RMSE_CMU Progress Report

Samarth Thopaiah ,Jacqueline Liao, Xueying Ding, Dewang Sun

(sthopaia, jfliao, xding2, dewangs)@andrew.cmu.edu

Stanza

- an open-source Python natural language processing toolkit by Stanford
  - From raw texts to annotations
  - Tokenization and sentence splitting
  - Pos tagging
  - Morphological feature tagging, lemmatizations
  - Named entity recognition
  - Multilingual: 66 languages
- NER Pipeline:
  - a forward and a backward character level LSTM language model
  - concatenate the representations at the end of each word position from both models with word embeddings
  - a standard one-layer Bi-LSTM sequence tagger with a conditional random field (CRF)-based decoder.

Example: BEIO tags

```
doc = nlp("Chris Manning teaches at
Stanford University. He lives in the Bay
Area.")
```

```
token: Chris   ner: B-PERSON
token: Manning ner: E-PERSON
token: teaches ner: O
token: at ner: O
token: Stanford     ner: B-ORG
token: University   ner: E-ORG
token: .  ner: O
token: He ner: O
token: lives   ner: O
token: in ner: O
token: the     ner: B-LOC
token: Bay     ner: I-LOC
token: Area    ner: E-LOC
token: .  ner: O
```

Spacy

- Industrialized NLP pipelines
  - Support for pretrained word vectors and embeddings
  - Linguistically-motivated tokenization
  - Named entity recognition,
  - POS tagging
  - text classification
  - lemmatization, morphological analysis
- NER
  - RoBERT a based pretrained model with transformers
  - Tok2vec
- Transformer: BILUO tags to BIO tags

Stanza Pipeline Results

- Combined sentence segmentation and tokenization model
- OntoNotes NER model

	Set 1 (football)	Set 2 (constellations)	Set 3 (languages)	Set 4 (movies)	**Total**
Sentences	3007	1333	3068	2604	**10012**
Tokens	85058	28287	71900	64714	**249959**
NERs	14025	3453	6354	7645	**31477**
PERs	3229	627	508	3247	**7611**

Spacy CPU model result

- On conll2003 test set

	Ground Truth	en_core_web_sm	en_core_web_lg	Accuracy
Tokens	46435	45193	45193	97.3%
PERs	1617	1202	1432	88.6%

- On project development set

	Set1	Set2	Set3	Set4	Total
Sentences	3620	1509	3619	2833	11581
Tokens	86802	29282	74092	66138	256314
PERs	2880	776	813	2912	7381

Transformer Pipeline with spaCy and HuggingFace

RoBERTa(Yinhan Liu et al., 2019) based pretrained model. Converted spaCy BILUO tags to BIO tags.
Test accuracy for NER on conll 2003 dataset is : 91.3%

	Set 1 (football)	Set 2 (constellations)	Set 3 (languages)	Set 4 (movies)	**Total**
Sentences	3020	1301	2945	2579	**9845**
Tokens	86864	29284	74119	66198	**256465**
NER (PER tags)	2999	199	364	3061	**6623**

Discoveries

- We notice that there are big differences of NER tags between Stanza and Spacy-CPU and Spacy-GPU for constellations and languages
- Mistagging
  - LEO(LOCATION) VS LEO(PERSON)
  - M95(LOCATION) VS M95(PERSON)
- Tuning the Stanza and Spacy-CPU models with labelled data
- Apply the Spacy-GPU to the next stage

Future Plans

- NER tags used for Questions Generation and Questions Answering
  - Search of relevant passages or paragraphs, tf-idf
- QGQA pipelines:
  - Hugging Face
  - NLTK
  - CoreNLP

References:

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In Association for Computational Linguistics (ACL) System Demonstrations. 2020. [pdf][bib]

Honnibal, M. & Montani, I., 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.