about 50% of the 10m pages)







Approach (II)

Koninklijke Bibliotheek Nationale bibliotheek van Nederland



 Use a machine learning tool (open source) developed by Stanford University, adapted for Europeana Newspapers by KBNL

https://github.com/KBNLresearch/europeananp-ner





Approach (III)

 Create (and release) training material by manually annotating named entities on OCR'd newspaper pages

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) iden met hunne welwillendhei ORG (key: o) voor eenige i de vrede", om te toonen da LOC (key: l) Door Engeland nannen werd deze gezindheid NOT KNOWN (key: n) alle we derwa bezorgde ons vrede in het (ons. Daarente and the first first mot we have been was het kabinet va





Challenges

- OCR quality
- Multiple (mixed) languages

EXMOUTH, July 13th. Since my last arrived the Pylades, Elliot; and the Minerva, Creese, from Neath, with cum; John, Mellow, from Plymouth, with sundries; Friendship, Shears, from Löndon, with wine and groceries; Salem, Baker, from Liverpool, with coals.--Salled the Alkion, Grace, for London, with sundries; "Wo Brothers, Wheaton; Pylades," Elliot; Harmony, Fishley; Lovely Peggy, Griffith, and the Jane, Redman, for Neath, Newport, and Milford; Two Brothers, Lackes, for Lyanington, with homschold furniture; Betsey, Perrian; Industry, Hodd r, and the Alexi, Jennings, for Plymouth, with timber, dec. Mary Anu, Hawkins, for Portsmouth, in hallast; Mary, Blundel, for Bristol; Resbuck, Finch, For Harlland, in ballast. la 112 B ik e my lat arrived the >Pylades,-. lliot; aod. Abe- Sineva, CNeee 4orn Neath, ' titch , cuim; ,'t;ohn_ IoMelwl fri ytiil SUn-.dieë; ,Frietndilp, St&ar, froniidon, 'Ui wine and grocerieu ;: ;aletn, Bker, from Liverpool,. witfi ecal.;' 4Stalled the AluidonG.: ceror' Lkndon, with sundries; : ;Two Rrothwsj'ê Whe-atn-;- Pylade', Eiot; Har'tinny,; ;: Fisbley:::Iiiveiy Peggy:-(flth add tie JAne, Redman, for eathly Newpot;agd llford; -Tw Br.otherAs, lawces, fos Lysixowjvithbinehol V pirI-ihzure;vi etsey, PerwIliti; iIudstry, ModA - -tbi, Al-t,'enniugs, for .:IPI-iOntI, StIth Ltu .c*ar An'l? Hawkinss foir ouck, + iii ballasto I ~ ~ ~ ~ ~~Ai

Historical spelling





- Stanford NER software is multi-threaded e.g. 4 CPU cores – 4x throughput
- Optimise the NER classifier by filtering noise and sentences without NE's marked
- Robust proven Java technology









First results (Dutch)

	Persons	Locations	Organizations
Precision	0.940	0.950	0.942
Recall	0.588	0.760	0.559
F-measure	0.689	0.838	0.671





First results (French)

	Persons	Locations	
Precision	0.529	0.548	* Score for organisations
Recall	0.834	0.216	omitted since not enough present in the source material
F-measure	0.622	0.310	





Outlook

- Q3: Release of training data for Named Entity Recognition in Dutch, German, French
- Q3: First results for German (Austrian, Italian/South Tirol), final results for Dutch, French
- Q4: Release of software (open source) for disambiguating and linking of NER results to DBPedia







www.europeana-newspapers.eu/

www.theeuropeanlibrary.org/tel4/newspapers

https://github.com/KBNLresearch/europeananp-ner

Thank you for your attention!



IFLA Newspaper Pre-Conference

14 August 2014, Geneva

Clemens Neudecker, SBB, @cneudecker