Real World Conversational Entity Linking Requires More Than Zero-Shots

Mohanna Hoveyda[1] **Arjen P. de Vries**[1] **Maarten de Rijke**[2] **Faegheh Hasibi**[1]

[1]Radboud University, The Netherlands

mohanna.hoveyda@ru.nl, arjen.devries@ru.nl, faegheh.hasibi@ru.nl

[2]University of Amsterdam, The Netherlands

m.derijke@uva.nl

Abstract

Entity linking (EL) in conversations faces notable challenges in practical applications, primarily due to the 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 setup and the dataset proposed in this research are made publicly available.[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 (Gerritse et al., 2022; Chatterjee and Dietz, 2022; Hasibi et al., 2016), question answering (Liu et al., 2023), and conversational search (Zamani et al., 2023).

The significance of EL 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., noisy text, informal language, and entity-related information spreading through turns (Joko et al., 2021; Joko and Hasibi, 2022). However, conversational EL has been less explored in prior research and is predominantly focused on techniques and benchmarks for long documents (Logeswaran et al., 2019) or stand-alone queries (Hasibi et al., 2015, 2017). El approaches also often assume the existence of ample training data (Cao et al., 2021; Van Hulst et al., 2020; Piccinno and Ferragina, 2014), 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.

Train and deployment of EL systems in general poses several other challenges as well. 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

[1]https://github.com/informagi/reddit_ConEL

Findings of the Association for Computational Linguistics ACL 2024, pages 13938–13946

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 this study, 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?*

The contributions of this paper are:

- Introducing evaluation scenarios to highlight gaps in zero-shot EL research and evaluation inadequacies specifically in conversational settings.
- Creating a conversational dataset to demonstrate real-world EL challenges empirically and to facilitate research in addressing practical EL challenges.
- Demostrating 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 and EL task

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 generalizability diverges from the definition adopted by (Logeswaran et al., 2019) and (Wu et al., 2020). In our approach, we strictly enforce that there is no overlap between the knowledge bases used for training and evaluation. Specifically, we use models that are exclusively trained on Wikipedia (G) and are only exposed to the Fandom knowledge base (G') during evaluation. This contrasts with the methodologies in the cited studies, where models receive training on some segments of the Fandom knowledge base before evaluation, even though these are distinct from the test segments.

Along unfamiliar KB scenario, we also assess the generalization of the EL systems to a new setting which is conversational EL, since these models have not been previously trained on this setting, we intend to assess their generalizability to this setting as well.

Adaptability to Conversational EL Task

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 setting (i.e., conversational), given that zero-shot EL models are typically trained for documents, and queries and not conversations.

	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 evaluating zero-shot EL methods in conversational setup and with unseen KBs, used for our 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 domains used in ZESHEL (Logeswaran et al., 2019) (introduced in 4.2). We extract subreddits that contain a ZESHEL domain title in their name. From each Reddit conversation, we extract its unique threads.

[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
Flair + BLINK Micro	.027	.255	.048	.026	.222	.047	.015	.147	.027	.130	.186	.153	.167	.232	.194	.064	.093	.076
Flair + BLINK Macro	.029	.269	.051	.029	.241	.051	.015	.156	.028	.136	.202	.162	.160	.237	.191	.057	.088	.069
ELQ Micro	.034	.205	.058	.015	.088	.025	.010	.062	.017	.135	.313	.189	.162	.367	.225	.069	.161	.097
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 and macro-averaged scores on Reddit dataset using Fandom as the knowledge base MD, ED, and EL show the relevant scores for mention detection, entity disambiguation, and entity linking. The scores indicate precision (P), recall (R), and f1-score (F). Only the corresponding domain knowledge base is used for each domain at inference time.

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., 2020; 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, and repetitive nonsensical tokens.
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
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 C). Dataset statistics can be found in Table 1. Samples of the dataset are included in Table 6.

4 Experimental Setup

We detail the entity linking models, datasets and knowledge bases used in our experiments, as well as experimental details of our analysis setups.

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., 2020), both BERT-based models that are pre-trained on Wikipedia for EL. ELQ is based on a biencoder model and 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 bi-encoder for initial entity ranking followed by a cross-encoder for candidate re-ranking. The cross-encoder's processing, while thorough, is slower compared to the biencoder in ELQ, which can be a disadvantage for applications requiring real-time response, such as conversational systems. Additionally, BLINK's segmented approach to entity linking, which involves separate processes for mention detection and entity disambiguation, further reduces its suitability for conversational scenarios.

It is crucial to note that the BLINK model we employ was trained using the Wikipedia KB and has not been exposed to the Fandom KB, ensuring no overlap with the knowledge base used in our evaluations.

4.2 Knowledge Bases

We have selected Fandom,[4] as the KB for our generalizability analysis. Fandom acts as a hub for 'fan-created wikis', covering a range of entertainment topics. We use a specific extraction of Fandom for zero-shot EL research called ZESHEL (Logeswaran et al., 2019) consisting of 16 Fandom domains and comprising approximately 500,000

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

entities. For our standard setup, we employ the Wikipedia dump from 2019-08-01,[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 test set of 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. The ConEL datasets, comprising ConEL 1 and 2, are derived from various sources and have been specifically curated and human-annotated for the entity linking task against Wikipedia. On the other hand, the Wikia dataset comprises documents featuring mentions of entities from Fandom, with entity annotations contributed by users of the Fandom website. Employing both these datasets allows us to effectively delineate the distinctions between conversational and traditional entity linking which mainly focuses on document-level entity linking.

4.4 Analysis Scenarios Setups

Generalizability. 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 generalizability, 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 the performance of EL methods in a standard conversational setting using ConEL datasets and Wikipedia as the KB. We then asses the adapatability of ELQ method by fine tuning it on the conversational data (ConEL-2 dataset) and compare its performance with the original mode.

[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). Evaluation is done on Reddit conversational dataset and on Wikia documents, against Fandom as the knowledge base. Scores are presented as micro-averaged and macro-averaged precision, the first aggregates true positives and false positives across all Fandom domains, while the latter is calculated by averaging domain-specific precision scores.

	ConEL1		ConEL2-Val		ConEL2-Test	
	MD	EL	MD	EL	MD	EL
GENRE	.350	.211	.290	.252	.320	.299
TagMe	.510	.375	.559	.478	.611	.504
WAT	.416	.336	.616	.539	.613	.519
REL	.462	.245	.304	.244	.279	.231
CREL	**.559**	.429	**.742**	**.651**	**.729**	.597
Flair + BLINK	.279	.166	.267	.216	.257	.200
ELQ	.533	**.431**	.596	.516	.642	.575
ELQ (FT)	.459	.358	.706	.617	.714	**.616**

Table 4: Entity linking results on ConEL datasets, reported by F_1-scores (top rows from Joko and Hasibi, 2022). Flair+BLINK and ELQ use Wikipedia for both training and inference. ELQ (FT) denotes fine tuning on conversational data (ConEL-2 train set).

5 Results and Discussion

5.1 Are Zero-Shot EL Models Generalizable?

We employed Flair+BLINK and ELQ as end-to-end zero-shot entity linking systems evaluating their generalizability 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 methods 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.

5.2 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 the top rows show common EL systems evaluated by Joko and Hasibi, 2022, with only CREL (Joko and Hasibi, 2022) being optimized for conversations and the rest (GENRE, TagMe, WAT and REL) are general-purpose EL systems. Results for Flair+BLINK and ELQ can be found in the second part of table. Flair underperforms in conversation mention detection, while fine tuned ELQ (adapted to conversational setup) excels in both mention detection and entity disambiguation, outdoing most models except CREL which is optimized for conversations. The adaptability of ELQ is likely due to the end-to-end MD and ED training, as well as similarity to the domain it is initially trained on.

6 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.

7 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 domains.

8 Acknowledgement

This publication is part of the project LESSEN with project number NWA.1389.20.183 of the research program NWA ORC 2020/21 which is (partly) financed by the Dutch Research Council (NWO).
We also would like to express our gratitude towards Emma Gerritse and Hideaki Joko for their insightful advice and inspiring discussions, which influenced the direction of the paper.

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 *In Proceddings of the 27th International Conference on Computational Linguistics COLING '18'*, pages 1638–1649.

Gizem Aydin, Seyed Amin Tabatabaei, George Tsatsaronis, and Faegheh Hasibi. 2022. Find the funding: Entity linking with incomplete funding knowledge bases. In *Proceedings of the 29th International*

Conference on Computational Linguistics (COLING '22)', pages 1937–1942.

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.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60.

Shubham Chatterjee and Laura Dietz. 2022. Bert-er: Query-specific bert entity representations for entity ranking. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1466–1477.

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.

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*.

Emma J Gerritse, Faegheh Hasibi, and Arjen P de Vries. 2022. Entity-Aware Transformers for Entity Search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, pages 1455–1465.

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.

Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. 2016. Exploiting Entity Linking in Queries for Entity Retrieval. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, pages 209–218.

Faegheh Hasibi, Krisztian Balog, and Svein Erik Bratsberg. 2017. Entity Linking in Queries: Efficiency vs. Effectiveness. In *Proceedings of the 39th European Conference on Information Retrieval*, pages 40–53.

Matthew Henderson, Paweł Budzianowski, Iñigo Casanueva, Sam Coope, Daniela Gerz, Girish Kumar, Nikola Mrkšić, Georgios Spithourakis, Pei-Hao Su, Ivan Vulić, and Tsung-Hsien Wen. 2019. A repository of conversational datasets. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 1–10.

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. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529.

Belinda Z. Li, Sewon Min, Srinivasan Iyer, Yashar Mehdad, and Wen-tau Yih. 2020. Efficient one-pass end-to-end entity linking for questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP '20)*, pages 6433–6441.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460.

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. 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.

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, November 12-15, 2018, Proceedings, Part II 19*, pages 325–334.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

A Replicating BLINK Results on Fandom

To ensure our results are comparable to those reported in (Wu et al., 2020), 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.

Dataset	R@64	Bi	Cross	All
Elder Scrolls	.896	.354	.4722	.423
Muppets	.819	.511	.650	.533
Ice Hockey	.857	.453	.484	.415
Coronation Street	.698	.208	.632	.442
Macro average	**.818**	**.382**	**.560**	**.453**

Table 5: Performance of BLINK on Wikia Validation Set. *R@64, Bi, Cross,* and *All* represent Biencoder Recall@64, Biencoder accuracy, Crossencoder normalized accuracy, and overall unnormalized accuracy, respectively. The scores reported align with the evaluation approach used in BLINK.

B 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;

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

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

C 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. We include samples of the dataset in Table 6.

Table 6: Sample Conversations from The Dataset

Conversation Sample #1

Utterances

User #1: *I know this is a far shot but I had an idea today that I thought I would share .*
User #2: *What if BB 8 is actually Boba Fett ? BoBa Fett , BB Eight .*
User #3: *His head is inside of BB 8!*
User #4: *Exactly ! BB 8 confirmed as an updated BT 16 droid : The B'omarr Monks used these droids to carry brains of those who had achieved enlightenment .*
User #5: *OMG THAT IS NOT CANON ANYMORE*

Mentions and Entities

Mention #1: BT 16 droid
Entity #1: https://starwars.fandom.com/wiki/BT-16_perimeter_droid

Conversation Sample #2

Utterances

User #1: *Deck building Hey Guys is the structure deck saga of the Blue eyes White Dragon Still viable to start with ? And is it worth to buy that structure deck 3 times ? Thanks for helping*
User #2: *Nope . Get Structure Deck : Seto Kaiba instead*
User #3: *Not even for casual Play ?*
User #4: *In that case get the Legendary Dragon Decks or Legendary Decks 2. Otherwise buy them singles instead*
User #5: *deck you mean This One ?*

Mentions and Entities

Mention #1: deck
Entity #1: https://yugioh.fandom.com/wiki/Legendary_Dragon_Decks

Conversation Sample #3

Utterances

User #1: *Should Konami release a small Link monster set Like the title says should Konami release maybe a 30 card set for just Link monsters when they drop that was my biggest complaint about synchros and XYZ that they didn't release a small set of just those card type that was mostly filled with generic monsters to help build the extra deck with .*
User #2: *They could release a links starter deck like they did for Synchros.*
User #3: *actually they did but it's garbage*
User #4: *I think it was good for learning how to synchro before they came out in a set . Should do the same for links .*
User #5: *Again , a link strater deck already exists . The problem is that it's crap .*

Mentions and Entities

Mention #1: for Synchros
Entity #1: https://yugioh.fandom.com/wiki/The_Duelist_Genesis
Mention #2: they did
Entity #2: https://yugioh.fandom.com/wiki/Starter_Deck:_Yu-Gi-Oh!_5D%27s
Mention #3: link strater deck
Entity #3: https://yugioh.fandom.com/wiki/Starter_Deck_2017

Conversation Sample #4

Utterances

User #1: *Secret New Hero : Jabba ? Jabba as a hero anyone ?*
User #2: *Was there actually ever a Hutt Jedi ? What about Tusken Raider ? Imagine Jabba being a bullet sponge with ATAT health .*
User #3: *Beldorian the Hutt was a Jedi , but fell to the darkside . Sharad Hett was a Jedi who left the Order and joined the Tuskens . His son A'Sharad Hett eventually became Darth Krayt .*
User #4: *Beldorian the Hutt was killed in a light saber duel by Leia Organa Solo . What .*
User #5: *Leia becomes a jedi in the old EU*

Mentions and Entities

Mention #1: Beldorian the Hutt
Entity #1: https://starwars.fandom.com/wiki/Beldorion
Mention #2: Sharad Hett
Entity #2: https://starwars.fandom.com/wiki/Sharad_Hett
Mention #3: A'Sharad Hett
Entity #3: https://starwars.fandom.com/wiki/Darth_Krayt
Mention #4: Beldorian the Hutt
Entity #4: https://starwars.fandom.com/wiki/Beldorion

Conversation Sample #5

Utterances

User #1: *[Question ?] Borreload Dragon and Eater of Millions Whats the interaction between these two cause bouth trigger at the begin of the damage step*
User #2: *Both cards trigger at the same time . Turn player's trigger is added to the chain first according to SEGOC . CL1: Borreload CL2: Eater Resolve the chain backwards : Eater banishes Borreload Borreload resolves without effect , as it no longer points at any zones*
User #3: *What about the other way around ? I . e . if the turn player controls Eater . CL1: Eater CL2: Borreload Resolve backwards : Borreload takes control of Eater Eater banishes Borreload ?*
User #4: *Borreload can only trigger when it's attacking , so it's controlled by the turn player . This means Borreload will be CL1 under SEGOC , the turn player's optional effects Borreload come before the non turn player's optional effects Eater .*
User #5: *Ahhhh I didn't realise it only triggered when attacking . Thanks !*

Mentions and Entities

Mention #1: SEGOC
Entity #1: https://yugioh.fandom.com/wiki/Simultaneous_Effects_Go_On_Chain

---