



## Accessing the Republic. Entity extraction from the resolutions of the Dutch States-General

Marijn Koolen,<sup>1,2</sup> Esger Renkema,<sup>1,2</sup> Nienke Groskamp,<sup>1</sup> Frank Smit,<sup>1</sup> Jirsi Reinders,<sup>1</sup> Ronald Sluijter,<sup>1</sup> Rik Hoekstra,<sup>1,2</sup> Jorid Oddens<sup>1</sup>

<sup>1</sup>Huygens Institute, Amsterdam, Netherlands <sup>2</sup>DHLab - KNAW Humanities Cluster, Amsterdam Netherlands

DH Benelux 2024 - 6 June 2024 - Leuven, Belgium

## REPUBLIC



- REsolutions PUBLished In a Computational environment
  - Resolutions of the States General of the Dutch Republic (1576-1796)
  - Long serial publication (220 years), over 500,000 pages
  - ~60,000 daily meeting, ~1 million resolutions (propositions and decisions)



## Motivation



#### • Reasons to tag entities

- Additional access points: alternative paths to navigate between documents
- Contextual information: quick assessment of relevance/interest for user
- Reasons not to tag entities
  - If quality is low, it can annoy users, induce a lack of trust
  - Added value may not outweigh required effort
- Which entities?
  - Which entities occur?
  - Which are interesting?
  - Which are taggable?

## Entities in the Resolutions



#### • Tagging 8 types of entities

- Person: person name including any attributions (title, job, legal status, ...)
- Attribution: person attribution (title, job, legal status, ...) if refers to specific entity
- Organisation: any organisation (incl. region name when it refers to governing body)
- Committee: members of the States General tasked to investigate a matter
- (Geo)Location Political entity: name of a geolocation when it refers to the place
- Date: absolute and relative dates (of submitted propositions and previous resolutions)
- Resolution reference: references to specific earlier resolutions
- Other: any remaining names

# Tagging Project

- Ground Truth
  - 1631 tagged resolutions of printed volumes 1705-1796
    - 370,560 tokens, 23,875 entities
  - 513 tagged paragraphs of handwritten volumes 1597-1702
    - 28,387 tokens, 2,347 entities
- Automated tagging
  - Train and evaluate taggers per entity type



## Nested Entities and Ground Truth



- Entities can be highly complex, with multiple levels of nesting
  - E.g. person name + attribution
  - Attribution can contain an organisation which can contain another organisation which can contain a location
- Examples
  - Henricus Gerhardus de Beveren Esveld, Predikant in de Gereformeerde Gemeente te Schoondyke onder het Classis van Walcheren, be roepen zynde tot Predikant in de Gemeente te Enkhuisen
  - Jan van Reusen Solliciteur van den heer Thibaut heer van St. Aechtekercke, Burgermeester der Stadt middelburch cum socijs, taeckende de Dijckagie genaempt de Polder benoorden Aerdenburch

# Training NER Taggers

huygens instituut

- Combine/compare types of embeddings
  - Character-level embeddings, trained on resolution corpus (150 million words)
  - FastText embeddings, trained on resolution corpus
  - GysBERT (Manjavacas & Fonteyn 2022)
- Use Python Flair package (Akbik et al. 2019)
  - <u>https://flairnlp.github.io/flair/</u>
  - Agnostic to type of embeddings: Character-, word-, sentence-level
  - Combine via stacked embeddings!

## Combined Model or Model per Type?



- Single model
  - Advantages:
    - Need to train only one model, no tag conflicts in applying
  - Disadvantages:
    - Model choices may not be optimal for all entity types
    - Difficult to determine hierarchy of nested entities (ground truth is flattened)
- One per type
  - Advantages:
    - pick optimal model per type, allow for partial overlapping entities
    - Ground truth contains all information
  - Disadvantages:
    - Need decide how to deal with tag conflicts (partial overlap)

## Quantitative Evaluation - Best Model Per Type





## Evaluation



| Entity type     | Tag repr.   | Embeddings                 | Prec. | Recall | F1   | Support |
|-----------------|-------------|----------------------------|-------|--------|------|---------|
| Person          | single type | GysBERT + Char.            | 0.81  | 0.69   | 0.75 | 405     |
| Person attr.    | single type | GysBERT + FastText + Char. | 0.57  | 0.56   | 0.56 | 573     |
| Organisation    | single type | GysBERT + FastText + Char. | 0.82  | 0.71   | 0.76 | 283     |
| Committee       | single type | Char.                      | 1.00  | 0.73   | 0.85 | 41      |
| Location        | single type | GysBERT + FastText         | 0.79  | 0.76   | 0.77 | 570     |
| Date            | single type | Char.                      | 0.90  | 0.88   | 0.89 | 249     |
| Resolution ref. | all types   | GysBERT + FastText         | 0.82  | 0.70   | 0.75 | 57      |
| Other names     | single type | GysBERT + FastText + Char. | 0.63  | 0.26   | 0.36 | 47      |

All best models use RNN instead of linear layer, CRF for prediction to capture dependencies in outputs



## Tagging All Resolutions

- 1.5 million paragraphs
- 8 million entities
- Mostly persons and attributions

| Entity type          | # distinct | # total   |
|----------------------|------------|-----------|
| Person               | 1,159,672  | 1,929,235 |
| Person attribution   | 1,176,039  | 1,743,086 |
| Organisation         | 287,022    | 743,860   |
| Committee            | 70,518     | 135,198   |
| Location             | 617,542    | 2,551,180 |
| Date                 | 411,477    | 873,202   |
| Resolution reference | 28,990     | 189,865   |
| Other names          | ~          | ~         |
| Total                | 3,751,260  | 8,165,626 |



of Aagged entities

& Jolentification of entities to recognise \* Resolution of entity descriptions to recognised entities



thousands! very many! & Jolentification of entities to recognise millions! \* Resolution of entity descriptions to recognized entities

Where and how to apply manual in put?

Sources of variation \* multiple names / lack of formulaicity \* political developments / evolution through time \* inconsistent spelling / abbreviations \* fext recognition errors Sources of cambiguity \* multiplicity of names / descriptions \* intra-sentual references ", the matter in question " ", aforementioned town"

Sources of variation \* multiple names / lack of formulaicity manual \* political developments / evolution through time intervention inconsistent spelling / abbreviations of fuzzy matching,
fext recognition errors of automatic re-writing Sources of ambiguity \* multiplicity of names / descriptions may be resolved \* intra-textual references by examining content ", the matter in question" ", aforementioned town"

curation: Jata

Fraditional model

NER

Sources of variation \* multiple names / lack of formulaicity manual + political developments / evolution through time intervention \* inconsistent spelling / abbreviations fussy matching, \* fext recognition errors automatic re-writing Sources of ambiguity \* multiplicity of names / descriptions may be resolved \* intra-sentual references by examining content " the matter in question"

Entity india Data Manual Damed cleanap candida uration descriptions.

curation: Jata Sources of variation \* multiple names / lack of formulaicity Fraditional model manual \* political developments / evolution through time intervention \* inconsistent spelling / abbreviations fussy matching , + text recognition errors automatic re-writing Sources of ambiguity A AT nocic \* multiplicity of names / descriptions may be resolved \* intra-sentual references by examining content " the matter in question" ner tagger algorithms, ... endless spreadsheets. snakes? Entity Data Manual inso lame cleanap descriptions candidates uration

Jata curation: Sources of variation \* multiple names / lack of formulaicity Fraditional model manual \* political developments / evolution through time intervention \* inconsistent spelling / abbreviations fussy matching , \* text recognition errors automatic re-writing Sources of ambiguity A AT DO \* multiplicity of names / descriptions may be resolved \* intra-sentual references by examining ground shuth " the matter in question " NER content 0 - prepared by volunteers tagger algorithms, endle spreadsheet snakes? Entity Data costide Manual named cleanap uration descriptions candidates Entro plain domain old-fashioned specific computer knowledge programming

Jata curation: Sources of variation \* multiple names / lack of formulaicity Fraditional model manual \* political developments / evolution through time intervention \* inconsistent spelling / abbreviations fussy matching, \* fext recognition errors automatic re-writing Sources of ambiguity \* multiplicity of names / descriptions may be resolved \* intra-sentual references by examining Auth ground " the matter in question" content C NER prepared volunteers tagge be cannot spreadshee snakes? Ensite Data Entidy Manual lame cleanap candidates uration descriptions decisions not 400 much & plain wak documente domain old-fashioned specific computer knowledge programming

Data curation: Data Sources of variation curation: . multiple names / lack of familiainty Fraditional model monual + political developments / evolution through dire J intervention + inconsistant spelling / abbreviations + fout reception errors ) furry makhing, antomatic re-saifing Sources of ambiguity ntegrated model multiplicity of names / classiphions
 intra-devolution / references 4 AT nos may be resolved ground thath prepared by volunteers . He nother is question" la samerica content 0 feat low NER agger cannot be repealed ... endless spreadsheets ... - hamed -Entity Data Entite Manual descriptions cleanap uration candidates Entities decisions not too much week domain - documented specific knowledge old-fashiored computer programming NER notice tions 0 ground

Data curation: Sources of variation Data curation: . multiple names / lack of formulainty Fraditional model monuel + political developments / evolution through time of intervention inconsistent spelling / abbreviations
/pet recepction errors furry makking , antomatic re-sailies Sources of ambiguity Integrated model multiplicity of normes / classiphions
 intra-decided references 4 AT mast may be resolved ground truth by examining content . the notion is question fent fent NER volunteers tagger cannot be repeated ... endless spreadsheets. - hamed -Data Entity Entidy Manual cleanap unation descriptions candidates Entities -Visual examination decisions not logical week documented domain specific knowledge programming miteria automated clescription groups NER sortina descriptions ground fruth

Data curation: Sources of variation Data unation: . malkiple names / lack of formulainty Fraditional model manual + political developments / evolution Moneyk dine & intervention + inconsident spatting / abbreviations + spat recepcition errors fuzzy mathing, automatic re-soliting Sources of ambiguity A AI noo multiplicity of names / chuniphing
 intra-doctual references Integrated model may be revolved ground buth by examining contract . the notion is queeks tent fent NER volunteers tagger cannot be repeated ... endless spreadsheets. 11111 Data Entity Entidy > hamed -Manual cleanap unation descriptions candidates Entities -Visual examination decisions not logical week documented domain specific knowledge iomputer progenming criferia automased description NER sortina descriptions groups tagaen cut-off ground Jamed Auth

Data curation: Data Sources of variation curation: + multiple somes / lack of familiants Fraditional model monuel + political developments / soulution through dine J intercomption + inconsistant gralling / abbraviations + post reception errors furry mathing, automatic re-soliting Sources of ambiguity Integrated model multiplicity of normer / charriptions
 intra-tember refuserces \* PI magic may be resolved ground thuth by examining Ale nother is question tent fent -NER prepared by volunteers tagger cannot be repeated no data ... endless spreadsheets. 111111 Data Entity Entidy > hamed -Manual Joins cleanap unation candidates descriptions. Entities -Visual decisions not week domain documented logical old-lashiored specific knowledge examination progenming criferia development automated path cycle Entit automased description NER sorting descriptions groups tagaen cut-off content-sensitive! self-documenting! ameet ground moderately fest!

Example : Admiralities of the Republic

Once gorded in the box admirality ", divide again on the following keywords :

Dohkum Flarlingen Vrieslandt Rafferdan Op de Mase Amst [a-Z] \*

Example : Admiralities of the Republic

Once jorted in the bon , Admirality ", divide again on the following keywords:

Dokkum Harlingen Zadmirally of Frisia Possible endities are known Vireslandt in advance. Rafferdam 7 admirally of Rosserdam Op de Mare J Amst [a-Z] \* - Udmiralby of Umsterdam > Nanow semantic domain enables improvements over OCR/HTR

sorting local governments Example :

Borgermien ende Regierders der Stadt Umstendam

Syndicus ende Raeds the von Geneve

hooch Baillu ende Schepenen von Gent

Regensen van Jelmont, Quartiere van Peellands, Meyerge van 's Herbogenbosch

sorting local governments Example :

Borgermien ende Regierders der Stadt amsterdam

Syndicus ende Raedt ten van Geneve

hooch Baillu ende Schepenen von Gent

Regensen van Jelmont, Quarfiere van Peellands, Meyerge van 's Glerbogenbosch

Example : sorting local governments

Bagermen ende Regierders der Stadt Umsterdam - Amsterdam Syndicus ende Raedt the von Geneve -3 Jyndicus .... 15 hooch Bailler ende Schepenen von Gent - Gend Regenten van Helmont, Quartière van Peellandt, }-> Helmond Meyerge van 's Herbogenbosch }-> Helmond - Organisations e athibutions .

## Conclusions



- Domain-specific NER tagging and training
  - Attributions are difficult
  - Ambiguity in instructions and in entities themselves

#### • Curation

- Identification and resolution of entity references by successive grouping
- (Logical) criteria for grouping form a separate dataset
- Nesting of entity types and partial overlap can be powerful tools
- Analysis
  - Decomposing complex entities allows for combining multiple dimensions of analysis
- Project results will be published by December 2024!

## Thank You!



We also thank the volunteers for their contributions to this project!

Questions?