TweetNERD
End to End Entity Linking Benchmark for Tweets

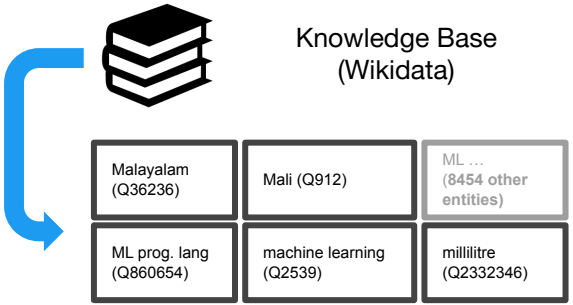
Shubhanshu Mishra, Aman Saini, Raheleh Makki, Sneha Mehta, Aria Haghighi, Ali Mollahosseini
Twitter, Inc.

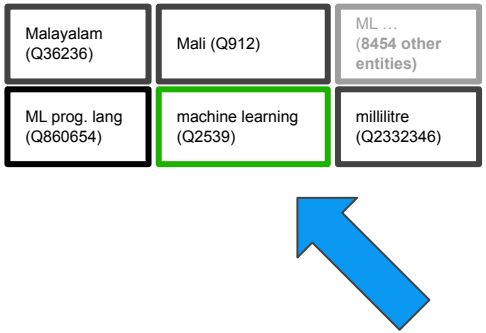
ArXiv: https://arxiv.org/abs/2210.08129
Dataset: https://doi.org/10.5281/zenodo.6617192
Code: https://github.com/twitter-research/TweetNERD

NeurIPS 2022 Datasets and Benchmarks

Named Entity Recognition and Disambiguation (NERD)

NeurIPS is the biggest ML conference. In 2022, it will be held in NOLA.





NeurIPS is the biggest **ML** conference. In 2022, it will be held in **NOLA**.

NER - Named Entity Recognition

NeurIPS is the biggest **ML** conference. In 2022, it will be held in **NOLA**.

Candidate Generation

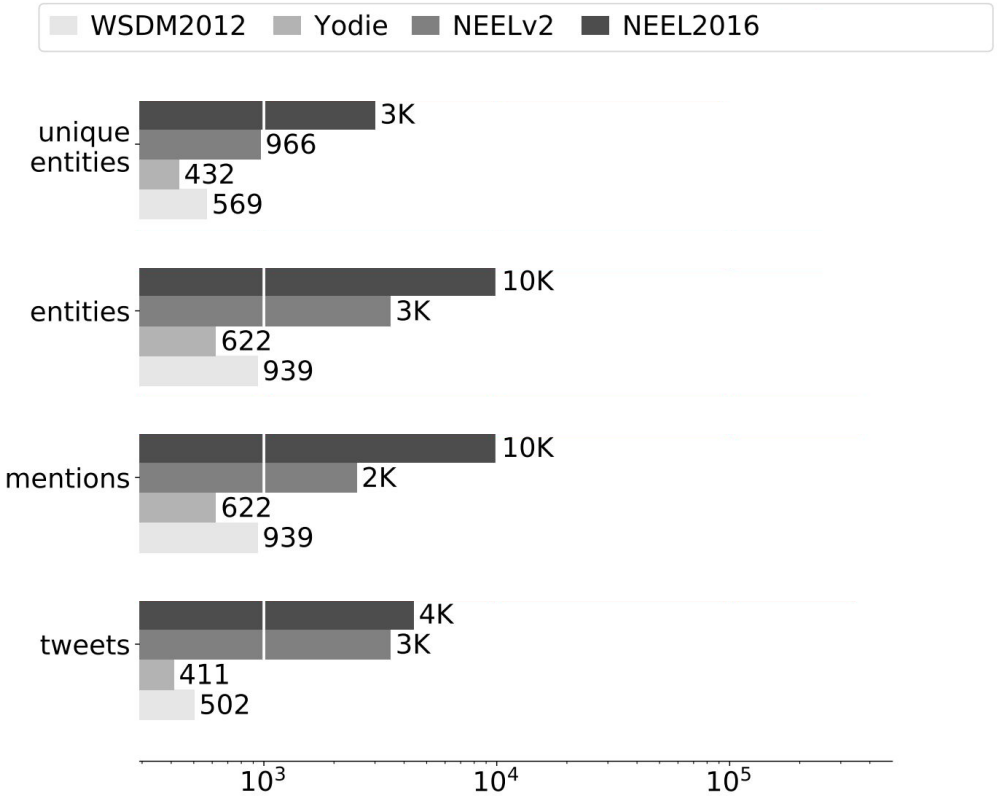
NeurIPS is the biggest **ML** conference. In 2022, it will be held in **NOLA**.

Entity Disambiguation

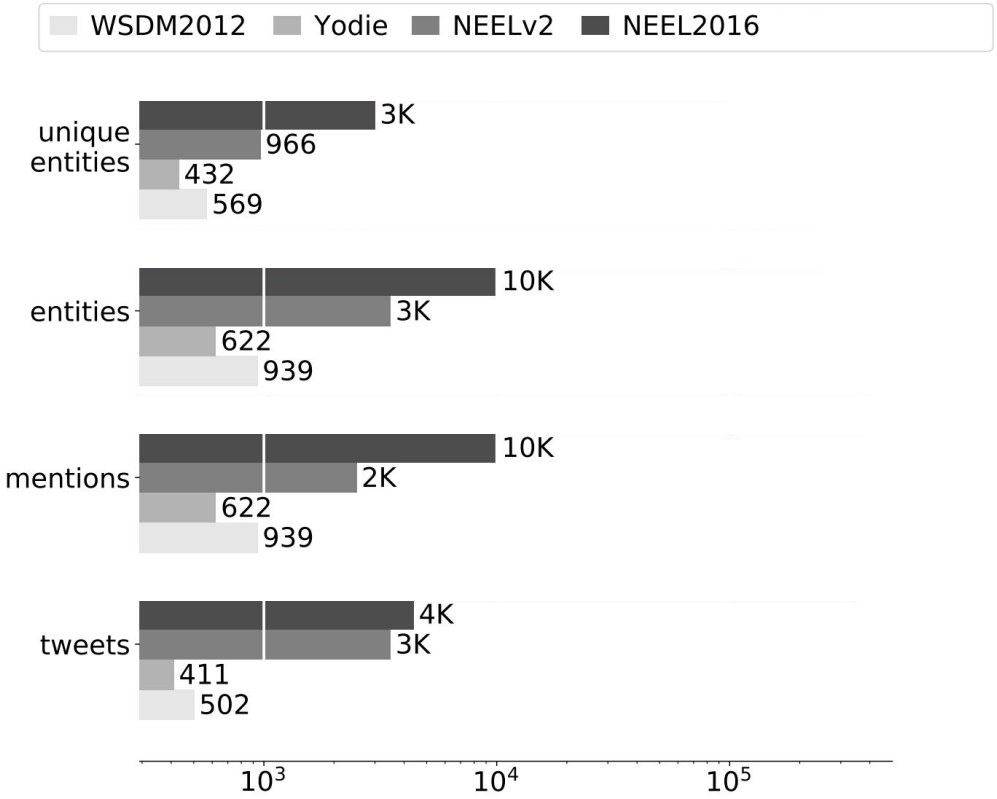
Challenges for NERD on Tweets

- Finding the correct entity could require context around the Tweet
- Users can use creative spellings
- Tweets are very short (max 280 chars)
- Small existing publicly available benchmarks for NERD on Tweets

Existing NERD datasets for Tweets



TweetNERD - Largest NERD dataset for Tweets



TweetNERD: Overview

- Available at: https://doi.org/10.5281/zenodo.6617192 under Creative Commons Attribution 4.0 International (CC BY 4.0) license.
- Scripts for processing TweetNERD can be found at: https://github.com/twitter-research/TweetNERD
- Entities linked to Wikidata via QID: Multilingual Public Knowledge Base.

Id	Start	End	Mention	Entity	Score
1	7	14	Twitter	Q918	3
2	0	5	Paris	Q90	3
3	0	4	Anil	AMB.	2

Released Data Format.
Should be hydrated via Public Twitter API.

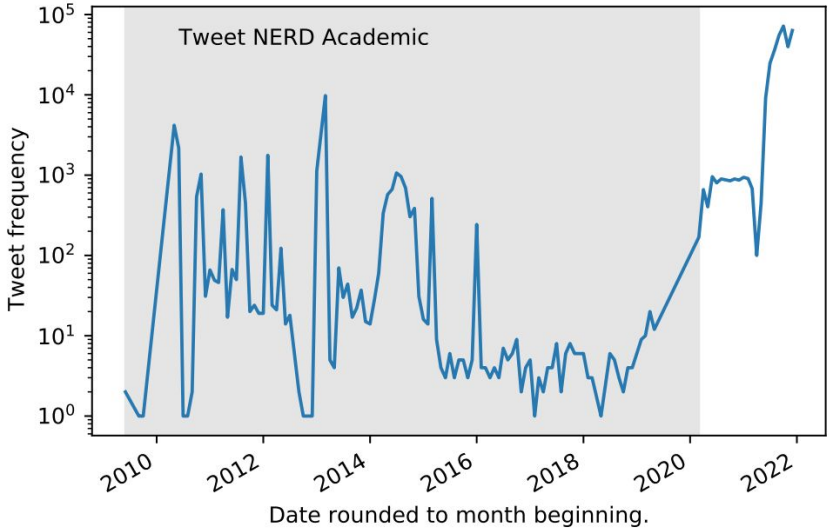
TweetNERD: Annotation

- Each Tweet annotated by 3 annotators. Majority is marked as gold.
- Annotators asked to use Tweet context, media, and time in making decision.
- Annotators select named entity span in Tweet and link it to Wikidata QID or AMBIGUOUS or NOT_FOUND.
- Entities Types: Person, Place, Organization, Products, Works of Art, Scientific Concepts
- Entities can be Hashtags, or words or phrases in Tweet.

Id=1: I love [**Twitter**]$_{[ENTITY]}$ Candidates: **Q918**, NOT FOUND, AMBIGUOUS
Id=2: [**Paris**]$_{[ENTITY]}$ is regarded as the world's fashion capital Candidates: **Q90**, Q79917, NOT FOUND, AMBIGUOUS
Id=3: [**Anil**]$_{[ENTITY]}$ is playing Candidates: NOT FOUND, **AMBIGUOUS**

TweetNERD: Data Sampling

- Random sampling gives low hit rate for Tweets with Entities.
- Sample Tweets which have NER, phrase match with Wikidata
- Use phrase entropy to select Tweets with easy and difficult phrases.
- 340K Tweets collected in rounds of 25K Tweets.
- Tweets from 2010 to 2022.



TweetNERD Test Splits: TweetNERD-OOD

25K Tweets used for evaluating existing named entity recognition and linking models.

TweetNERD-OOD is sampled in equal proportion based on the entropy of the contained NER mentions.

TweetNERD Data Splits: TweetNERD-Academic

Includes 30K Tweet IDs used in existing Tweet based NERD benchmarks, re-annotated using TweetNERD guidelines.

Many Tweets from existing benchmarks are now deleted (Found rate ~60%)

Good for temporal generalization evaluation.

dataset	Tasks	Total Tweets	Found Tweets	Found %
Tgx [Dredze et al., 2016]	CDCR	15,313	9,790	63.9
Broad [Derczynski et al., 2016]	NER	8,633	6,913	80.1
Entity Profiling [Spina et al., 2012]	NER	9,235	6,352	68.8
NEEL 2016 [Rizzo et al., 2016]	NERD	9,289	2,336	25.1
NEEL v2 [Yang and Chang, 2015]	NERD	3,503	2,089	59.6
Fang and Chang [2014]	NERD	2,419	1,662	68.7
Twitter NEED [Locke, 2009]	NERD & IR	2,501	1,549	61.9
Ark POS [Gimpel et al., 2011]	POS	2,374	1,313	55.3
WikiD	NED	1,000	504	50.4
WSDM2012 [Meij et al., 2012]	Relevance	502	415	82.7
Yodie [Gorrell et al., 2015]	NERD	411	288	70.1

Performance of existing NERD systems

Model	OOD	Academic
Spacy	0.377	0.454
StanzaNLP	0.421	0.503
SocialMediaIE	0.153	0.245
BERTweet WNUT17	0.278	0.46
TwitterNER	0.424	0.522
AllenNLP	0.454	0.552

(a) NER `strong_mention_match` F1 scores.

- Evaluate popular systems for NER, Linking given True Spans, and End to End Linking.
- Evaluations are tokenizer independent
- Performed using **neleval** software

Model	`entity_match`		`strong_all_match`	
	OOD	Academic	OOD	Academic
GENRE	0.469	0.636	0.39	0.624
REL	0.463	0.614	0.387	0.56
Lookup	0.621	0.645	0.584	0.617

(b) Entity Linking given true spans (EL) F1 scores.

Model	`entity_match`		`strong_all_match`	
	OOD	Academic	OOD	Academic
DBpedia	0.292	0.399	0.231	0.347
NLAI	0.522	0.568	0.313	0.494
TAGME	0.402	0.431	0.293	0.381
REL	0.344	0.484	0.27	0.444
GENRE[4]	0.307	0.458	0.223	0.379

(c) End to End Entity Linking (End2End) F1 scores.

Suggested applications

Suggested tasks:

- NER,
- Candidate Generation,
- Linking given Gold Spans, and
- End to End Entity Linking.

Evaluate NERD systems for Tweet using standard Train/Dev/Test split

Assess Temporal Generalization of NERD models

Thank You

Mishra, Shubhanshu, Aman, Saini, Raheleh, Makki, Sneha, Mehta, Aria, Haghighi, and Ali, Mollahosseini. "TweetNERD - End to End Entity Linking Benchmark for Tweets.". In Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. 2022.

ArXiv: https://arxiv.org/abs/2210.08129

Dataset: https://doi.org/10.5281/zenodo.6617192

Code: https://github.com/twitter-research/TweetNERD

Wikidata page: https://www.wikidata.org/wiki/Q114825474

Questions: @TheShubhanshu