Corpus Annotation
Co-reference for Named-Entities in Trump rallies speech

Yixuan Wu, Sooyeon Cho, Mohamed Ouji

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Contents

- Overview of the project
- Workflow
- Results of the project
- Issues
- Perspectives

Overview of the project

- Goal
  - Develop a process and create an annotated corpus for the analysis of characters of a certain figure presented in speeches.
  - Go through every step of annotation project introduced in this course (data preparation, automatic annotation, manual annotation, curation, agreement)
- Dataset
  - Donald Trump’s speeches
  - 35 rally speeches given by Trump from 2019 to 2020

Link to dataset: https://www.kaggle.com/christianlillelund/donald-trumps-rallies

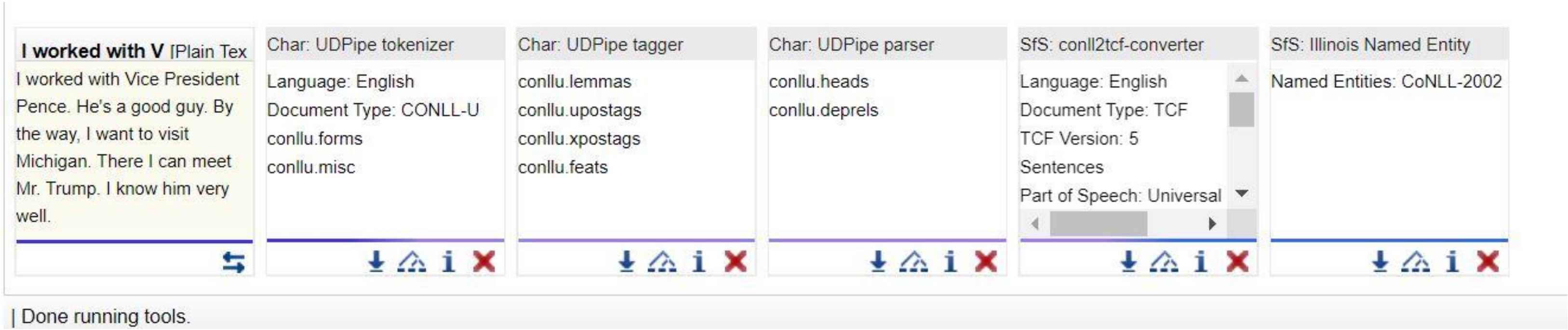
Workflow

Annotation Layers

1. Lemma
2. Part of speech
3. Named entity
4. Coreference

- W1: Tried different tools like Spacy, Stanza, Corenlp, Webanno ‘automation project’
- W2: Automatic annotation of 35 speeches using **Weblicht**
  Chose annotation guidelines (MUC7, UD Tagset, Schäfer(2012))
  Started working with **Mykonos WebAnno**
- W3 & W4: Manual correction and annotation of lemma, POS, named entity and coreference

Workflow



I worked with V [Plain Tex

I worked with Vice President Pence. He's a good guy. By the way, I want to visit Michigan. There I can meet Mr. Trump. I know him very well.

Char: UDPipe tokenizer

Language: English
Document Type: CONLL-U
conllu.forms
conllu.misc

Char: UDPipe tagger

conllu.lemmas
conllu.upostags
conllu.xpostags
conllu.feats

Char: UDPipe parser

conllu.heads
conllu.deprels

SfS: conll2tcf-converter

Language: English
Document Type: TCF
TCF Version: 5
Sentences
Part of Speech: Universal

SfS: Illinois Named Entity

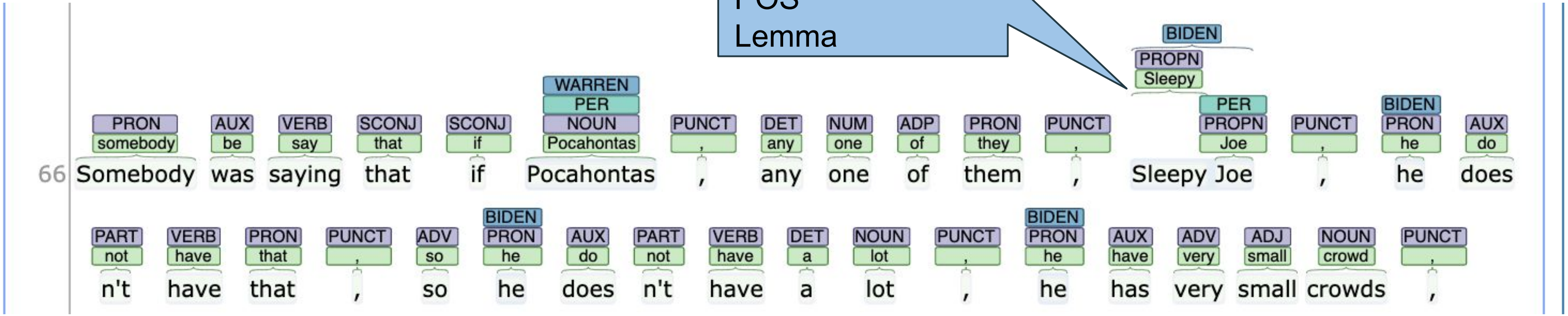
Named Entities: CoNLL-2002

| Done running tools.

WebLicht chain for automatic annotation

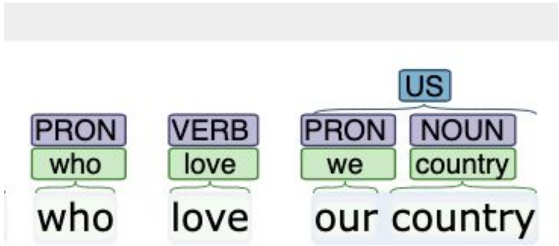
Achievements

4 Layers:
Coreference
Named Entity
POS
Lemma



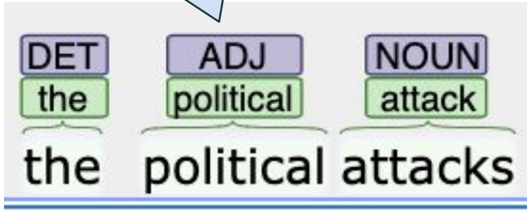
66 Somebody was saying that if Pocahontas , any one of them , Sleepy Joe , he does n't have that , so he does n't have a lot , he has very small crowds ,

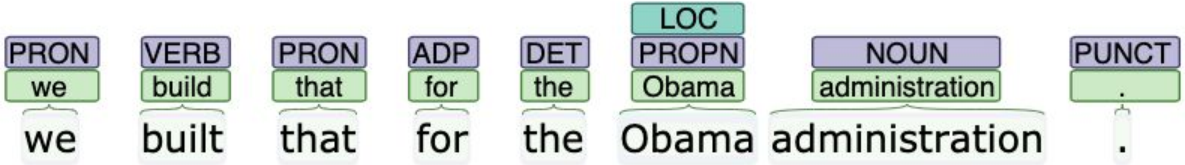
Achievements



Auto POS Error:
Our -> possessive determiner -> DET✓

With POS as ADJ, “political” remains intact

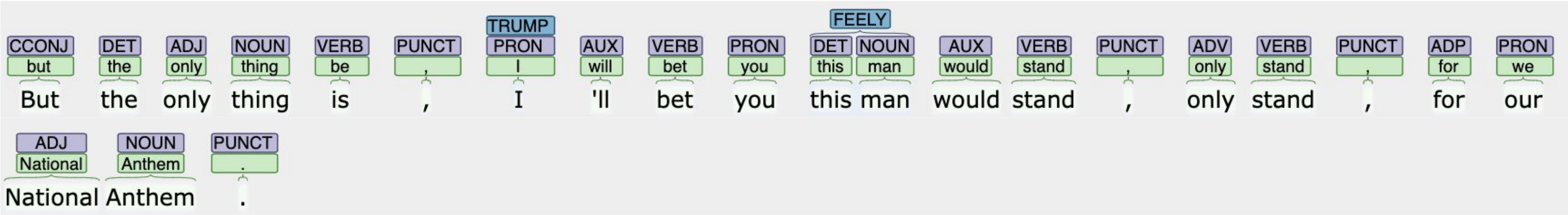




Auto NER error: “Obama” -> PER

Achievements

With manual coreference, all mentioned related entities are linked together...

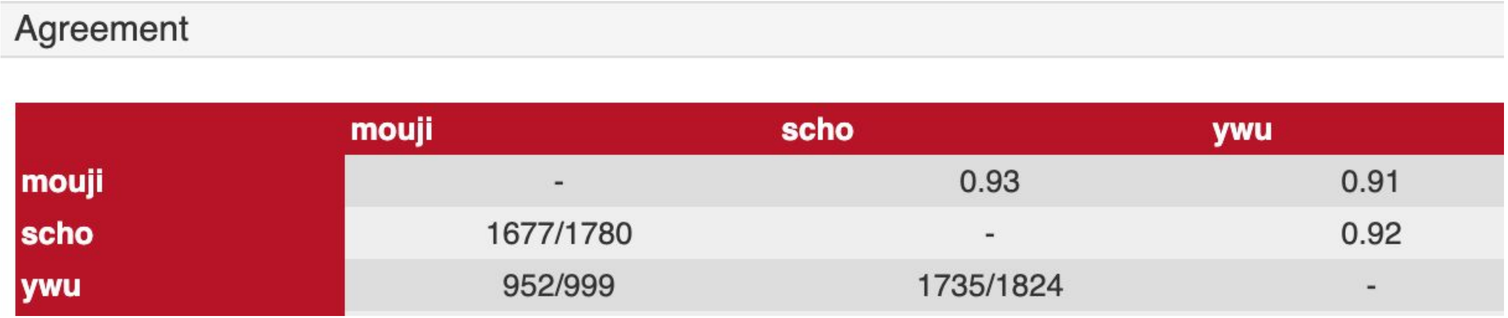


But the only thing is , I 'll bet you this man would stand , only stand , for our National Anthem .

Quantity & Agreements

- Apply automatic annotation for 35 docs (**Layers: Lemma/POS/NER**)
- We selected 3 docs (with sentence number about 1500)
- We manually corrected/annotated all **Lemma/POS/NER/Coref** for 1 doc (TexasSep23_2019.txt)
- We manually corrected/annotated of **NER/Coref** for 2 docs (YumaAug18_2020.txt, CharlotteMar2_2020.txt)

Quantity & Agreements



Agreement

	mouji	scho	ywu
mouji	-	0.93	0.91
scho	1677/1780	-	0.92
ywu	952/999	1735/1824	-

An example of NER agreement from WebAnno

Statistics

Coref:

Modi: a great man / a great leader

…

Joe Biden: Sleepy Joe / Joe …

Hillary Clinton: crooked Hillary...

ADVs:
(YumaAug18_2020.txt)

...

20 back
21 now
26 very
32 never
39 so

NOUNs:
(YumaAug18_2020.txt)

...

23 thing
29 border
32 year
36 country
37 people

Issues

- Automatic annotation of coreference resolution.
  - Hard to use for manual correction task.
- NER tag MISC elimination uncovered later useful use.
- Ambiguity with the personal pronoun “we” in the speeches.
- The depth of fine graining the coreference annotation.

Perspectives

- Manually corrected data is useful for enhancing the NLP pipeline performance on spoken data.
- The data is useful for a machine learning approach for named entity coreference resolution in spoken data.
- Named entity coreference resolution is a useful feature for authorship profiling.

Thank you!