HS Corpus Annotation WS2020/21

Project Description

Yixuan Wu, Sooyeon Cho, Mohamed Ouji

Goal:

The goal of this project is to create an annotated corpus for the analysis of characters of a certain figure presented by speeches. We decided to analyse Donald Trump's speeches because they have distinctive unique styles. The basic linguistic annotations like tokenization, lemma, and part-of-speech will be done to extract the most frequently used adjectives/nouns/verbs lemmas in the speeches. A coreference annotation will also be included in this project to find out the "Trump-styled" expression on certain people or other entities (e.g., Joe Biden as "sleepy Joe"). The sub-goal of the project is to calculate annotation agreement, thus all annotators will annotate the same corpus.

Data:

The data of this project is based on an existing well-made corpus (https://www.kaggle.com/christianlillelund/donald-trumps-rallies) called Donald Trump's Rallies. In this corpus, 35 rally speeches given by Trump were documented in plain text. The time span of these speeches are from 2019 to 2020.

Revised Work Flow (Where member names not written, we did all together)

1. develop algorithm for automatic annotation
2. choose tools and set-up the project
3. automatically annotate (sentence split, tokenization) pos, lemma, and named entity for all 35 files (each member did ⅓ of the files)
4. choose 3 documents to manually correct (TexasSep23_2019.txt, YumaAug18_2020.txt, CharlotteMar2_2020.txt)
5. choose and refine some annotation guidelines (MUC7 for named entity, UD for pos)
6. manually correct and curate lemma, pos and named entity on TexasSep23_2019.txt to reach an agreement on the annotation schema (all annotators work on the same file together)
7. manually add coreference layer on the curated TexasSep23_2019.txt, then curate the coreference layer (Yixuan & Sooyeon did the annotation of coreference, Mohamed did the curation)
8. manually correct the named entity layer on YumaAug18_2020.txt, CharlotteMar2_2020.txt and curate the two files. (Yixuan & Sooyeon did named entity correction, Mohamed did named entity curation)

9. add coreference layer on YumaAug18_2020.txt, CharlotteMar2_2020.txt (Yixuan & Sooyeon did coreference correction, Mohamed did coreference curation)
10. extract the coreference list and most frequently used 10 lemmas for verb, noun and adjectives from the three curated documents using the command line interface.

Some issues in this project:

1. With the limited time, we can only fully annotate 1 file (TexasSep23_2019.txt) and partially annotate 2 files (YumaAug18_2020.txt, CharlotteMar2_2020.txt)
2. We removed the MISC tag in the named entity layer from the 3 annotated files because we are mainly interested on PER, LOC, and ORG entities and the coreference based on these three entity tags. If we had enough more time, it would be good to also include the MISC tag to make the corpus more fine grained. (ex. We have a nice [Indian American community]. This hardworking community is ….)
3. In the coreference layer, we added and annotated predicative descriptive noun phrases. For example: "[Trump] is [a nice guy]". Here "a nice guy" is actually the property of Trump, but we annotated "a nice guy" in the coreference layer as Trump, because in our corpus data, Trump used sentence structure like "A is B" very often. In this case, we didn't want to lose too much information which we are interested in, so we modified our coreference schema so it fits better with our research goal.

Code for extracting the data:

Find out the most frequent Nouns/Adjectives/Adverbs.. etc:

awk '$(select column number) ~/NOUNs/ {print $(interested column Name)}' CURATED-FILE-NAME.tsv | sort | uniq -c | sort

Find out the coreference and linked expression:

(take coref "MODI" as an example)

awk '$10 ~/MODI/ {print $3, $10}' CURATED-FILE-NAME.tsv