Robust Candidate Generation for Entity Linking on Short Social Media Texts

(Researcher) (Actor)

Liam H is a great ML

Entity Disambiguation

User @user

Researcher!

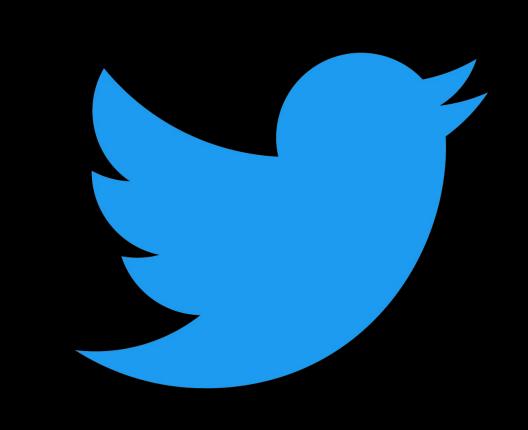
Example Tweets are not real.

7:35 PM · Oct 13, 2022

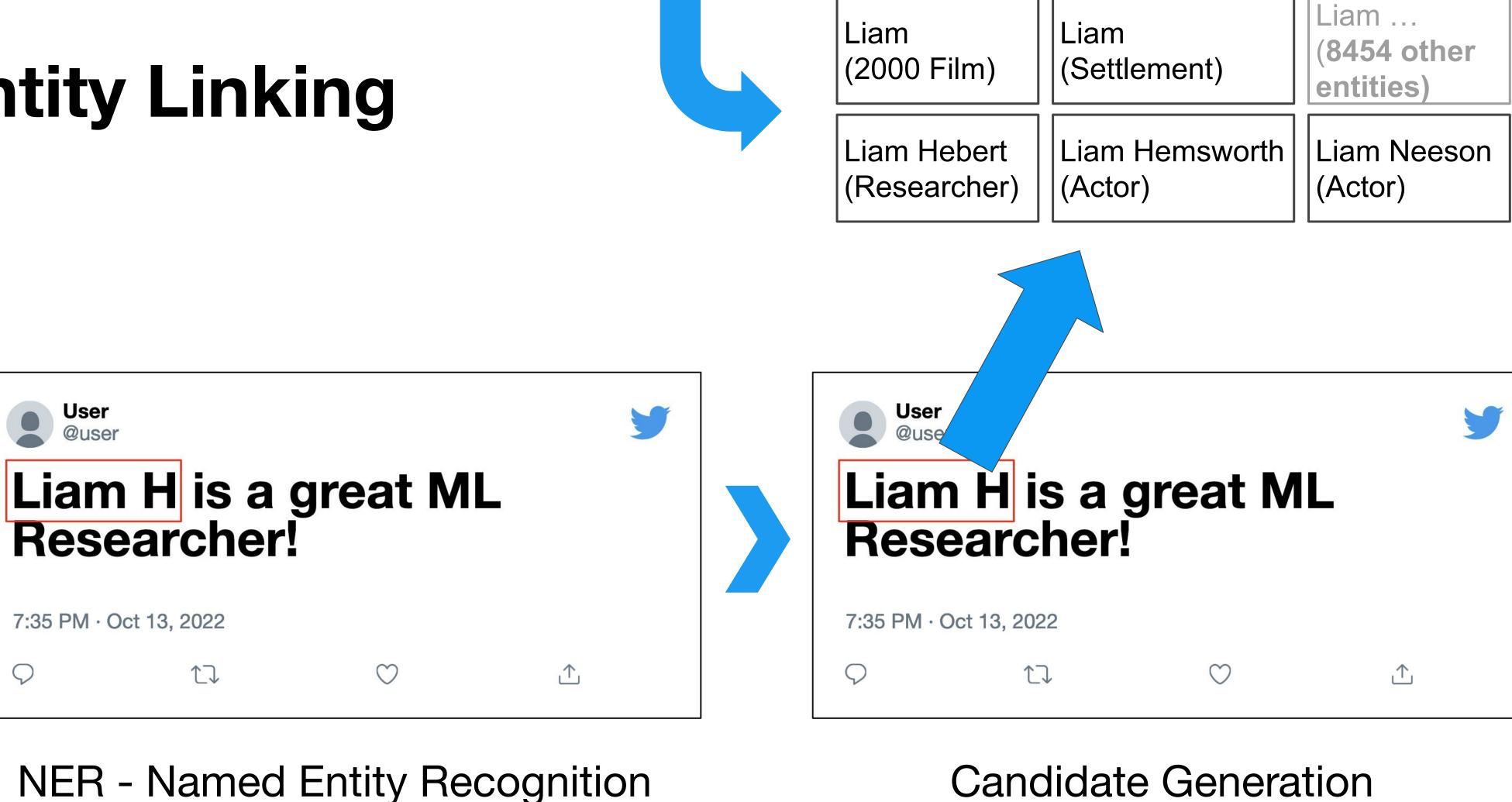
(Actor)

Liam Hebert¹*, Raheleh Makki², Shubhanshu Mishra², Hamidreza Saghir², Anusha Kamath², Yuval Merhav² ¹University of Waterloo, ²Twitter, Inc., *Work done during internship at Twitter, Inc.

2022 The 8th Workshop on Noisy User-generated Text (W-NUT)



Entity Linking



Challenges?

- Finding the correct entity could require context
- Users can have creative spelling

Recall@k Using Gold Spans

- Lookup tables have to be maintained with aliases
- Good Candidate Generation relies on accurate NER.



--- hybrid (0.96)

- lookup (0.76)

dense (0.86)

Motivation

- Can we improve candidate generation in presence of noisy NER?
- Can we scale EL without storing all possible surface forms?
- Can we use context to guide candidate generation?
- Can we utilize embeddings?



Zero-Shot Dense Retrieval

Methodology



Knowledge Base: July 2022 Wikipedia - 6.5M Entities. Filtered to remove miscellaneous pages using

Wikidata

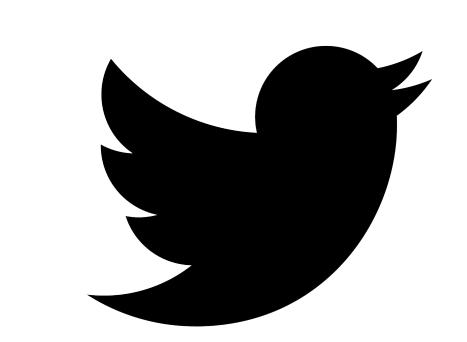


Lookup Retrieval:

Alias Table using Wikidata Aliases and Labels, ranked using probability of entity given surface form



Hybrid Retrieval: Combine candidates from both Dense and Lookup



TweetNERD Dataset[2]:

Dataset of over 340k+ Labeled Tweets. Evaluated on Academic and OOD Split



Dense Retrieval:

Pre-trained BLINK[1] Encoders, embeddings indexed using FAISS. First 4 sentences of Wikipedia and annotated spans

[1] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable Zero-shot Entity Linking with Dense Entity Retrieval. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6397–6407, Online. Association for Computational Linguistics [2] Mishra, Shubhanshu, Saini, Aman, Makki, Raheleh, Mehta,

Sneha, Haghighi, Aria, & Mollahosseini, Ali. (2022). TweetNERD

End to End Entity Linking Benchmark for Tweets (0.0.0) [Data

set]. Zenodo. https://doi.org/10.5281/zenodo.6617192

Results

Recall@16 Using Gold Spans

Data Split	Dense	Lookup	BM25	Hybrid	0.90
Academic	0.783	0.741	0.221	0.916	
OOD	0.772	0.847	0.556	0.933	∠ 0.85
Overall	0.779	0.717	0.362	0.930	0.80
					ě o 75

Recall@16 Using NER Spans

Data Split	Dense	Lookup	BM25	Hybrid
Academic	<u>0.761</u>	0.613	0.164	0.880
OOD	0.754	0.757	0.440	0.903
Overall	0.759	0.715	0.245	0.887

Note: Lookup, we retrieve all exact match candidates since they are not explicitly ranked. As a result, the performance of Lookup reflects an upper-bound of the performance of that method.

Unique Retrieved Candidates

NER Spans

Data Split	Dense	Lookup	BM25
Academic	8,362	4,711	983
OOD	1,263	2,448	1,496
Overall	9,625	7,159	2,479

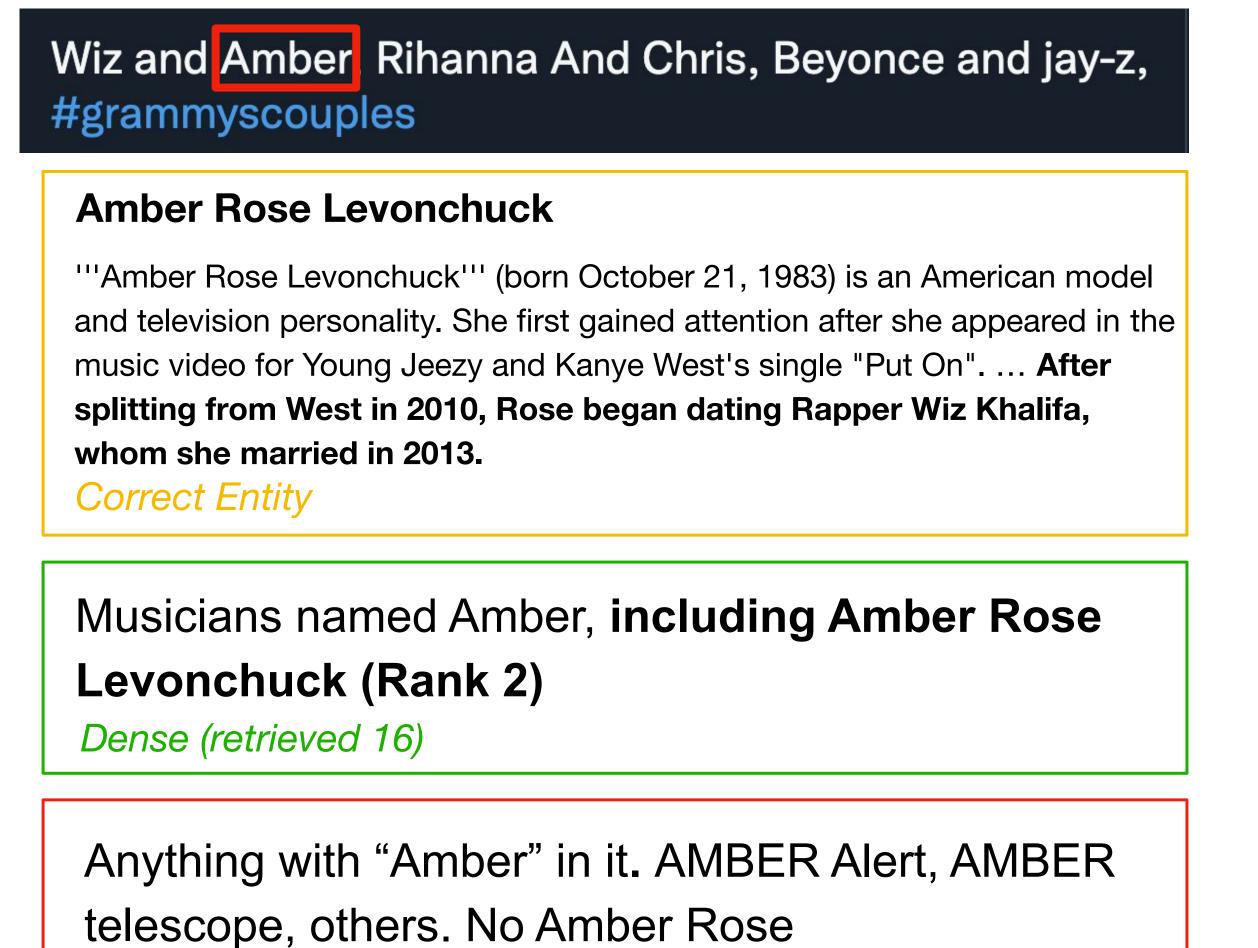
Gold Spans

Data Split	Dense	Lookup	BM25
Academic	7,719	5,268	1,043
OOD	1,055	2,664	1,495
Overall	8,774	7,932	2,538

Liam Hebert, Raheleh Makki, Shubhanshu Mishra, Hamidreza Saghir, Anusha Kamath, and Yuval Merhav. 2022. Robust Candidate Generation for Entity Linking on Short Social Media Texts. In Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022), pages 83–89, Gyeongju, Republic of Korea. Association for Computational Linguistics. https://aclanthology.org/2022.wnut-1.8/

Quantitative Analysis

Context



No one here remembers The Marine and 12 rounds.

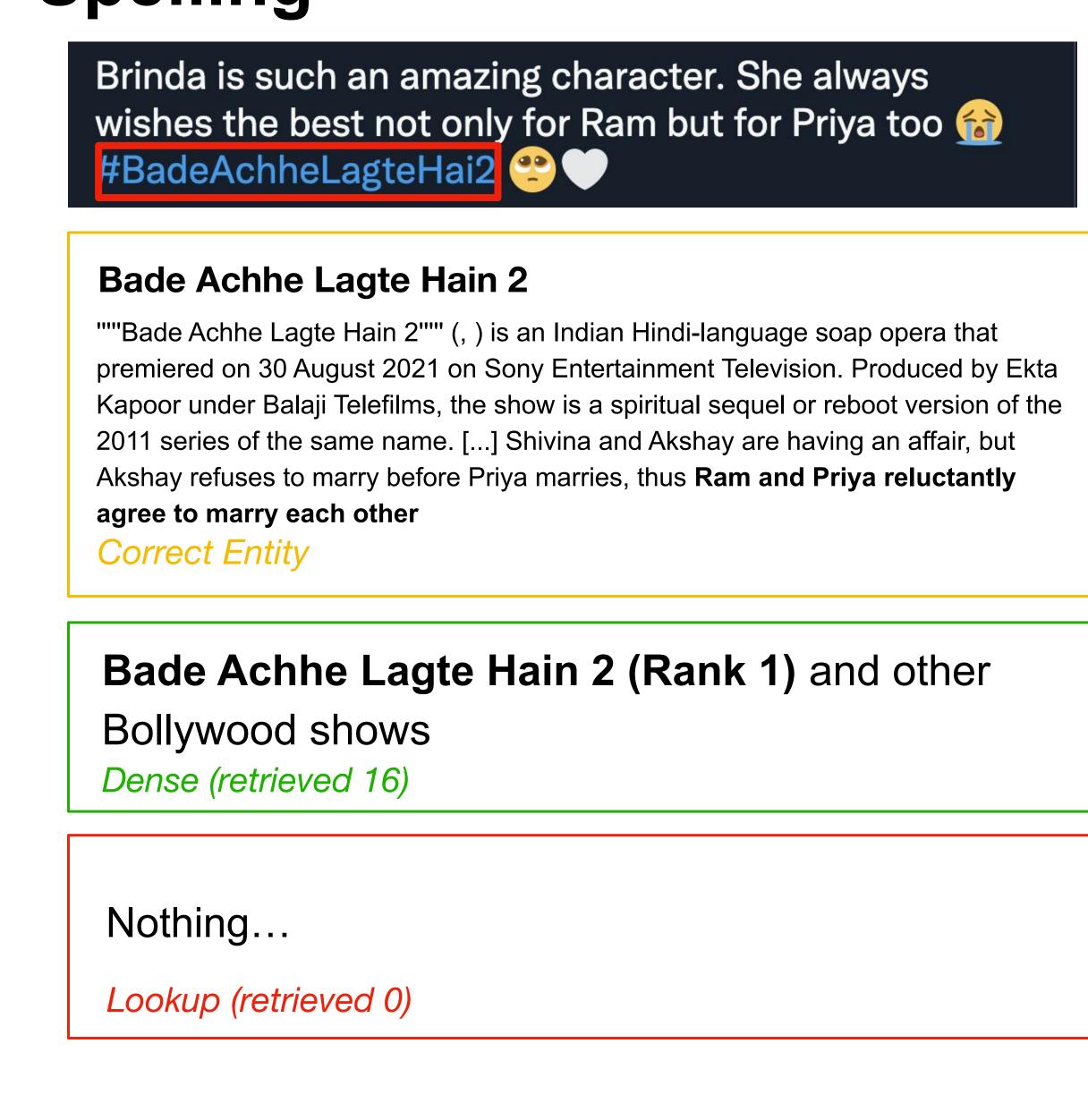
12 Rounds (Film)

""12 Rounds"" is a 2009 American action film directed by Renny Harlin and produced by WWE Studios. The cast is led by John Cena, alongside Aidan Gillen, Steve Harris, Gonzalo Menendez, Brian J. White, Ashley Scott, and Taylor Cole.

Various ammunition wikipedia pages, including Military 12-gauge cartridges Dense (retrieved 16)

12 Rounds (Band) and 12 Rounds (Film) Lookup (retrieved 2)

Spelling



Insights

- Dense is really good at entities that require context and finds related entities
- Dense retrieval can easily scale to more entities without retraining
- NER Impacts Lookup more than Dense Retrieval
- Dense is highly dependent on Context