Real World Conversational Entity Linking Requires More Than Zero-Shots

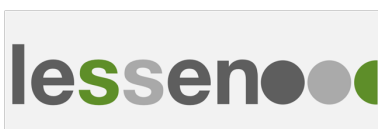
Mohanna Hoveyda[1] **Arjen P. de Vries**[1] **Maarten de Rijke**[2] **Faegheh Hasibi**[1]

[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



Entity Linking

Task of matching vague mentions in text to their respective entities in a knowledge base (KB)

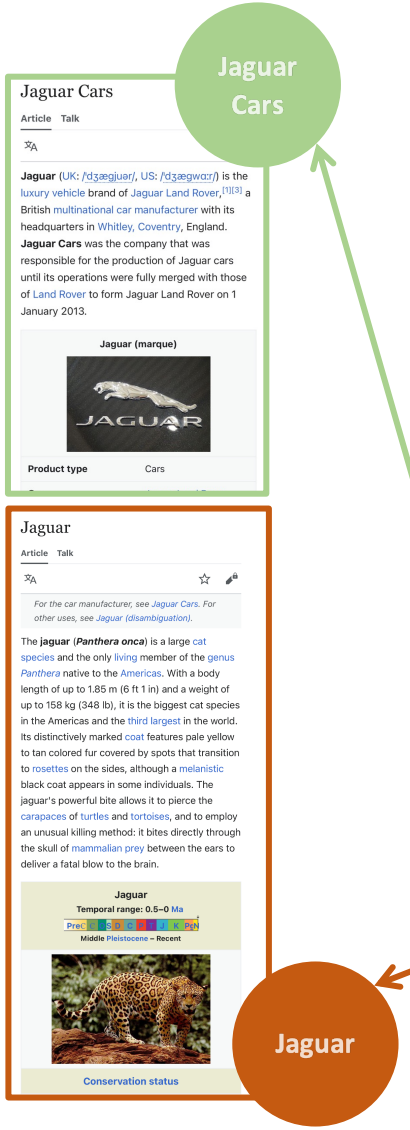
Textual Input (Document)

*".... Last weekend, I had the opportunity to test drive a beautiful **Jaguar**. The sleek design and its handling on the road was impressive..."*

*".... During my visit to the wildlife sanctuary, I was fortunate to spot a magnificent **jaguar** in its natural habitat. This elusive creature..."*

Pivotal/Complementary Step in

- Semantic Search,
- Question Answering,
- Conversational Search,
- Retrieval Augmented Generation



Previous Work on Entity Linking

1 Conversational EL

Conversations' **specific characteristics** compared to documents [1];

- *Informal text*
- *Information spreading through multiple turns*
- *Need for linking more versatile entity types*

These characteristics make general-purpose EL models performance suboptimal for conversational EL. [2]

👱 **User:** *"Planning Paris trip! Any restaurant suggestions?"*

🤖 **Assistant:** *"Definitely try **Le Comptoir du Relais** for a great meal. Also, Eiffel Tower and Louvre are must-visit landmarks."*

👱 **User:** *"Heard **of Le Caveau de la Huchette**? Jazz club recommendation from a friend."*

🤖 **Assistant:** *"Yes! It is a fantastic jazz spot. Enjoy the music! 🎷"*

👱 **User:** *"Any unique boutiques for shopping?"*

🤖 **Assistant:** *"Check out **Colette** for fashion and **Shakespeare and Company** for books. Happy shopping! "*

2 Zero-Shot EL

Definition**: Task of linking mentions to **entities not encountered during training

ZESHEL [3] dataset is

- document-based benchmark introduced for **Z**ero-**Sh**ot **EL**
- curated based on Fandom KB

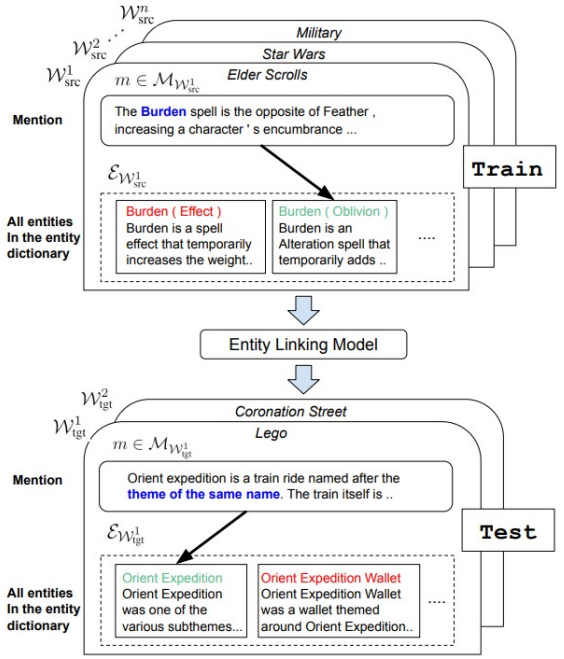


Figure taken from [3]

[1] Personal Entity, Concept, and Named Entity Linking in Conversations (Joko & Hasibi 2022)
[2] Conversational Entity Linking: Problem Definition and Datasets (Joko et al., 2020)
[3] Zero-Shot Entity Linking by Reading Entity Descriptions (logeswaran et al., 2019)

Motivations & Contributions

Focus

Real-World Zeroshot Conversational EL in Face of Data Scarcity

Research questions

1. Are zero-shot EL models able to **generalize effectively to a whole new KB**, that was absent in their initial training?

2. How much can zero-shot EL models **adapt to conversational settings** without prior training?

To answer the questions, we:

1. **Evaluation Scenarios**
   Designed for more realistic evaluation

2. **Reddit ConEL**
   A REDDIT-based dataset introduced for conversational zero-shot EL

3. **Showed zero-shot EL is yet to be effective in real-world tasks**

Introducing *Reddit ConEL*

Reddit ConEL dataset specifically curated for evaluating zero-shot EL methods in conversational setup

PushShift Reddit Dataset

(948,169 subreddits from Convokit)

Preprocessing

Extracting **subreddits with ZESHEL [1] domains' subjects** only

Extract gold mention-entity pairs by relying on **user's hyperlinks to Fandom website**

- Data **cleaning**
- **Filtering** of profanities, extremely long/short utterances, nonsensical mentions, extremely short threads, ….
- **Augmenting** user annotations

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

Sample of a Conversation from the Final Dataset

	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

Final Dataset Statistics

[1] Zero-Shot Entity Linking by Reading Entity Descriptions (logeswaran et al., 2019)

Evaluation Scenarios & Results (1)

Scenario 1

Generalization to Unfamiliar KB (no pre-training on the entities)

Models: *ELQ & FLAIR+BLINK*
Knowledge base: *Fandom*
Datasets: *Conversational (Reddit ConEL) and Documents (Wikia [1])*

	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.

- **Significantly low (MD and EL) performance** (both in conversations & documents)
  - **MD: Numerous text spans considered as possible mentions by Flair/ELQ, many of which do not align with the gold mentions in the Wikia/Reddit datasets**
  - **EL: ELQ and FLAIR+BLINK fail to generalize in a totally new KB setup**

	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.

Zero-shot **conversational** entity disambiguation is more challenging for **BLINK** than in **documents**

[1] Zero-Shot Entity Linking by Reading Entity Descriptions (logeswaran et al., 2019)

Evaluation Scenarios & Results (2)

Scenario 2

Adaptability to Conversational EL Task

Models: *ELQ (w/o fine-tuning) & FLAIR+BLINK*
Knowledge base: *Wikipedia*
Datasets: *Conversational (ConEL 1&2 [1,2])*

- Many general-purpose EL systems **outperform FLAIR+BLINK,**
- **However, ELQ shows significant performance** w/o fine-tuning on **conversational EL** task (potentially due to its **one-pass approach**)

	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).

[1] Personal Entity, Concept, and Named Entity Linking in Conversations (Joko & Hasibi 2022)
[2] Conversational Entity Linking: Problem Definition and Datasets (Joko et al., 2020)

Conclusion & Discussion

Existing zero-shot EL models:

Are not reliable as real-world solutions in practice

Can not effectively model entity saliency
so that mention predictions are relevant and align with the user expectations

One-pass EL (MD+ED) seems promising for conversational EL

For **future work**, our experiments' results indicate a need for the followings ;

- **Rethinking evaluation of zero-shot EL models** for **more realistic assessment**
- Design and training of EL models capable of handling **real-world conversational** and **data scarcity** settings

Resources

- Code & *Reddit ConEL* Dataset: https://github.com/informagi/reddit_ConEL