Robust Candidate Generation for Entity Linking on Short Social Media Texts

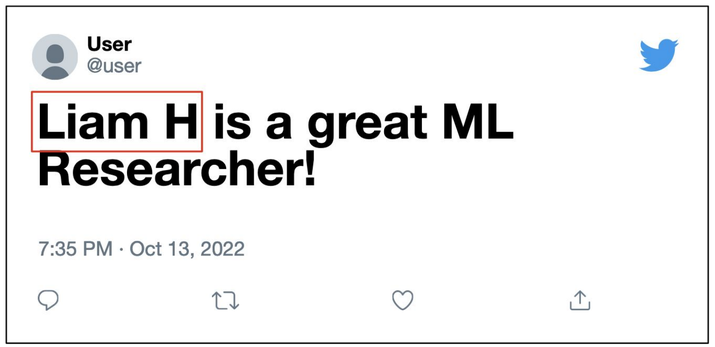
Liam Hebert[1]*, Raheleh Makki[2], Shubhanshu Mishra[2], Hamidreza Saghir[2], Anusha Kamath[2], Yuval Merhav[2]
[1]University of Waterloo, [2]Twitter, Inc.,
*Work done during internship at Twitter, Inc.

2022 The 8th Workshop on Noisy User-generated Text (W-NUT)
Oct 16, 2022 — collocated with COLING 2022.

Entity Linking

Knowledge Base

Liam (2000 Film)	Liam (Settlement)	Liam … (8454 other entities)
Liam Hebert (Researcher)	Liam Hemsworth (Actor)	Liam Neeson (Actor)



NER - Named Entity Recognition

Candidate Generation

Liam (2000 Film)	Liam (Settlement)	Liam … (8454 other entities)
Liam Hebert (Researcher)	Liam Hemsworth (Actor)	Liam Neeson (Actor)

Entity Disambiguation

Example Tweet is not real.

What are the challenges?

- Finding the correct entity could require context
- Users can have creative spelling
- Lookup tables have to be maintained with aliases
- Performance of Candidate Generation relies on accuracy of NER.

Example Tweet is not real.

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?

YES

Zero-Shot Dense Retrieval

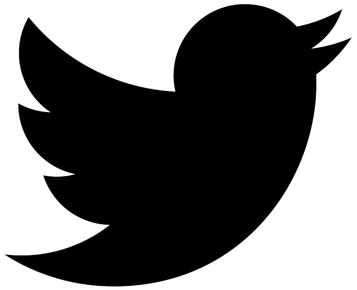
Methodology

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

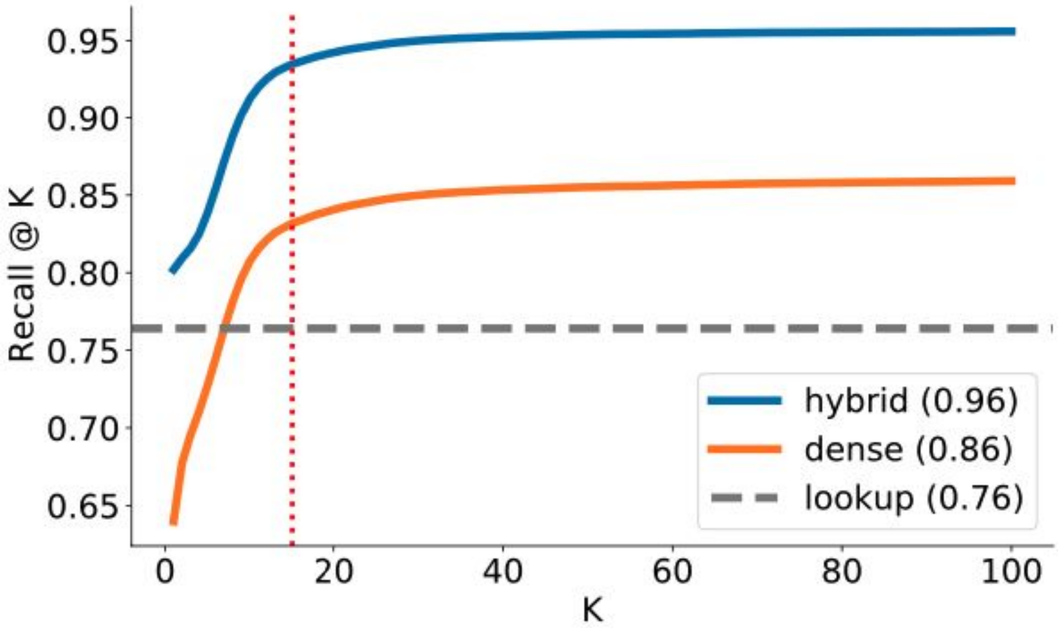
Dense Retrieval:
Pre-trained BLINK[1] Encoders, embeddings indexed using FAISS. First 4 sentences of Wikipedia and annotated spans



TweetNERD Dataset[2]:
Dataset of over 340k+ Labeled Tweets. Evaluated on Academic and OOD Split

[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

Recall@k Using Gold Spans



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.

Recall@16 Using Gold Spans

Data Split	Dense	Lookup	BM25	Hybrid
Academic	0.783	0.741	0.221	**0.916**
OOD	0.772	0.847	0.556	**0.933**
Overall	0.779	0.717	0.362	**0.930**

Recall@16 Using NER Spans

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

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

a continuing trend of complementary results between Dense retrieval and Lookup

Qualitative Analysis (Context)

Wiz and Amber, Rihanna And Chris, Beyonce and jay-z, #grammyscouples

Amber Rose Levonchuck

'''Amber Rose Levonchuck''' (born October 21, 1983) is an American model and television personality. She first gained attention after she appeared in the music video for Young Jeezy and Kanye West's single "Put On". … **After splitting from West in 2010, Rose began dating Rapper Wiz Khalifa, whom she married in 2013.**

Correct Entity

Musicians named Amber, **including Amber Rose Levonchuck (Rank 2)**

Dense (retrieved 16)

Anything with “Amber” in it. AMBER Alert, AMBER telescope, others. No Amber Rose

Lookup (retrieved 46)

Qualitative Analysis (Lack of 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.

Correct Entity

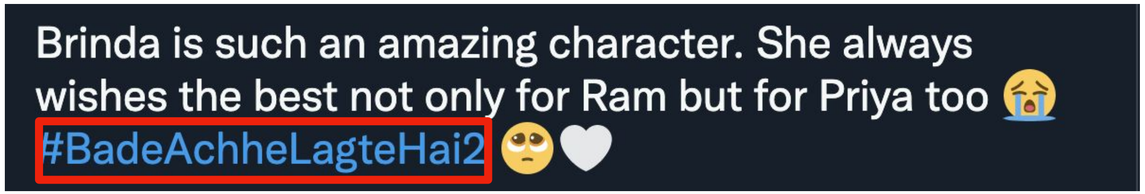
Various ammunition wikipedia pages, including **Military 12-gauge cartridges**

Dense (retrieved 16)

12 Rounds (Band) and **12 Rounds (Film)**

Lookup (retrieved 2)

Qualitative Analysis (Spelling)



Bade Achhe Lagte Hain 2

"""Bade Achhe Lagte Hain 2""" (,) is an Indian Hindi-language soap opera that premiered on 30 August 2021 on Sony Entertainment Television. Produced by Ekta Kapoor under Balaji Telefilms, the show is a spiritual sequel or reboot version of the 2011 series of the same name. [...] Shivina and Akshay are having an affair, but Akshay refuses to marry before Priya marries, thus **Ram and Priya reluctantly agree to marry each other**

Correct Entity

Bade Achhe Lagte Hain 2 (Rank 1) and other Bollywood shows

Dense (retrieved 16)

Nothing…

Lookup (retrieved 0)

Insights Gained From Candidate Generation

- Dense is really good at entities that require context
  - It also 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
- Hybrid combines the best of Lookup and Dense.

Questions?

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/