Real World Conversational Entity Linking Requires More Than Zero-Shots

Anonymous ACL submission

Abstract

Entity linking (EL) in conversations faces notable challenges in practical applications, primarily due to scarcity of entity-annotated conversational datasets and sparse knowledge bases (KB) containing domain-specific, long-tail entities. We designed targeted evaluation scenarios to measure the efficacy of EL models under resource constraints. Our evaluation employs two KBs: Fandom, exemplifying real-world EL complexities, and the widely used Wikipedia. First, we assess EL models' ability to generalize to a new unfamiliar KB using Fandom and a novel zero-shot conversational entity linking dataset that we curated based on Reddit discussions on Fandom entities. We then evaluate the adaptability of EL models to conversational settings without prior training. Our results indicate that current zero-shot EL models falter when introduced to new, domain-specific KBs without prior training, significantly dropping in performance. Our findings reveal that previous evaluation approaches fall short of capturing real-world complexities for zero-shot EL, highlighting the necessity for new approaches to design and assess conversational EL models to adapt to limited resources. The evaluation frame-work and dataset proposed are tailored to facilitate this research.[1]

1 Introduction

Entity Linking (EL) is the process of detecting and resolving ambiguous mentions of entities in a given text by accurately associating them with their corresponding entries in a knowledge base (Kolitsas et al., 2018; Sevgili et al., 2022).

This is a pivotal step in many downstream tasks such as semantic search (Balog, 2018), question answering (Liu et al., 2023), and conversational search (Zamani et al., 2023). EL's significance particularly comes to the fore in the realm of conversational systems as it helps to enhance the accuracy and relevance of the information provided to users during a dialogue session. As these systems are becoming increasingly prevalent in various applications, their ability to ground discussions in real-world knowledge is indispensable for maintaining the integrity and usefulness of the system (Ahmadvand et al., 2019; Fan et al., 2023; Kandpal et al., 2023). Conversations possess characteristics that render common EL models suboptimal (e.g. noisier text, informal language use, entity-related information spreading through turns, etc.) (Joko et al., 2021; Joko and Hasibi, 2022). However, conversational EL has been less explored in prior research, which predominantly concentrates on techniques and benchmarks for long static documents (Logeswaran et al., 2019) or stand-alone queries (Hasibi et al., 2015). On the other hand, traditional EL often presupposes the existence of ample training data (De Cao et al., 2020; Ferragina and Scaiella, 2010; Piccinno and Ferragina, 2014; Van Hulst et al., 2020) , a similar distribution of entities in KB during training and at inference time, and a structurally/textually rich KB for training. These assumptions, however, do not usually hold in real-world EL scenarios, especially in a conversational context, making EL in practice more challenging. Creating an entity-annotated training dataset can be prohibitively exhaustive, or the data might be unavailable due to privacy concerns (Sui et al., 2023). In addition, the distribution of train and test entities might differ as knowledge bases may expand with time, and new entities can be added to the KB which results in an incomplete KB at training time (Aydin et al., 2022; Zhang et al., 2018). Lastly, real-world KBs do not often come with dense structural/textual entity information. As a result, zero-shot entity linking (Logeswaran et al., 2019; Bhargav et al., 2022) was introduced to address some of these challenges. This setup is aimed to allow disambiguating mentions of previously unseen entities by relying on pre-trained models. In

[1]The dataset and relevant experiment codes will be shared

this study, however, we design an evaluation framework and a dataset, addressing the gap between real-world conversational EL and the existing zero-shot EL studies, showing that current zero-shot models do not adequately address practical challenges. We pose our research questions as **RQ1)** *Are zero-shot EL models able to generalize effectively when introduced to a whole new KB, not included in their initial training?* **RQ2)** *How much can zero-shot EL models adapt to conversational settings without prior training?*

We summarize our contributions as:

- Introduced evaluation scenarios to highlight gaps in zero-shot EL research and evaluation inadequacies specifically in conversational settings.
- Created a conversational dataset to demonstrate real-world EL challenges empirically and to facilitate research into methods addressing practical challenges.
- Showed that current zero-shot EL models significantly underperform when applied to new, domain-specific KBs without prior exposure to their entities, emphasizing that zero-shot EL is yet to be effective in solving real EL tasks.

2 Analysis Scenarios

To assess models based on practical constraints we perform the following groups of analysis;

Generalization to Unfamiliar KB

This set of experiments is aimed to assess how well EL models are capable of generalizing to a new KB at inference time. Given G and G' as KBs, models are previously trained on G and encounter G' only at the evaluation step. Particularly selecting G' to ensure the frequency of domain-specific and long-tail entities, makes the task more challenging. Our definition of generalisability differes from that used by (Logeswaran et al., 2019; Wu et al., 2019) in the sense that we do not do training on any part of the new KB.

Adaptability to Conversational Context

In the second set of evaluation experiments, we examine how well EL models perform in a conversational setting. We formulate this as a zero-shot EL task since it tests the model's adaptability to a new domain, given that zero-shot EL models are typically trained for documents, queries, or question-answering settings.

	Train	Test
Conversations	5352	745
Threads	8026	745
All utterances	49695	4557
Annotations	10263	965
Utterances with Annotations	8787	833
Average thread length	6.19	6.11

Table 1: Reddit Conversational Data Statistics

3 Reddit Conversational Dataset for Zero-shot EL

We introduce the Reddit Conversational EL dataset, specifically curated for generalization analysis scenarios.

To curate this dataset we used the Convokit's Reddit corpus[2] (Chang et al., 2020), which includes subreddit posts and comments until October 2018, sourced from the broader Pushshift Reddit dataset[3] (Baumgartner et al., 2020). Convokit offers 948,169 subreddits, among which, we only opt for the discussions around each of the 16 ZESHEL domains (Logeswaran et al., 2019). We extract the subreddits with a ZESHEL's domain title in their name. From each Reddit conversation, we extract its unique threads. In this context, a thread is a distinct path in a hierarchical structure of user utterances, beginning with an original post (the root) and encompassing all subsequent replies until the last reply (the leaf) (Zhang et al., 2019; Henderson et al., 2019). To create gold mention spans along with their gold Fandom entities, we rely on instances where users include hyperlinks to the Fandom website as a way of disambiguating their mention of an entity in their utterance. Next, several preprocessing, pruning, and augmentation steps were performed:

1. Removed URLs, special symbols, non-English characters, repetitive nonsensical tokens, etc.
2. Pruned utterances including profanity keywords (based on a publicly available profanity list (Harel et al., 2022)) and utterances with less than 5 or more than 70 tokens
3. Excluded annotations with nonsensical mentions (e.g. "here", "this link", "link" etc.)
4. Augmented user annotations in cases where the exact mention text is annotated by the user in some occurrences but not others

[2] https://convokit.cornell.edu/documentation/subreddit.html

[3] https://pushshift.io/

	Wikia									Reddit								
	MD			ED			EL			MD			ED			EL		
	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
1 **FLAIR + BLINK** Micro	.027	.255	.048	.026	.222	.047	.015	.147	.027	.130	.186	.153	.167	.232	.194	.064	.093	.076
2 **FLAIR + BLINK** Macro	.029	.269	.051	.029	.241	.051	.015	.156	.028	.136	.202	.162	.160	.237	.191	.057	.088	.069
3 **ELQ** Micro	.034	.205	.058	.015	.088	.025	.010	.062	.017	.135	.313	.189	.162	.367	.225	.069	.161	.097
4 **ELQ** Macro	.036	.223	.062	.019	.117	.033	.013	.081	.022	.123	.285	.171	.142	.323	.197	.057	.134	.080

Table 2: Entity linking micro-averaged scores on Reddit dataset using Fandom as the knowledge base. For each domain, at inference time only the corresponding domain knowledge base is used.

5. Excluded threads with less than 5 utterances and threads with no annotations

We checked the extracted annotations for instances where the gold mention and entity were exact matches. To avoid trivial disambiguation tasks, following (Logeswaran et al., 2019), we ensured no more than 5% of our threads have such annotations. Splitting the final data to train and test sets, we relied on conversation timestamps and annotation density (details in Appendix A). Dataset statistics can be found in Table 1.

4 Experimental Setup

4.1 Entity Linking Models

We focus on assessing two of the very few models purported to facilitate zero-shot entity linking; ELQ (Li et al., 2020) and BLINK (Wu et al., 2019), both BERT-based models that are pre-trained on Wikipedia for EL. ELQ, a biencoder, performs mention detection and entity disambiguation simultaneously in a single pass showing promise in zero-shot QA contexts. Our analysis, however evaluates its ability to adapt to conversations. BLINK, on the other hand, specializes in entity disambiguation, requiring either predefined mention spans or an external mention detection module. It uses a BERT-based biencoder for initial entity ranking followed by a cross-encoder for candidate reranking. The cross-encoder's slower processing and BLINK's segmented approach to entity linking make BLINK less suited for conversational applications.

4.2 Knowledge Bases

Fandom[4], primarily a host for fan-created wikis covering a range of entertainment topics, is the KB used in our generalisability analysis. We use an specific extraction of Fandom for zero-shot EL research called ZESHE Logeswaran et al., 2019 consisting of 16 Fandom domains and comprising approximately 500,000 entities. For our standard setup, we employ the Wikipedia 2019-08-01 dump [5], encompassing more than 5 million entities. This version of Wikipedia serves as the standard KB against which ELQ and BLINK are benchmarked.

4.3 Datasets

Along with the zero-shot conversational Reddit dataset introduced in Section 3, we perform experiments using ConEL datasets (Joko et al., 2021; Joko and Hasibi, 2022) and Wikia[6] documents. This helps contrast conversational and traditional EL settings.

4.4 Analysis Scenarios Setups

Generalisability ELQ and BLINK share the same entity encoder which is trained on Wikipedia (for language understanding and also for EL) but not on Fandom. To assess their generalisability, the mentioned encoder is used to encode Fandom entities using the first 128 tokens of each entity description. Our assessment leverages two distinct data sources; our conversational Reddit data and Wikia validation set. As BLINK does not support mention detection, we evaluated BLINK's performance in two ways. Once we detected potential mentions using FLAIR (Akbik et al., 2018) and provided these mentions to BLINK for entity disambiguation. Next, to assess BLINK's zero-shot entity disambiguation capabilities, we supply it with gold mention spans of the Wikia validation and Reddit test sets and compare it to a naive baseline (Levenshtein distance).

Conversational Context Adapatability This scenario aims to evaluate the EL models' adaptability in a new setting; conversational EL. We evaluate performance in a standard conversational setting using ConEL datasets and the original Wikipedia

[4] https://www.fandom.com/

[5] https://github.com/facebookresearch/BLINK/tree/main/elq

[6] https://github.com/lajanugen/zeshel

	Reddit		Wikia	
	micro	macro	micro	macro
GT + **Edit Distance**	.168	.161	.108	.113
GT + **BLINK**	.288	.233	.446	.457

Table 3: Entity disambiguation performance scores given the ground truth mention spans (GT). Performance is measured in terms of micro and macro averaged precisions across different domains in Reddit and Wikia.

	ConEL1-all		ConEL2-Val		ConEL2-Test	
	MD	EL	MD	EL	MD	EL
1 GENRE	.350	.211	.290	.252	.320	.299
2 TagMe	.510	.375	.559	.478	.611	.504
3 WAT	.416	.336	.616	.539	.613	.519
4 REL	.462	.245	.304	.244	.279	.231
5 CREL	.559	.429	.742	.651	.729	.597
FLAIR + BLINK						
6 WP	.279	.166	.267	.216	.257	.200
ELQ						
7 *WP*	.533	.431	.596	.516	.642	.575
8 **ft_WP** + *WP*	.459	.358	.706	.617	.714	.616

Table 4: Entity linking results on ConEL datasets, reported by F_1-scores (rows 1-5 from Joko and Hasibi, 2022). For the ELQ related results, the italics and bolds respectively depict inference and fine-tuning settings. WP = Wikipedia at inference, ft_WP = fine-tuned on Wikipedia

catalogue as the KB. We further assess ELQ by fine-tuning it on ConEL-2.

5 Are Zero-Shot EL Models Generalisable to New KBs?

We employed FLAIR+BLINK and ELQ as end-to-end zero-shot entity linking systems evaluating their generalisability on Reddit coversations and Wikia documents. Results in Table 2 reveal a significantly low performance when these systems are tested against Fandom without any pre-training on this specific KB, in both documents and conversations. This stark underperformance raises questions regarding the practicality and reliability of these systems as zero-shot EL solutions when confronted with novel, domain-specific knowledge bases in the real-world. The results depict substantial scope for improvement in the mention detection capabilities of both FLAIR and ELQ. By inspecting the predictions, we realized that numerous text spans are considered as possible correct mentions by FLAIR/ELQ, many of which do not align with the gold mentions in the Wikia and Reddit datasets. Given that annotations in both datasets is done by users, this raises the question of whether these models can model entity saliency so that predictions are relevant and align with the user expectations.

Considering table 3 we observe that even given the gold mention spans, correctly linking entities in conversations is more challenging for BLINK than in documents, highlighting the complexity of this environment. This highlights the need for better entity disambiguation techniques that consider and leverage conversational characteristics for improved disambiguation.

6 Are Zero-Shot EL Models Adaptable to Conversational EL Task?

We analyzed adaptability of end-to-end EL systems, specifically FLAIR+BLINK and ELQ, for disambiguating entity mentions in conversations without prior training in this context—a zero-shot setup. Findings are summarized in Table 4, where rows 1-5 show common EL systems evaluated by Joko and Hasibi, 2022, with only CREL being optimized for conversations. Results for FLAIR+BLINK and ELQ can be found in rows 6 and 8 respectively. FLAIR underperforms in conversation mention detection, while ELQ excels in both mention detection and entity disambiguation, outdoing most models except CREL which is optimized for conversations. This adaptation is probably due to the integrated MD and ED operation of ELQ. This highlights the efficacy of end-to-end EL approach in conversational settings and specifically when training resources for new tasks are limited.

7 Conclusions and Future Work

This study re-examined the efficacy of current EL models in conversational scenarios with limited data and KB resources. Motivated by the real-world challenges frequent when integrating EL components into conversational assistants, we recognized overlooked practical limitations in zero-shot EL research. We showed that current zero-shot EL models critically underperform when introduced to a new KB at inference time, due to shortcomings in both mention detection and entity disambiguation functions. These results highlight the need for designing better end-to-end zero-shot EL systems that are reliable in various tasks and KB constraint scenarios. We conclude that the evaluation approaches being used so far in EL literature to evaluate zero-shot EL models are quite naive and not representative of the user's perspective on entity saliency, a crucial point when in interactive systems. For future work, we will leverage our curated dataset to advance model capabilities.

8 Limitations

Our experiment setup involves the use of a new KB, however, the number of EL systems allowing such a use case is very limited. On the other hand, end-to-end EL systems capable of integrating mention detection and entity disambiguation is also limited. These made our choice of models to evaluate quite restricted. Additionally, to test the capabilities of models in zero-shot conversational setup, we needed a conversational dataset that is annotated by entities in a specific-domain KB with long-tail entities. Such data is usually proprietary and not open-access, thereby we had to simulate such a scenario. It would be interesting to assess whether our results hold for other domain-specific settings.

References

Ali Ahmadvand, Harshita Sahijwani, Jason Ingyu Choi, and Eugene Agichtein. 2019. Concet: Entity-aware topic classification for open-domain conversational agents. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1371–1380.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Gizem Aydin, Seyed Amin Tabatabaei, Giorgios Tsatsaronis, and Faegheh Hasibi. 2022. Find the funding: Entity linking with incomplete funding knowledge bases. *arXiv preprint arXiv:2209.00351*.

Krisztian Balog. 2018. *Entity-oriented search*. Springer Nature.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

GP Shrivatsa Bhargav, Dinesh Khandelwal, Saswati Dana, Dinesh Garg, Pavan Kapanipathi, Salim Roukos, Alexander Gray, and L Venkata Subramaniam. 2022. Zero-shot entity linking with less data. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1681–1697.

Jonathan P Chang, Caleb Chiam, Liye Fu, Andrew Z Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. *arXiv preprint arXiv:2005.04246*.

Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. 2013. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22nd international conference on World Wide Web*, pages 249–260.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

Yue Fan, Kevin K Bowden, Wen Cui, Winson Chen, Vrindavan Harrison, Angela Ramirez, Saaket Agashe, Xinyue Gabby Liu, Neha Pullabhotla, NQJ Bheemanpally, et al. 2023. Athena 3.0: Personalized multimodal chatbot with neuro-symbolic dialogue generators. In *Alexa Prize SocialBot Grand Challenge 5 Proceedings*.

Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628.

Itay Harel, Hagai Taitelbaum, Idan Szpektor, and Oren Kurland. 2022. A dataset for sentence retrieval for open-ended dialogues. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2960–2969.

Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. 2015. Entity linking in queries: Tasks and evaluation. In *Proceedings of the 2015 international conference on the theory of information retrieval*, pages 171–180.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, et al. 2019. A repository of conversational datasets. *arXiv preprint arXiv:1904.06472*.

Hideaki Joko and Faegheh Hasibi. 2022. Personal entity, concept, and named entity linking in conversations. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4099–4103.

Hideaki Joko, Faegheh Hasibi, Krisztian Balog, and Arjen P de Vries. 2021. Conversational entity linking: problem definition and datasets. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2390–2397.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. *arXiv preprint arXiv:1808.07699*.

Belinda Z Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient one-pass end-to-end entity linking for questions. *arXiv preprint arXiv:2010.02413*.

Shuo Liu, Gang Zhou, Yi Xia, Hao Wu, and Zhufeng Li. 2023. A data-centric way to improve entity linking in knowledge-based question answering. *PeerJ Computer Science*, 9:e1233.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. *arXiv preprint arXiv:1906.07348*.

Francesco Piccinno and Paolo Ferragina. 2014. From tagme to wat: a new entity annotator. In *Proceedings of the first international workshop on Entity recognition & disambiguation*, pages 55–62.

Özge Sevgili, Artem Shelmanov, Mikhail Arkhipov, Alexander Panchenko, and Chris Biemann. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3):527–570.

Xuhui Sui, Ying Zhang, Kehui Song, Baohang Zhou, Xiaojie Yuan, and Wensheng Zhang. 2023. Selecting key views for zero-shot entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1303–1312.

Johannes M Van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P de Vries. 2020. Rel: An entity linker standing on the shoulders of giants. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2197–2200.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.

Hamed Zamani, Johanne R Trippas, Jeff Dalton, Filip Radlinski, et al. 2023. Conversational information seeking. *Foundations and Trends® in Information Retrieval*, 17(3-4):244–456.

Shaohua Zhang, Jiong Lou, Xiaojie Zhou, and Weijia Jia. 2018. Entity linking facing incomplete knowledge base. In *Web Information Systems Engineering–WISE 2018: 19th International Conference, Dubai, United Arab Emirates, November 12-15, 2018, Proceedings, Part II 19*, pages 325–334. Springer.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

A Zero-Shot Conversational EL Reddit Data

Our final threads timeline spans from April 27, 2010, to October 31, 2018. Threads dated up to January 1, 2015, were allocated to the training set. For the test set, we selected the densest thread from conversations post-January 1, 2015, as the test thread, incorporating the rest into the training set.

B Replicating BLINK Results on Fandom

To ensure our results are comparable to those reported in (Wu et al., 2019), we used their Wikipedia-trained bi-encoder and cross-encoder model (the only trained models they released) and evaluated it on Wikia's validation set using the evaluation approaches and metrics employed by BLINK's authors. We included the results in Table 5. As this model is only trained on Wikipedia and the scores in BLINK paper are based on a Fandom-trained model, the performance is close but still lower than the ones reported by the authors.

C Evaluation Metrics

We evaluate the performance of the EL systems across three aspects; mention detection (MD), entity disambiguation (ED) (Cornolti et al., 2013), and entity linking (EL). To assess mention detection (MD) we employ a strict matching criterion, where a predicted span is deemed accurate only if it has complete overlap with the corresponding gold standard mention span. Given the entity catalogue E, let T and $\hat{T}$ be the set of gold and predicted mention and entity pairs respectively. Consequently, with our matching criterion, the set of final true positives for entity linking will be defined as;

$$\begin{aligned} \text{C} = \{ \text{ e} \in E \mid [m_s, m_e] = [\hat{m}_s, \hat{m}_e], \\ (e, [m_s, m_e]) \in T, (e, [\hat{m}_s, \hat{m}_e]) \in \hat{T} \} \end{aligned}$$

We report precision (p), recall (r) and F1-score (F_1) for the three aspects whenever it is relevant. For generalisability experiments, both micro and macro averaging are used to report the scores across multiple Fandom domains.

Dataset	Biencoder Accuracy	Recall@64	Crossencoder Norm. Acc.	Overall Unnorm. Acc.
Elder Scrolls	0.3539	0.8959	0.4722	0.4232
Muppets	0.5113	0.8195	0.6500	0.5330
Ice Hockey	0.4532	0.8571	0.4841	0.4151
Coronation Street	0.2077	0.6981	0.6325	0.4419
Macro average	.382	.818	.560	.453

Table 5: Performance of BLINK on Wikia Validation Set. The scores reported align with the evaluation approach used in BLINK