Who mentions whom? Recognizing political actors in proceedings

Kerkvliet, L.; Kamps, J.; Marx, M.

Publication date
2020

Document Version
Final published version

Published in
Creating, Using and Linking of Parliamentary Corpora with Other Types of Political Discourse (ParlaCLARIN II)

License
CC BY-NC

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)
Who Mentions Whom? Recognizing Political Actors in Proceedings

Lennart Kerkvliet, Jaap Kamps, Maarten Marx
Universiteit van Amsterdam
lennart.kerkvliet@student.uva.nl, {kamps, maartenmarx}@uva.nl

Abstract

We show that it is straightforward to train a state of the art named entity tagger (spaCy) to recognize political actors in Dutch parliamentary proceedings with high accuracy. The tagger was trained on 3.4K manually labeled examples, which were created in a modest 2.5 days work. This resource is made available on github. Besides proper nouns of persons and political parties, the tagger can recognize quite complex definite descriptions referring to cabinet ministers, ministries, and parliamentary committees. We also provide a demo search engine which employs the tagged entities in its SERP and result summaries.

Introduction

Parliamentary proceedings containing edited verbatim transcripts of debates can be seen as semistructured documents which convey who said what (to whom) and in which capacity. Much of the work around ParlaClarin has been devoted to 1) extracting this structure from un- or partially structured textual formats like PDF, HTML or Word, and 2) defining an XML schema in which this structure can be encoded in an efficient manner preferably fitting every possible proceedings format. Several projects have shown that this extraction process is feasible for large diachronic corpora (Beelen et al., 2017), (Palmirani and Vitali, 2011) (Marx et al., 2010) (Priser et al., 2018) (Blätte and Blessing, 2018). The main reason that this can be done is that parliamentary proceedings are predominantly based on the Hansard model, and tend to change very little during the years.

So the problem of who says what has been solved and we can ask complex information retrieval queries like when did members from party X start speaking about immigration? But we cannot yet ask the similar structured query which asks for speeches about immigration and mentioning a member of party X. This query is difficult because of two reasons: 1) member of party X refers to a set of persons which changes over time, and 2) these persons can be named in a number of different ways. The first problem can be solved using a parliamentary database indicating who was a member of which party in which period. The second is harder, especially in politics in which speakers can be rather creative in naming fellow MPs.

The problem of “who mentions whom” can be solved using named entity recognition techniques (Cardie and Wilkerson, 2008) (Lample et al., 2016) and then these recognized entities can be used in search systems following techniques from entity oriented information retrieval (Balog, 2018). In this paper we show that this process is feasible with off-the-shelf NLP and IR technology with a modest investment in creating training material. Concretely, the research described here is done to answer the following two questions:

1. How effective and accurate are off-the-shelf named entity recognition techniques when applied to parliamentary proceedings?
2. How easily can named entity annotations in parliamentary proceedings be integrated in a parliamentary search engine (both indexing and SERP), and how useful is this extra layer of meta data?

Main findings

An out of the box NER engine like spaCy trained on a newspaper corpus performs very poorly on Dutch proceedings. However, creating a training corpus which generalizes well and leads to quite acceptable recognition scores is feasible in a few days work and can be done by non experts.

The recognized entities are not just named entities but also the more interesting (especially in proceedings) definite descriptions referring to political actors like ministers of X or committees on Y, which are annotated as persons and organizations, respectively.

Next to political actors we also showed that recognizing other entities like monetary amounts and laws is feasible.

Incorporating named entity information in an existing Elastic Search search engine is straightforward both for indexing and improving the search engine result page (SERP). Easily implemented features include faceted search using entities, highlighting of entities in search snippets, entity histograms (“entity clouds”), and diachronic histograms of entities (like Google ngram viewer).

A panel of 3 academic parliamentary historians rated the extra search engine functionality based on the extracted political actors on average with a 7.5 on a scale between 1 and 10.

Resources

The manually annotated set of sentences used to train the NER tagger is available at https://github.com/maartenmarx/DutchParlNer. A search engine for Dutch parliamentary proceedings which employs the tagged entities is located at http://ner.politicalmashup.nl
Related work

Named entity recognition (NER) is a key task for many information processing algorithms like question answering and relation extraction, and has been studied in several evaluation platforms like CoNLL and MUC. We refer to (Nadeau and Sekine, 2007) for an older survey covering the traditional techniques (rule based, learned classifiers, conditional random fields) and to (Lample et al., 2016) for a survey covering the newer neural approaches based on transfer learning and word embeddings.

A common problem with NER taggers is that they perform often very well on the type of data on which they are trained, but perform (very) poorly on data from another domain. This has been observed for the special domain of parliamentary proceedings and led to several papers describing special approaches, which we will reference here. (Grover et al., 2008) describes a rule based system developed for OCRed historical documents which is tested on UK Hansard proceedings from the period 1814-1817. (Bick, 2004) contains a rule based system developed for modern Danish proceedings, and (Bojars et al., 2019) for modern Latvian proceedings. The latter use NER tagging and linking to Wikipedia to turn the proceedings into a linked data graph. NER tagging with the Stanford parser has been applied to German (Faruqui et al., 2010) and Slovenian (Pančur and Šorn, 2016) proceedings. A semantic web information retrieval system for proceedings of theuropean parliament built on top of MongoDB is presented in (Onyimadu et al., 2012). A combination of entity recognition and linking (to Wikipedia) on Dutch proceedings is presented in (Olieman et al., 2015) which builds a system on top of DBpedia Spotlight and the Dutch NER tagger FROG.

Method

Data set

We created a hand labeled dataset consisting of 5,536 sentences taken from Dutch plenary and committee proceedings from 2018-2019. These sentences consisted of 86,206 tokens and contained 3,579 named entities, often consisting of multiple tokens. These sentences are available in the spaCy train format in the github repository belonging to this paper. Table lists the number of manually labeled entities per class.

Creating the data set

The data set was created using active learning. This greatly sped up the annotation process which took in total 20 hours for one person. To bootstrap the process we used a list of current MPs and cabinet members and a list of ministries and parliamentary committees, turned this into a regular expression and matched that on the sentences in the corpus. We then trained spaCy with these automatically annotated examples and then manually checked and corrected the output. After correcting a few hundred sentences, the model was trained again with the corrected (now hand-labeled) examples, and this process was repeated a number of times until all sentences were hand-labeled. This way of working saves time because acknowledging that an example is correct or not can be done with one click, in contrast to marking entities in a text manually. In addition, the annotator perceived that the NER tagger improved after each round, and that this made his work more enjoyable and rewarding. All location names (e.g. cities, countries, mountains etc.) were considered Geo-political entities (GPE class) irrespective of the context they appeared in. The annotations were done by one annotator.

Named entity recognition with spaCy

spaCy features an extremely fast statistical entity recognition system, that assigns labels to contiguous spans of tokens. The default model identifies a variety of named and numeric entities, including persons, companies, locations, organizations and products.

For Dutch, a single statistical model (nl_core_news_sm) is available, trained on the Lassy corpus (van Noord et al., 2013). We used this as our baseline model. We then retrained this model with the additional training data described in section .

Political Actor Centric SERP

We created a search engine for 10 years of Dutch plenary and committee proceedings using Elasticsearch with a standard Search Engine Result Page (SERP) based on recommendations in (Hearst, 2009). Using the recent mapper annotated text plugin in Elasticsearch it is easy to index and use tagged ngrams. Elasticsearch indexes the additional list of “tags” as occurring at exactly the same position as the original string in the text and so these tags can be used in several tasks like highlighting, faceted search, and complex “phrase” queries combining tags and keywords (e.g. entity oriented search like “Europe Cabinet_Member”).

We added three extra features based on the tagged political actors (Balog, 2018):

1. Highlighting of political actors in result snippets. See Figure
2. Diachronic histograms of number of mentions per year per political actor in the returned hits given a query. See Figure
3. “Word” clouds containing the top 30 most mentioned political actors in the returned hits given a query.

In the interface we used consistent color coding to indicate the class of the entity, e.g., always blue for persons, red for organizations and yellow for locations.
Table 1: Number of manually annotated entities per entity type in our data set.

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSON</td>
<td>People, including fictional</td>
<td>1163</td>
</tr>
<tr>
<td>NORP</td>
<td>Nationalities or religious or political groups</td>
<td>149</td>
</tr>
<tr>
<td>FAC</td>
<td>Buildings, airports, highways, bridges, etc</td>
<td>11</td>
</tr>
<tr>
<td>ORG</td>
<td>Companies, agencies, institutions, etc</td>
<td>954</td>
</tr>
<tr>
<td>GPE</td>
<td>Countries, cities, states</td>
<td>395</td>
</tr>
<tr>
<td>EVENT</td>
<td>Named hurricanes, battles, wars, sports events, etc</td>
<td>31</td>
</tr>
<tr>
<td>LAW</td>
<td>Named documents made into laws</td>
<td>76</td>
</tr>
<tr>
<td>DATE</td>
<td>Absolute or relative dates or periods</td>
<td>261</td>
</tr>
<tr>
<td>PERCENT</td>
<td>Percentage, including &quot;%&quot;</td>
<td>59</td>
</tr>
<tr>
<td>MONEY</td>
<td>Monetary values, including unit</td>
<td>75</td>
</tr>
<tr>
<td>ORDINAL</td>
<td>“first”, “second”, etc</td>
<td>142</td>
</tr>
<tr>
<td>CARDINAL</td>
<td>Numerals that do not fall under another type</td>
<td>262</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>**</td>
<td><strong>3579</strong></td>
</tr>
</tbody>
</table>

Figure 1: Result snippet with highlighting of political actors.

El Yassini
VVD

Voorzitter **PERSON**. Als ik het goed begrijp verzoekt een D66-Kamerlid om een debat over een uitspraak van een oud-fractievoorzitter **PERSON** van **ORG** van D66,**ORG** die de minister van Onderwijs **PERSON** van D66,**ORG** op de vingers tikt. Voor dat D66-feestje ga ik niet liggen, dus: steun.

Results

Quantitative evaluation

We compare the precision (P), recall (R) and the F1 score obtained by the trained NER tagger to the out of the box spaCy baseline (trained using the *nl_core_web_sm* model, currently the only model available for Dutch). We use the following evaluation setup: we created a random 80% train, 20% test split; trained the model on the train set and computed P, R and F1 scores on the test set. We repeated this 8 times and compute for each NER class the average score over these 8 trials. The score for the baseline was computed in the same manner. The evaluation uses the strict method: an annotation of an NE is correct if it exactly coincides with the gold standard annotation. In other words, if both annotations have the same starting B token, and the same sequence of following I tokens.

Table 2 contains the scores for each NER class for the baseline and the trained model. The micro averaged F1 score increases from 49.2 to 90.1. All increases, except for the classes DATE and PERCENT are significant. Concentrating on the political actor classes PERSON and ORG, we see that the untrained baseline model is cautious with a low recall and a bit higher precision, while recall and precision are on par for the trained model. The often quite complex organization names are harder to recognize than the person.

Table 2: Micro averaged scores for the base-line and the trained model.

<table>
<thead>
<tr>
<th>Class</th>
<th>F1 Score Baseline</th>
<th>F1 Score Trained</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSON</td>
<td>49.2</td>
<td>90.1</td>
</tr>
<tr>
<td>ORG</td>
<td>50.0</td>
<td>89.5</td>
</tr>
<tr>
<td>TOTAL</td>
<td>50.0</td>
<td>89.5</td>
</tr>
</tbody>
</table>

Qualitative evaluation

A panel of 3 academic parliamentary historians were asked to rate the three extra political actor centered features, which were shown in section : highlighting of entities in result snippets, entity word clouds, and entity time lines. They did this during a real task on a search engine filled with 10 years of Dutch plenary and committee proceedings.

They rated the extra search engine functionality based on the extracted political actors on average with a 7.5 on a scale between 1 and 10. The raw scores were (8, 8, 7) for the entity highlighting, (7, 8, 8) for the entity clouds, and (9, 7, 8) for the entity time lines.

Conclusions

We have shown that an off-the-shelf named entity recognizer trained with an easy to obtain set of examples performs rather well on parliamentary proceedings. Of interest is that it is able to learn to detect complex definite descriptions of committees, ministries and other parliamentary bodies, usually consisting of several tokens.

Employing these tagged named entities in an existing search
Figure 2: Google ngram-viewer style display of the number of mentions per year for a number of political actors.

Table 2: Precision (P), Recall (R) and F1 scores for each NER class for the baseline and trained models.

(a) Baseline Model

<table>
<thead>
<tr>
<th>Type</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSON</td>
<td>16.1</td>
<td>6.2</td>
<td>9.0</td>
</tr>
<tr>
<td>ORG</td>
<td>58.8</td>
<td>37.1</td>
<td>45.5</td>
</tr>
<tr>
<td>MONEY</td>
<td>17.9</td>
<td>4.8</td>
<td>7.5</td>
</tr>
<tr>
<td>LAW</td>
<td>50.0</td>
<td>4.3</td>
<td>7.8</td>
</tr>
<tr>
<td>GPE</td>
<td>60.4</td>
<td>87.5</td>
<td>71.4</td>
</tr>
<tr>
<td>NORP</td>
<td>50.2</td>
<td>89.7</td>
<td>64.1</td>
</tr>
<tr>
<td>DATE</td>
<td>80.3</td>
<td>86.5</td>
<td>83.2</td>
</tr>
<tr>
<td>PERCENT</td>
<td>97.3</td>
<td>96.4</td>
<td>96.8</td>
</tr>
<tr>
<td>ORDINAL</td>
<td>91.5</td>
<td>95.9</td>
<td>93.6</td>
</tr>
</tbody>
</table>

(b) Trained Model

<table>
<thead>
<tr>
<th>Type</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>PERSON</td>
<td>93.8</td>
<td>91.9</td>
<td>92.8</td>
</tr>
<tr>
<td>ORG</td>
<td>82.9</td>
<td>85.3</td>
<td>84.1</td>
</tr>
<tr>
<td>MONEY</td>
<td>95.1</td>
<td>93.0</td>
<td>93.9</td>
</tr>
<tr>
<td>LAW</td>
<td>81.7</td>
<td>76.5</td>
<td>78.6</td>
</tr>
<tr>
<td>GPE</td>
<td>92.5</td>
<td>90.5</td>
<td>91.4</td>
</tr>
<tr>
<td>NORP</td>
<td>85.0</td>
<td>85.7</td>
<td>85.2</td>
</tr>
<tr>
<td>DATE</td>
<td>89.3</td>
<td>90.9</td>
<td>90.1</td>
</tr>
<tr>
<td>PERCENT</td>
<td>97.6</td>
<td>96.4</td>
<td>97.0</td>
</tr>
<tr>
<td>ORDINAL</td>
<td>96.9</td>
<td>95.4</td>
<td>96.1</td>
</tr>
</tbody>
</table>

Future work

A nice feature of spaCy is the ease with which one can train new types of entities. It would be of interest to see whether it is possible to distinguish political actors within the classes of persons and organizations.

Previous work has shown that using transfer learning techniques with unsupervised learned embeddings like BERT or ELMO can significantly outperform state of the art NER approaches (Peng et al., 2019; Devlin et al., 2018). We expect that similar gains can be reached with proceedings data.

We have not touched upon the obvious next step which is the reconciliation of entities, that is, linking the named entity in a text to the correct unique object that it refers to. Even though proceedings seem to be often cleaned compared to the verbatim transcript (with the official names of actors like ministries replacing more colloquial mentions), there is still quite some variance in ways of referring to the same object. Also political entities like ministries or committees frequently change their name. Keeping the links from entities to unique objects up to date, correct and complete is one of the hardest things in a dynamic political information system, simply because it asks for continuous manual labor by experts.

In the Netherlands, the ministry of Justice changed its name from Veiligheid en Justitie to Justitie en Veiligheid, which costed 2M Euro (Volkskrant, 2020).
Bibliographical References


