Annotating the Tweebank Corpus on Named Entity Recognition and Building NLP Models for Social Media Analysis

HARVARD MEDICAL SCHOOL
MIT center for constructive communication
Hang Jiang*, Yining Hua*, Doug Beeferman, Deb Roy | {hjian42, ninghua, dougb5, dkroy} @ mit.edu

Introduction

Processing the noisy and informal language of social media is challenging for traditional NLP tools because such messages are usually short in length and irregular in spelling and structure. Liu et al. (2018) introduced a tweet-based Tweebank V2 (TB2), including tokenization, part-of speech (POS) tags, and Universal Dependencies, but there is no NER benchmark on TB2. Annotating named entities in TB2 allows researchers to not only train multi-task learning models but also study linguistic relationship between named entities and syntactic labels.

Contributions

- Create the *Tweebank-NER* benchmark
- Train and release the *Twitter-Stanza* pipeline.
- Compare *Twitter-Stanza* against existing models, showing simple neural architecture is effective and suitable for Tweet processing.
- Train Transformer-based models to establish a strong baseline on the *Tweebank-NER* benchmark.
- Release our data, models, and code, including *Twitter-Stanza* and Hugging Face BERTweet models.

Why do we need Tweebank-NER?

- *Tweebank-NER* is still challenging for current NER models (e.g. models pre-trained on WNUT17).
- It makes TB2 a complete dataset for multi-task learning.

Annotate Named Entities in Tweebank v2.0

- Follow CoNLL 2003 guidelines
- Use Qualtrics platform + Amazon Mechanical Turk
- Two-stage annotation
  - 3 annotators annotate Tweets
  - Tweets without consensus to be re-annotated by the first two authors
- Adopt token-level pairwise F1 score (70.7) calculated without the O label

Dataset statistics

Dataset	Train	Dev	Test
Tweets	1,639	710	1,201
Tokens	24,753	11,742	19,112
Avg. token per tweet	15.1	16.6	15.9
Annotated spans	979	425	750
Annotated tokens	1,484	675	1183
Avg. token per span	1.5	1.6	1.6

Table 1: Annotated corpus statistics.

Label	Quantity	F1
PER	777	84.6
LOC	317	74.4
ORG	541	71.9
MISC	519	50.9
Overall	2,154	70.7

Table 2: Number of span annotations per entity type and Inter-annotator agreement scores in pairwise F1.

Methods for NLP Modeling

Models

- Stanza
- Hugging Face (BERTweet + Token Classification)
- spaCy, FLAIR, spaCy-transformer

Questions

- How do Stanza models perform compared with other NLP frameworks on the core Tweet NLP tasks?
- How do transformer-based models perform compared with traditional models on these tasks?

Performance on Tweebank-NER

Main findings

- The best non-transformer model: Stanza NER model (TB2+W17)
- The best transformer model: HuggingFace-BERTweet (TB2+W17)
- TB2 and WNUT17 training sets boost the performance

Training data	TB2	WNUT17	F1 Drop
spaCy	52.20	44.93	7.27↓
FLAIR	62.12	55.11	7.01↓
HgFace-BERTweet	73.71	59.43	14.28↓
spaCy-BERTweet	73.79	60.77	13.02↓
Stanza	60.14	56.40	3.74↓

Table 5: Comparison among NER models trained on TB2 vs. WNUT17 on TB2 test in entity-level F1. “Hg-Face” stands for “HuggingFace”.

Systems	F1
spaCy (TB2)	52.20
spaCy (TB2+W17)	53.89
FLAIR (TB2)	62.12
FLAIR (TB2+W17)	59.08
HuggingFace-BERTweet (TB2)	73.71
HuggingFace-BERTweet (TB2+W17)	**74.35**
spaCy-BERTweet (TB2)	73.79
spaCy-BERTweet (TB2+W17)	74.15
Stanza (TB2)	60.14
Stanza (TB2+W17)	62.53

NER

Performance on Syntactic NLP Tasks

Tokenization + Lemmatization

- Stanza (TB2) achieves the SOTA performance
- Combining TB2 + UD English-EWT hurt performance

System	F1
Twokenizer	94.6
Stanford CoreNLP	97.3
UDPipe v1.2	97.4
Twpipe	98.3
spaCy (TB2)	98.57
spaCy (TB2+EWT)	95.57
Stanza (TB2)	**98.64**
Stanza (TB2+EWT)	98.59

Tokenization

System	F1
NLTK	88.23
spaCy	85.28
Flair (TB2)	96.18
Flair (TB2+EWT)	84.54
Stanza (TB2)	**98.25**
Stanza (TB2+EWT)	85.45

Lemmatization

POS Tagging + Dependency Parsing

- **POS**: HuggingFace-BERTweet (TB2+EWT) achieves the SOTA
- **Parsing**: spaCy-XLM-RoBERTa (TB2) achieves the SOTA
- Stanza achieves competitively against non-transformer models

System	UPOS
Stanford CoreNLP	90.6
Owoputi et al. (2013) (greedy)	93.7
Owoputi et al. (2013) (CRF)	94.6
Ma and Hovy (2016)	92.5
BERTweet (Nguyen et al., 2020)	95.2
spaCy (TB2)	86.72
spaCy (TB2+EWT)	88.84
FLAIR (TB2)	87.85
FLAIR (TB2+EWT)	88.19
HuggingFace-BERTweet (TB2)	95.21
HuggingFace-BERTweet (TB2+EWT)	**95.38**
spaCy-BERTweet (TB2)	87.61
spaCy-BERTweet (TB2+EWT)	86.31
spaCy-XLM-RoBERTa (TB2)	93.90
spaCy-XLM-RoBERTa (TB2+EWT)	93.75
Stanza (TB2)	93.20
Stanza (TB2+EWT)	93.53

POS Tagging

System	UAS	LAS
Kong et al. (2014)	81.4	76.9
Dozat et al. (2017)	81.8	77.7
Ballesteros et al. (2015)	80.2	75.7
Liu et al. (2018) (Ensemble)	83.4	**79.4**
Liu et al. (2018) (Distillation)	82.1	77.9
spaCy (TB2)	66.93	58.79
spaCy (TB2 + EWT)	72.06	63.84
spaCy-BERTweet (TB2)	76.32	71.72
spaCy-BERTweet (TB2+EWT)	76.18	69.28
spaCy-XLM-RoBERTa (TB2)	**83.82**	**79.39**
spaCy-XLM-RoBERTa (TB2+EWT)	81.02	75.43
Stanza (TB2)	79.28	74.34
Stanza (TB2 + EWT)	82.10	77.60

Dependency Parsing

Future Work

Develop multi-task Tweet NLP models, and design human-in-the-loop methods to identify bad annotation and improve the quality of Tweet NLP datasets.

References

- Liu, Y., Zhu, Y., Che, W., Qin, B., Schneider, N., and Smith, N. A. (2018). Parsing tweets into universal dependencies. NAACL.
- Nguyen, D. Q., Vu, T., and Nguyen, A. T. (2020). Bertweet: A pre-trained language model for english Tweets. ACL Demo.