Proceedings of ParlaCLARIN II Workshop, pages 35–39
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

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*[1]. 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; Fišer 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.[2]

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[3] 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.

[1] We put the *to whom* in between brackets because this is not in all cases obvious. In the Dutch proceedings of plenary debates it is however clear who is *interrupting* whom and complete interuptions (and answers from the MP or cabinet member who was interupted) are transcribed.

[2] For instance, in the UK, "the Honourable Member for…" followed by the name of their constituency or as either "the Honourable gentleman" or "the Honourable lady". If the MP being addressed is a member of the same party they are referred to as "my Honourable friend". See http://news.bbc.co.uk/democracylive/hi/guides/newsid_82000/82149.stm.

[3] https://spacy.io/

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 CoNNL 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 (Bojārs 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[4] 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 the european 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[5] in the github repository belonging to this paper. Table 1 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[6].

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[7] 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 1.
2. Diachronic histograms of number of mentions per year per political actor in the returned hits given a query. See Figure 2.
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.

[4] https://nlp.stanford.edu/software/lex-parser.shtml

[5] For example, `["Geert Wilders is partijleider van de PVV", [[0, 13, "PERSON"], [37, 40, "ORG"]]`

[6] https://spacy.io/usage/linguistic-features#named-entities

[7] https://www.elastic.co/blog/search-for-things-not-strings-with-the-annotated-text-plugin

Table 1: Number of manually annotated entities per entity type in our data set.

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

Figure 1: Result snippet with highlighting of political actors.

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 staring 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[8]. 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.

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

[8]We did not train the classes FAC, EVENT and CARDINAL.

Figure 2: Google ngram-viewer style display of the number of mentions per year for a number of political actors.

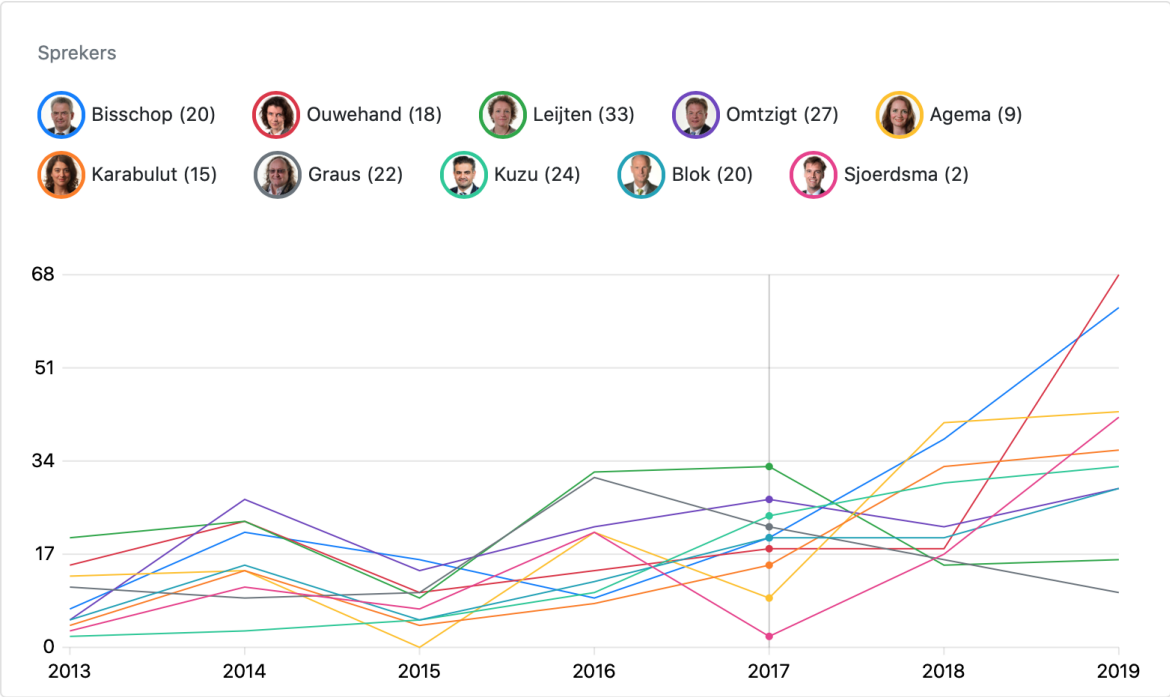

Table 2: Precision (P), Recall (R) and F1 scores for each NER class for the baseline and trained models.

(a) Baseline Model

Type	P	R	F_1
PERSON	16.1	6.2	9.0
ORG	58.8	37.1	45.5
MONEY	17.9	4.8	7.5
LAW	50.0	4.3	7.8
GPE	60.4	87.5	71.4
NORP	50.2	89.7	64.1
DATE	80.3	86.5	83.2
PERCENT	97.3	96.4	96.8
ORDINAL	91.52	95.90	93.60

(b) Trained Model

Type	P	R	F_1
PERSON	93.8	91.9	92.8
ORG	82.9	85.3	84.1
MONEY	95.1	93.0	93.9
LAW	81.7	76.5	78.6
GPE	92.5	90.5	91.4
NORP	85.0	85.7	85.2
DATE	89.3	90.9	90.1
PERCENT	97.6	96.4	97.0
ORDINAL	96.9	95.4	96.1

engine is easy and was evaluated positively by professional users.

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[9]. 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.

[9]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

Balog, K. (2018). *Entity-oriented search*. Springer.

Beelen, K., Thijm, T. A., Cochrane, C., Halvemaan, K., Hirst, G., Kimmins, M., Lijbrink, S., Marx, M., Naderi, N., Rheault, L., et al. (2017). Digitization of the Canadian parliamentary debates. *Canadian Journal of Political Science/Revue canadienne de science politique*, 50(3):849–864.

Bick, E. (2004). A named entity recognizer for danish. In *LREC*. Citeseer.

Blätte, A. and Blessing, A. (2018). The GermaParl corpus of parliamentary protocols. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Bojārs, U., Darģis, R., Lavrinovičs, U., and Paikens, P. (2019). Linkedsaeima: A linked open dataset of latvia's parliamentary debates. In *International Conference on Semantic Systems*, pages 50–56. Springer.

Cardie, C. and Wilkerson, J. (2008). Text annotation for political science research. *Journal of Information Technology & Politics*, 5(1):1–6.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Faruqui, M., Padó, S., and Sprachverarbeitung, M. (2010). Training and evaluating a german named entity recognizer with semantic generalization. In *KONVENS*, pages 129–133.

Darja Fišer, et al., editors. (2018). *Proceedings of LREC2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora.*, Paris, France, May. European Language Resources Association (ELRA).

Grover, C., Givon, S., Tobin, R., and Ball, J. (2008). Named entity recognition for digitised historical texts. In *LREC*.

Hearst, M. A. (2009). *Search User Interfaces*. Cambridge University Press, USA, 1st edition.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016). Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Marx, M., Aders, N., and Schuth, A. (2010). Digital sustainable publication of legacy parliamentary proceedings. In *Proceedings of the 11th Annual International Digital Government Research Conference on Public Administration Online: Challenges and Opportunities*, pages 99–104.

Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Olieman, A., Kamps, J., Marx, M., and Nusselder, A. (2015). A hybrid approach to domain-specific entity linking. *CoRR*, abs/1509.01865.

Onyimadu, O., Nakata, K., Wang, Y., Wilson, T., and Liu, K. (2012). Entity-based semantic search on conversational transcripts semantic. In *Joint International Semantic Technology Conference*, pages 344–349. Springer.

Palmirani, M. and Vitali, F. (2011). Akoma-Ntoso for legal documents. In *Legislative XML for the semantic Web*, pages 75–100. Springer.

Pančur, A. and Šorn, M. (2016). Smart big data: use of slovenian parliamentary papers in digital history. *Contributions to Contemporary History*, 56(3):130–146.

Peng, Y., Yan, S., and Lu, Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.

van Noord, G., Bouma, G., Van Eynde, F., de Kok, D., van der Linde, J., Schuurman, I., Sang, E. T. K., and Vandeghinste, V., (2013). *Large Scale Syntactic Annotation of Written Dutch: Lassy*, pages 147–164. Springer, Berlin, Heidelberg.

Volkskrant. (2020). Van Veiligheid en Justitie naar Justitie en Veiligheid: naamswijziging ministeries kost kabinet miljoenen. https://www.volkskrant.nl/nieuws-achtergrond/van-veiligheid-en-justitie-naar-justitie-en-veiligheid-naamswijziging-ministeries-kost-kabinet-miljoenen b88129af. Accessed: 2020-02-01.