Saarland University

Bosch Center for AI

Cross-Domain Neural Entity Linking

End of Thesis Presentation

Authors:
Hassan Soliman

Co-supervised by:
Prof. Dietrich Klakow

Heike Adel

Mohamed Gad-Elrab

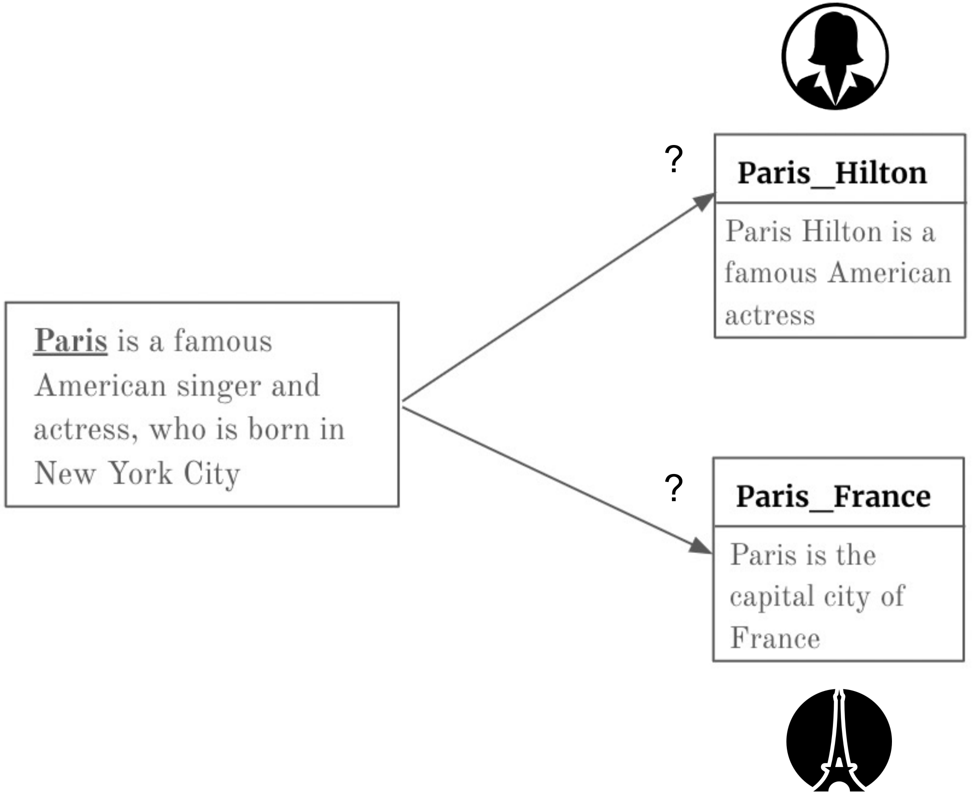
Dragan Milchevski

Agenda

- Motivation and Challenges
- Background and Related Work
- Methodology and Datasets
- Experimental Evaluation
- Conclusion and Summary

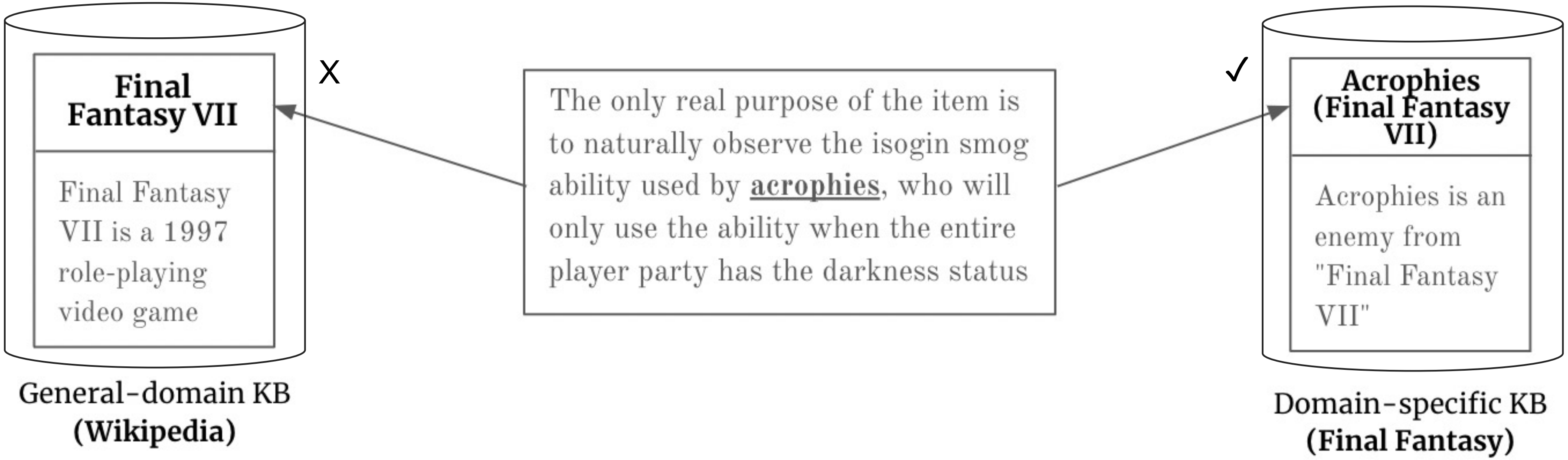
Motivation

- Entity Linking (EL) is the task of **linking a mention** to an **entity** in a **knowledge base (KB)**
- Neural Entity Linking (NEL) gains popularity in the recent years
  - Information extraction capabilities
  - Semantic text understanding
- Existing NEL systems focus on developing models that are typically **domain-dependent**
  - Robust only to a **particular KB** on which they **have been trained**



Necessity of training on multiple KBs

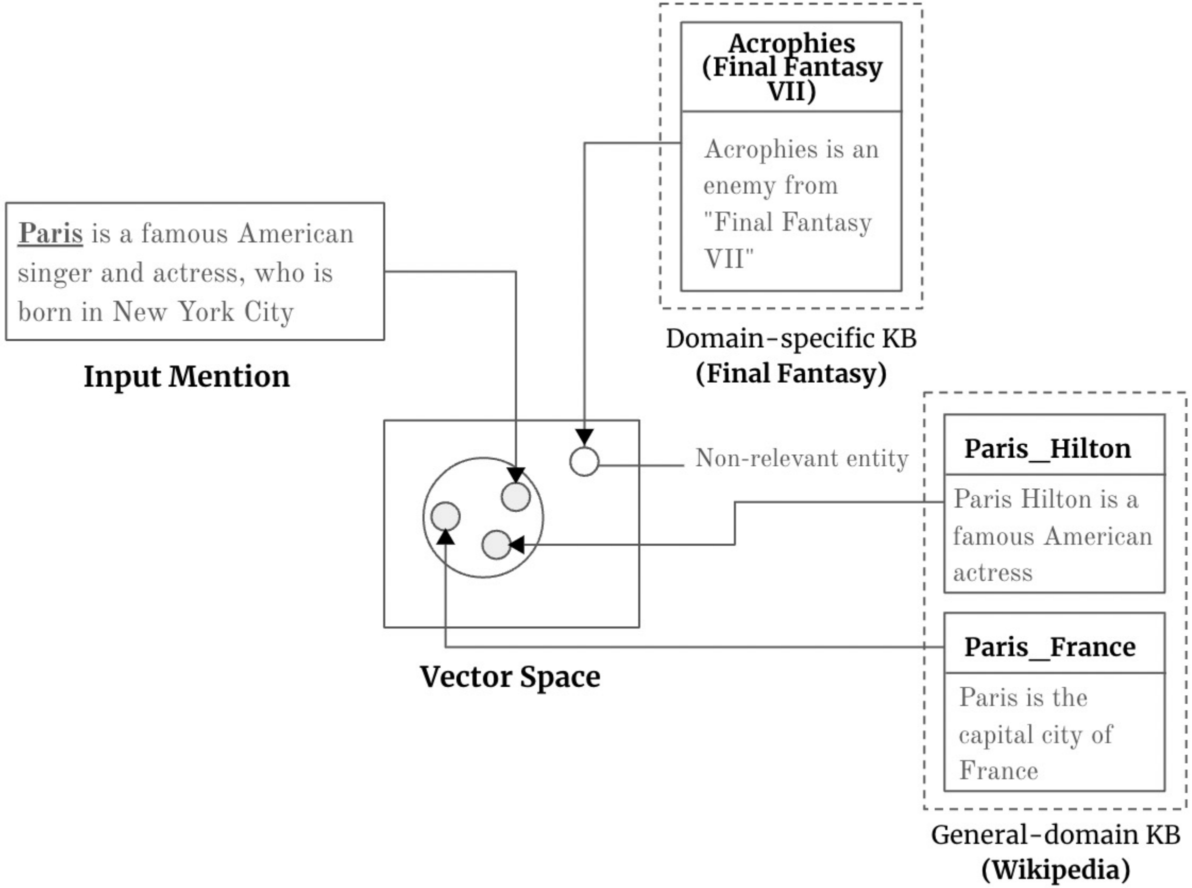
- Mentions in **domain-specific** documents **can not be linked** to a **general-domain** KB
    - They should be **linked** to entities **in a domain-specific** KB
    - Example: **Wikipedia** is the **general-domain** KB combined with a **domain-specific** KB, e.g., **Final Fantasy**



Hassan Soliman | CR/PJ-AI-R26 | 26.07.2021

Problem Definition

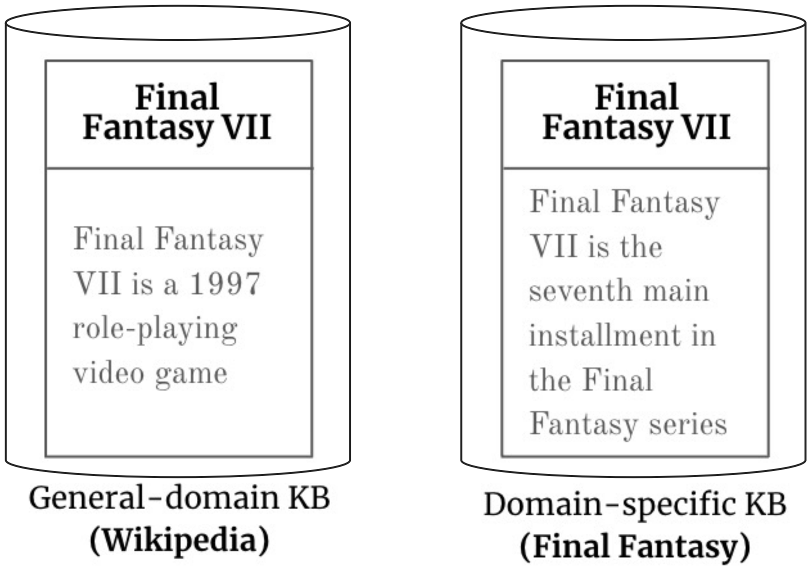
- Develop a **more accurate NEL model** across **different domains**
  - Easy to **expand** to **new domains** by including **a new domain-specific** KB
  - **Enable simultaneous** linking to **two** (**multiple**) **KBs**



BOSCH

Challenges

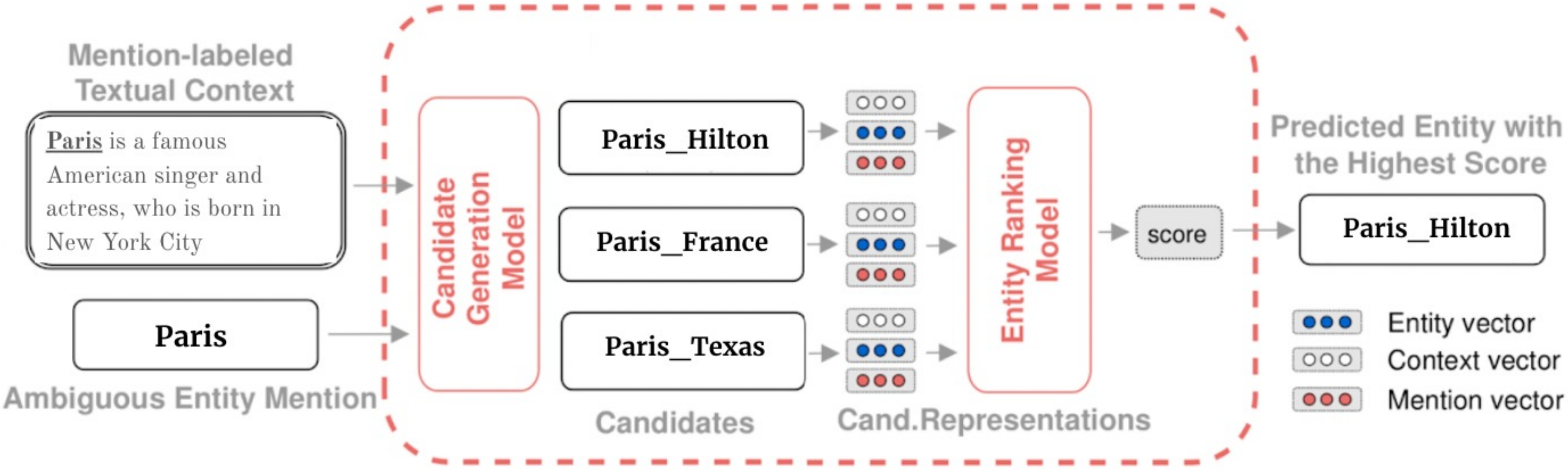
- **Alignment of entities in KBs**
  - When **combining two or more** KBs, entities may be identical (**overlapping**)
- **Having a single representation of KBs**
  - Learn a new **latent representation space** that can represent entities from **multiple** KBs
- **Overfitting to domain-specific KB**
  - Fine-tuned **neural** entity linking models are likely to **overfit** to the **domain-specific** KB



Background and Related Work

Neural Entity Linking

- **General** architecture for a **NEL** system as stated by [Sevgili et al. (2021)[1]]



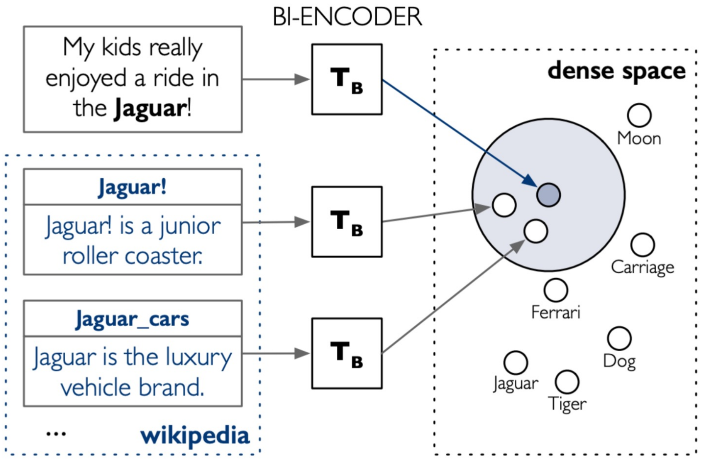
Candidate Generation Module

Candidate Ranking Module

Related Work

- **Domain-adaptive pre-training** by [Logeswaran et al. (2019)[2]]
  - Their **candidate ranking** module is **BERT-based**
  - They **constructed** a new **entity linking dataset** (**Zeshel**) from **Fandom**[1]
  - They use **BM25** which is a variation of **TF-IDF** in their **candidate generation** module

- **Zero-shot entity linking (BLINK)** by [Wu et al. (2020)[3]]
  - Their **BERT-based** model learns **a single space** of **mentions** and **entities**
  - It **only** encodes **entities** of the **domain-specific** KB **without re-training**
  - They **do not** incorporate the **overlapping entities** between **Wikipedia** and a **domain-specific** KB

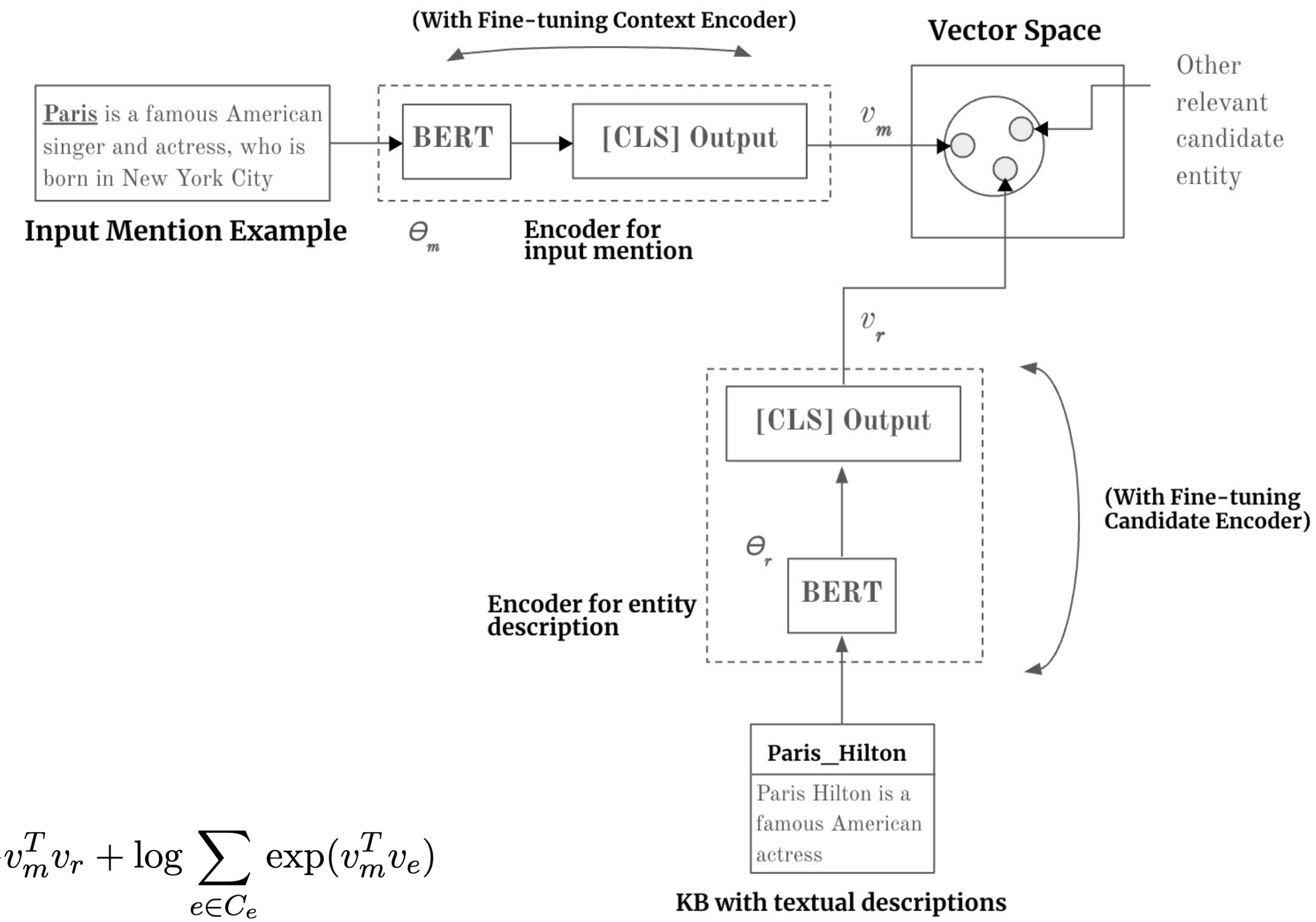


Candidate Generation

[1]**Fandom**, https://www.fandom.com.

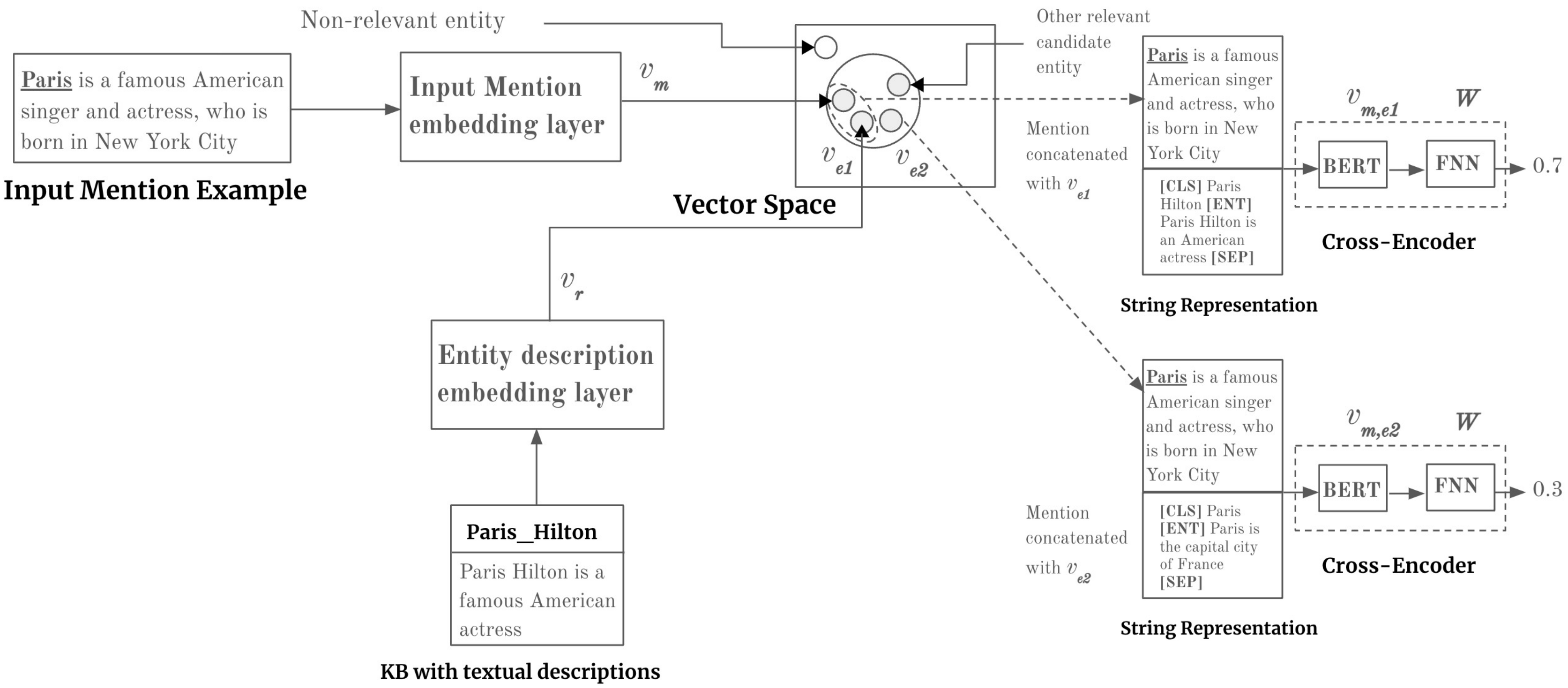
Methodology

BLINK (Base Model): Candidate Generation Phase



$$\mathcal{L}(v_m, v_r)_{\theta_m,\theta_r} = -v_m^T v_r + \log \sum_{e \in C_e} \exp(v_m^T v_e)$$

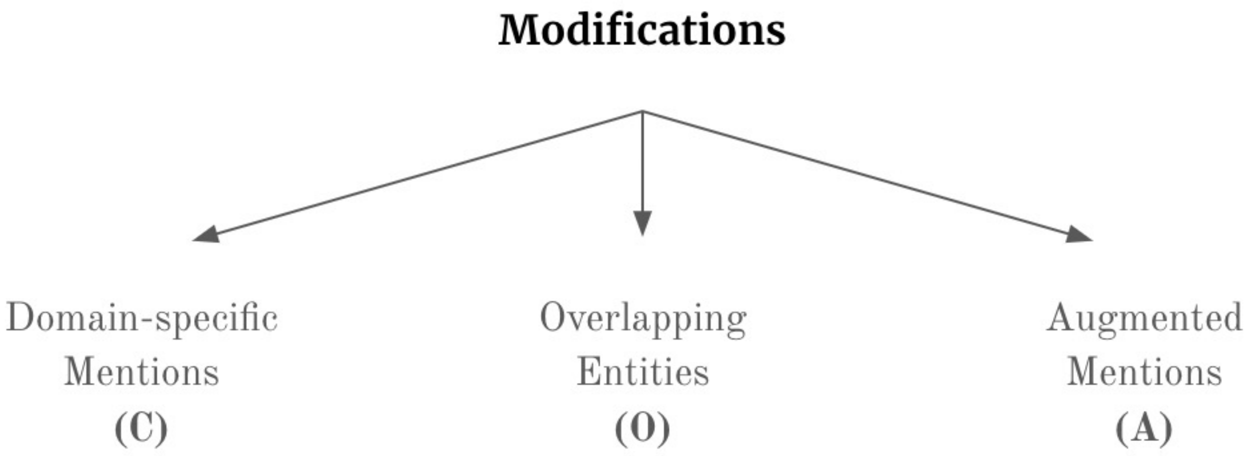
BLINK (Base Model): Candidate Ranking Phase



$$\mathcal{L}(v_{m,e})_W = -v_{m,e}W + \log \sum_{k \in C_e} \exp(v_{m,k}W)$$

CDNEL: Contributions

- Our framework (**CDNEL**) builds on **BLINK** to improve its results, specifically when **linking** to entities from **domain-specific** KBs
  - The key idea is to **fine-tune BLINK** using various proposed **modifications** in this section
  - The goal of these **modifications** is to **better represent** entities from **multiple** KBs to help in the downstream task of **entity linking**



CDNEL: Fine-tuning on the domain-specific KB (C)

- **Challenge**
  - Learn a **new latent representation space** that can represent **entities** from **two or more** KBs

- **Approach**
  - **Fine-tuning** the context and candidate encoders on **mentions with context annotated on the domain-specific KB**
  - The aim of **fine-tuning** is to make the **representation** for the **input mention** with context and the **representation** of the **correct entity** from the **domain-specific** KB **close** together

$$L_\theta = \sum_{m,r \in T} \left(- v_m^\top v_r + \log \sum_{e \in C_e} \exp(v_m^\top v_e) \right)$$

CDNEL: Fine-tuning on the overlapping entities (O)

- **Challenge**
  - When **combining two or more** KBs, entities may be identical (**overlapping**)

- **Approach**
  - **Further fine-tuning** the context and candidate encoders on **the overlapping entities between the general-domain KB and the domain-specific KB**
  - This **learning** is done by **maximizing** the **dot product** between $\mathbf{v}_{\mathbf{o1}}$ and $\mathbf{v}_{\mathbf{o2}}$ representing each **entity** in an **overlapping entity pair**

$$L_\theta = \sum_{o1,o2 \in O} \left(-v_{o1}^\top v_{o2} + \log \sum_{p \in C_p} \exp(v_{o1}^\top v_p) + \log \sum_{q \in C_q} \exp(v_{o2}^\top v_q) \right)$$

CDNEL: Fine–tuning on augmented (additional) mentions (A)

- **Challenge**
  - **Fine-tuned neural** entity linking models are likely to **overfit** to the **domain-specific** KB

- **Approach**
  - **Fine-tuning** the context and candidate encoders on **additional mentions with context annotated on the general-domain KB for augmenting the data**
  - These **general-domain mentions** act as **augmented data** to reduce **overfitting** to the **domain-specific** KB

$$L_\theta = \sum_{m,r \in T} \left(-v_m^\top v_r + \log \sum_{e \in C_e} \exp(v_m^\top v_e) \right)$$

Datasets

- Entity linking dataset (**Zeshel**) constructed by Logeswaran et al. (2019)[2]
  - For each domain, there are **entities** with **textual descriptions** and **labeled mentions** about that domain
  - To get **overlapping entities**, we used a **Sentence-Transformer** model (**Roberta-large**) by Reimers et al. (2019) [4]

Domain	Entities	Mentions			Overlapping Entities	
		Fine-tuning		Test	Matching	Filtered
		Train	Dev			
American Football	31,929	3,000	320	578	24,074	22,928
Doctor Who	40,281	6,360	640	1,334	10,458	3,611
Fallout	16,992	2,500	320	466	2,876	752
Final Fantasy	14,044	4,360	640	1,041	1,495	413
Wikipedia (Reddit)	5,903,538	7,711	409	1,328	-	-

General-domain KB (Wikipedia)

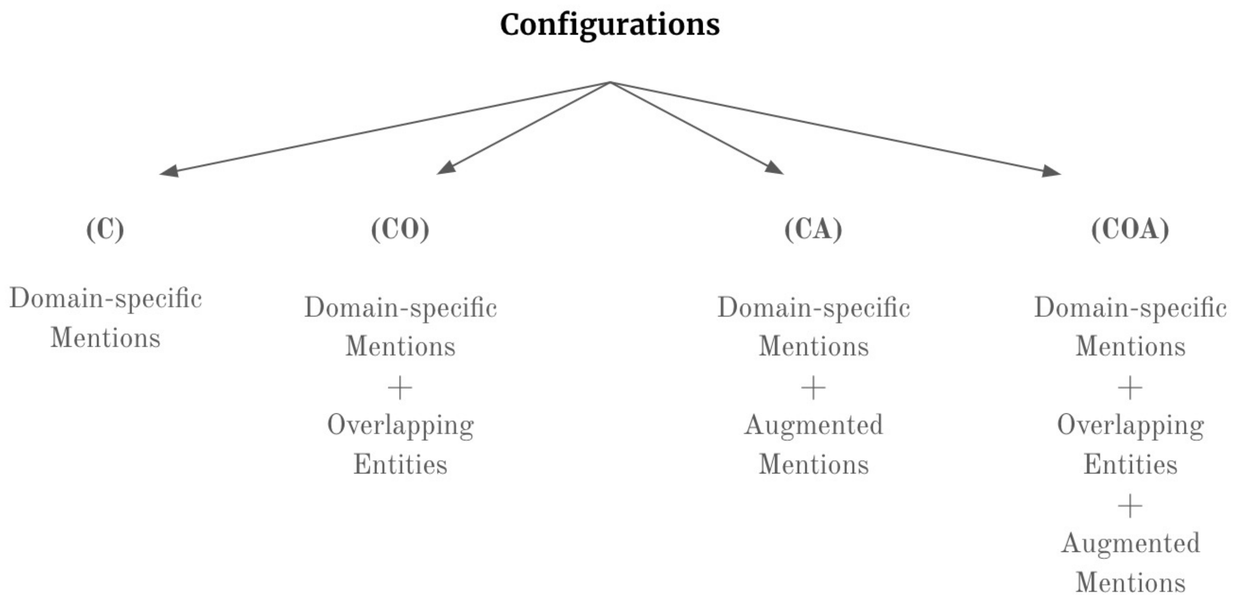
Domain-specific KB (Final Fantasy)

- **Reddit** entity linking dataset by Botzer et al. (2021) [5]
  - Provide **additional mentions** from **Reddit** annotated on the **general-domain** KB (**Wikipedia**)

Experimental Evaluation

CDNEL: Variants

- We aim at **testing** the following **configurations (variants)** and **compare** them with **BLINK** using the **Zeshel** dataset



- They have **in common** the **fine-tuning** on mentions with context annotated on the **domain-specific** KB (**C**)
- We perform **further fine-tuning** on the **overlapping entities** between the **general-domain** KB and a **domain-specific** KB (**O**)
- We used **additional** mentions in the **fine-tuning** that are annotated on the **general-domain** KB (**Wikipedia**) (**A**)

Intrinsic Evaluation of Embeddings

- We evaluate the **joint representation** space of the **entities** from the **combined** KBs
  - The **evaluation** is done on the **overlapping entities** of **each domain-specific** KB and the **general-domain** KB (**Wikipedia**)

	American Football		*Doctor Who*		*Fallout*		*Final Fantasy*	
	MRR	*ACS*	*MRR*	*ACS*	*MRR*	*ACS*	*MRR*	*ACS*
BLINK	0.4991	0.9938	0.4607	0.9650	0.4071	0.9603	0.3623	0.9532
C	0.4982	0.9892	0.3926	0.9095	0.3533	0.9317	0.4136	0.9515
CO	0.4990	0.9919	**0.4932**	0.9784	**0.4558**	0.9680	**0.4400**	0.9628
CA	**0.4999**	**0.9958**	0.4323	0.9605	0.4223	0.9676	0.4072	0.9746
COA	0.4995	0.9896	0.4619	**0.9830***	0.4534*	**0.9820***	0.4209*	**0.9791***

C: Fine-tuning on the domain-specific KB

O: Fine-tuning on the overlapping entities

A: Fine-tuning on the augmented mentions

Intrinsic evaluation of overlapping entities between each domain-specific KB and Wikipedia KB

* shows statistically different results of **COA** in comparison to **BLINK** (randomization test, significance level of **0.05**)

- It has a better **representation space** of the **overlapping entities** in which they are the **most similar (closest)**

BOSCH

Evaluation of Entity Linking

- For **fine-tuning** on **mentions** annotated on the **domain-specific** KB, we address the **research question**
  - "To what extent does **fine-tuning** on **domain-specific** datasets **affect** the results of **simultaneous entity linking**?"

	American Football		*Doctor Who*		*Fallout*		*Final Fantasy*	
	AP@1	*MAP@10*	*AP@1*	*MAP@10*	*AP@1*	*MAP@10*	*AP@1*	*MAP@10*
BLINK	0.1747	0.4104	0.4108	0.4810	0.3412	0.4444	0.3833	0.5179
C	**0.2042**	**0.4578**	**0.6184**	0.6927	0.4313	0.5506	0.3871	0.5407
CO	0.1834	0.4061	0.5735	0.6574	0.4485	0.5590	0.3429	0.4871
CA	**0.2042***	0.4577*	0.6154*	**0.7140***	**0.4657***	**0.5916***	**0.4121***	**0.5710***
COA	0.1626	0.3334	0.5382	0.6165	0.4227	0.5404	0.3900	0.5489

C: Fine-tuning on the domain-specific KB

O: Fine-tuning on the overlapping entities

A: Fine-tuning on the augmented mentions

Evaluation on **mentions** annotated on each **domain-specific** KB

* shows statistically different results of **CA** in comparison to **BLINK** (randomization test, significance level of **0.05**)

- It improves upon **BLINK** for all domains and achieves an **average** gain of **9.5%** with respect to **AP@1**

Hassan Soliman | CR/PJ-AI-R26 | 26.07.2021

Evaluation of Entity Linking

- For **overfitting** of the **fine-tuned** models to the **domain-specific** KB, we address the **research question**
    - Does the fine-tuned models **overfit** to the **domain-specific** KB?

	American Football		*Doctor Who*		*Fallout*		*Final Fantasy*	
	AP@1	*MAP@10*	*AP@1*	*MAP@10*	*AP@1*	*MAP@10*	*AP@1*	*MAP@10*
BLINK	0.8479	0.8973	0.8509	0.8985	0.8509	**0.8987**	0.8494	0.8987
C	0.8517	**0.8974**	0.8170	0.8529	0.8042	0.8452	0.8577	0.8970
CO	0.8517	0.8940	0.8524	0.8859	0.8321	0.8676	0.8592	0.8930
CA	**0.8614**	0.8964	0.8622	0.8992	0.8592	0.8959	0.8773*	**0.9076**
COA	0.8163	0.8387	**0.8773**	**0.9063**	**0.8637**	0.8859	**0.8810**	0.9045

C: Fine-tuning on the domain-specific KB

O: Fine-tuning on the overlapping entities

A: Fine-tuning on the augmented mentions

Evaluation on Reddit mentions annotated on the general-domain KB (Wikipedia)

* shows statistically different results of **CA** in comparison to **BLINK** (randomization test, significance level of **0.05**)

- It proves the **comparability** of **CDNEL** with **BLINK**, which performs **well** by default on **general-domain** mentions

Mentions Qualitative Assessment

- We take a closer look at the **mentions** annotated by **BLINK** and our framework (**CDNEL**)
  - Assessing how the **fine-tuned** model (**CDNEL**) learned to link **mentions** to the **general-domain** and **domain-specific** KBs

Mention	Domain	BLINK	CDNEL
Putin can't do much. Russia has no leverage over us and are already feeling huge pressure from American and EU sanctions (one big reason Putin threw his hat in with Trump and the GOP, to try and lift those sanctions).	Reddit	Russia	**Vladimir Putin**
The only real purpose of the item is to naturally observe the isogin smog ability-used by **acrophies**, who will only use the ability when the entire player party has the darkness status.	Final Fantasy	Acrophies (Final Fantasy V)	**Acrophies (Final Fantasy VII)**

Hassan Soliman | CR/PJ-AI-R26 | 26.07.2021

Conclusion and Summary

Conclusion

- The main goal of our framework (**CDNEL**) is to have **a single system** that allows **simultaneous linking** to **more than one** KB
    - In our case, these are the **general-domain** KB (**Wikipedia**) and the **domain-specific** KB, such as **Final Fantasy**
- For **fine-tuning** on a **domain-specific** dataset to perform **simultaneous entity linking** to **both** KBs
    - We recommend including **additional data (A)** in the form of **general-domain** mentions annotated on **Wikipedia**
    - Adding **overlapping entities (O)** and **augmented data (A)** helps **reduce overfitting** of the **fine-tuned** models
- Further **fine-tuning** on the **overlapping entities** helps **better represent** their **embeddings**, which are as **close** as possible in the **joint representation space**
- Possible actions of this work can go in the direction of **combining more than two** KBs and enabling **simultaneous linking** to them **within the same system**

References

[1] O. Sevgili, A. Shelmanov, M. Arkhipov, A. Panchenko, and C. Biemann. Neural Entity Linking: A Survey of Models Based on Deep Learning.arXiv:2006.00575[cs], Jan. 2021. arXiv: 2006.00575.

[2] Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. Zero-shot entity linking by reading entity descriptions. In Anna Korhonen, David R. Traum, and Llu ́ıs M`arquez, editors, Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 3449–3460. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1335. URL https://doi.org/10.18653/v1/p19-1335.

[3] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettle-moyer. Scalable zero-shot entity linking with dense entity retrieval. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, pages 6397–6407. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.519. URL https://doi.org/10.18653/v1/2020.emnlp-main.519.

[4] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 3980–3990. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1410. URL https://doi.org/10.18653/v1/D19-1410.

[5] Nicholas Botzer, Yifan Ding, and Tim Weninger. Reddit entity linking dataset. Inf. Process. Manag., 58(3):102479, 2021. doi: 10.1016/j.ipm.2020.102479. URL https://doi.org/10.1016/j.ipm.2020.102479.

BOSCH

[CLS] Thanks ! [SEP]