MuVER: Improving First-Stage Entity Retrieval with Multi-View Entity Representations

Xinyin Ma[◇‡], **Yong Jiang**[†*], **Nguyen Bach**[†], **Tao Wang**[†],
Zhongqiang Huang[†], **Fei Huang**[†], **Weiming Lu**[◇*]
[◇] College of Computer Science and Technology , Zhejiang University
[†] DAMO Academy, Alibaba Group
{maxinyin, luwm}@zju.edu.cn, yongjiang.jy@alibaba-inc.com

Abstract

Entity retrieval, which aims at disambiguating mentions to canonical entities from massive KBs, is essential for many tasks in natural language processing. Recent progress in entity retrieval shows that the dual-encoder structure is a powerful and efficient framework to nominate candidates if entities are only identified by descriptions. However, they ignore the property that meanings of entity mentions diverge in different contexts and are related to various portions of descriptions, which are treated equally in previous works. In this work, we propose **Mu**lti-**V**iew **E**ntity **R**epresentations (MuVER), a novel approach for entity retrieval that constructs multi-view representations for entity descriptions and approximates the optimal view for mentions via a heuristic searching method. Our method achieves the state-of-the-art performance on ZESHEL and improves the quality of candidates on three standard Entity Linking datasets[1].

1 Introduction

Entity linking (EL) refers to the task that disambiguates the mentions in textual input and retrieves the corresponding unique entity in large Knowledge Bases (KBs) (Han et al., 2011; Ceccarelli et al., 2013). The majority of neural entity retrieval approaches consist of two steps: Candidate Generation (Pershina et al., 2015; Zwicklbauer et al., 2016), which nominates a small list of candidates from millions of entities with low-latency algorithms, and Entity Ranking (Yang et al., 2018; Le and Titov, 2019; Cao et al., 2021), which ranks those candidates to select the best match with more sophisticated algorithms.

In this paper, we focus on the Candidate Generation problem (a.k.a. the first-stage retrieval). Prior works filter entities by alias tables (Fang et al., 2019) or precalculated mention-entity prior probabilities, e.g., $p(entity|mention)$ (Le and Titov, 2018). Ganea and Hofmann (2017) and Yamada et al. (2016) build entity embedding from the local context of hyperlinks in entity pages or entity-entity co-occurrences. Those embedding-based methods were extended by BLINK (Wu et al., 2020) and DEER (Gillick et al., 2019) to two-tower dual-encoders (Khattab and Zaharia, 2020), which encode mentions and descriptions of entities into high-dimensional vectors respectively. Candidates are retrieved by nearest neighbor search (Andoni and Indyk, 2008; Johnson et al., 2019) for a given mention. Solutions that require only entity descriptions (Logeswaran et al., 2019) are scalable, as descriptions are more readily obtainable than statistical or manually annotated resources.

Although description-based dual-encoders can compensate for the weakness of traditional methods and have better generalization ability to unseen domains, they aim to map mentions with divergent context to the same high-dimensional entity embedding. As shown in Figure 1, the description of "Kobe Bryant" mainly concentrates on his professional journey. As a result, the embedding of "Kobe Bryant" is close to the context which describes the career of Kobe but is semantically distant from his helicopter accident. Dual-encoders are trained to encode those semantically divergent contexts to representations that are close to the embedding of "Kobe Bryant". The evidence relies on the Figure 2 (section 3.2) that the previous method (Wu et al., 2020) is good at managing entities with short descriptions but seems troubling to retrieve entities with long descriptions, which contains too much information to be encoded into a single fixed-size vector.

To tackle those issues, we propose to construct multi-view representations from descriptions. The contributions of our paper are as follows:

[*] Corresponding authors.

[‡] Work was done when Xinyin Ma was interning at Alibaba DAMO Academy.

[1] Our source code is available at https://github.com/Alibaba-NLP/MuVER.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2617–2624
November 7–11, 2021. ©2021 Association for Computational Linguistics

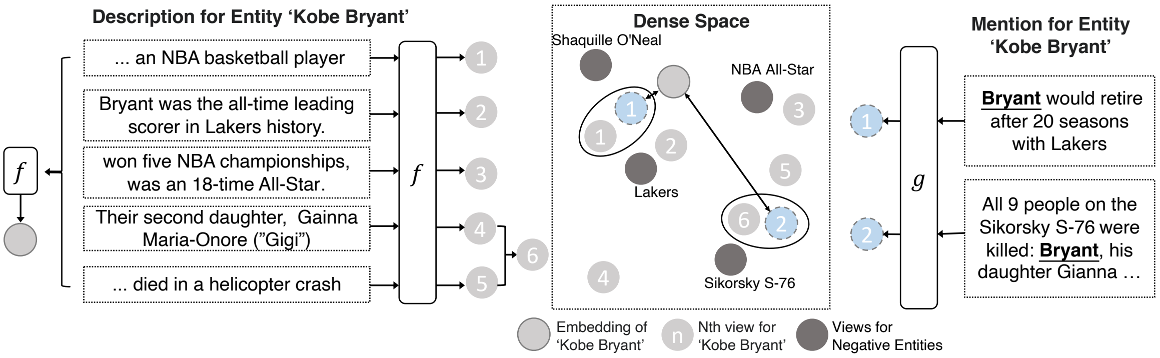
Description for Entity 'Kobe Bryant'
... an NBA basketball player
Bryant was the all-time leading scorer in Lakers history.
won five NBA championships, was an 18-time All-Star.
Their second daughter, Gainna Maria-Onore ("Gigi")
... died in a helicopter crash
f
f
1
2
3
4
5
6
Dense Space
Shaquille O'Neal
NBA All-Star
Lakers
Sikorsky S-76
Mention for Entity 'Kobe Bryant'
Bryant would retire after 20 seasons with Lakers
All 9 people on the Sikorsky S-76 were killed: Bryant, his daughter Gianna ...
g
Embedding of 'Kobe Bryant'
Nth view for 'Kobe Bryant'
Views for Negative Entities

Figure 1: An illustration of our MuVER framework. (i) The contextual information of the given document for the same mention may differ widely (Right), resulting in a large discrepancy between their representations (Blue circles with dashed borders) and the embedding of "Kobe Bryant" has trouble in getting close to both of them. (ii) We refer to each sentence as a view for descriptions to form a view set V (Gray circles with number) and merge views to approximate the optimal views for mentions (points enclosed by ellipses).

- We propose an effective approach, MuVER, for first-stage entity retrieval, which models entity descriptions in a multi-view paradigm.
- We define a novel distance metric for retrieval, which is established upon the optimal view of each entity. Furthermore, we introduce a heuristic search method to approximate the optimal view.
- MuVER achieves state-of-the-art performance on ZESHEL and generates higher-quality candidates on AIDA-B, MSNBC and WNED-WIKI in full Wikipedia settings.

2 Methods

2.1 Problem Setup

Formally, given an unstructured text D with a recognized mention m, the goal of entity linking is to learn a mapping from the mention m to the entity entry e in a knowledge base $\mathcal{E} = \{e_1, e_2, \ldots, e_N\}$, where N can be extremely large (for Wikipedia, $N = 5.9M$). In the literature, existing retrieval methods address this problem in a two-stage paradigm: (i) selecting the top relevant entities to form a candidate set $\mathcal{C}$ where $|\mathcal{C}| \ll |\mathcal{E}|$; (ii) ranking candidates to find the best entity within $\mathcal{C}$. In this work, we mainly focus on the first-stage retrieval, following Logeswaran et al. (2019)'s setting to assume that for each $e \in \mathcal{E}$, entity title t and description d are provided in pairs.

2.2 Multi-View Entity Representations

Dual-encoders We tackle entity retrieval as a matching problem, where two separated encoders, entity encoder f and mention encoder g, are deployed. We consider BERT (Devlin et al., 2019) as the architecture to encode textual input, which can be formulated as:

$$f(t, d) = T_1([CLS]\ t\ [ENT]\ d\ [SEP])$$
$$g(m) = T_2([CLS]\ \text{ctx}_l\ [M_s]\ m\ [M_e]\ \text{ctx}_r\ [SEP])$$

where t, d, m, ctx_l, ctx_r refer to word-pieces tokens of the entity title, the entity description, the mention and the context before and after the mention correspondingly. Besides, we use $[M_s]$ and $[M_e]$ to denote the *start of mention* and *end of mention* identifiers respectively. The special token $[ENT]$ serves as the delimiter of titles and descriptions. T_1 and T_2 are two independent BERT, with which we estimate the similarity between mention m and entity e as $sim(m, e) = f(t, d) \cdot g(m)$.

Multi-view Description Our method matches a mention to the appropriate entity by comparing it with entity descriptions. Motivated by the fact that mentions with different contexts correspond to different parts in descriptions, we propose to construct multi-view representations for each description. Specifically, we segment a description into several sentences. We refer to each sentence as a view v, which contains partial information, to form a view set $\mathcal{V}$ of the entity e. Figure 1 illustrates an example that constructs a view set $\mathcal{V}$ for "Kobe Bryant".

Multi-view Matching Given a view set $\mathcal{V} = \{v_1, v_2, \ldots, v_k\}$ for entity e, determining whether a mention m matches the entity e requires a metric space to estimate the relation between m and $\mathcal{V}$, which can be defined as

$$d(m, \mathcal{V}) = \|g(m) - f(t, [v_1, v_2, ..., v_k, v_i \in \mathcal{V}])\| \quad (1)$$

where $[v_1, v_2, ..., v_k]$ refers to an operation that concatenates tokens in views following the sentence order in descriptions and t is the corresponding entity title for $\mathcal{V}$. Note that this metric can be applied to the subset of $\mathcal{V}$ to focus on partial information of the description. As mentioned before, for m in different contexts, only a part of the views are related. For each mention-entity pair (m, e) and the view set $\mathcal{V}$ of e, we define its optimal Q^* as:

$$Q^*(m, e) \triangleq \arg\min_{Q \subseteq \mathcal{V}} d(m, Q) \quad (2)$$

where Q is a subset of $\mathcal{V}$ and Q^* has the minimal distance to current mention m. We define the distance $d(m, Q^*(m, e))$ as the matching distance between e and m. To find the optimal entity for mention m, we select the entity that has minimal matching distance:

$$e^* = \arg\min_{e \in \{e_1, e_2, ..., e_N\}} d(m, Q^*(m, e)) \quad (3)$$

Distance Metric & Training Objectives The above retrieval process requires an appropriate metric space to estimate the similarity between views and mentions. The metric space should satisfy that similar inputs are pulled together and dissimilar ones are pushed apart. To achieve this, we introduce an NCE loss (van den Oord et al., 2018) to establish the metric space :

$$\mathcal{L}_{NCE} = \mathop{\mathbb{E}}_{\mathcal{E}'}\left[\log \frac{\exp(d(m, Q^*(m, e)))}{\sum_{e_i \in \mathcal{E}'} \exp(d(m, Q^*(m, e_i)))}\right]$$

where $\mathcal{E}' = \{e\} \cup \{e_1, \ldots, e_{n-1}\}$. Mention-entity pairs (m, e) are pulled together and randomly sampled $n-1$ negatives $\{e_1, \ldots, e_{n-1}\}$ are pushed apart from m, based on their matching distance in the current metric space. Unfortunately, $Q^*(m, e)$ is intractable due to the non-differentiable subset operation in Equation 2. Besides, it is time-consuming to obtain the optimal view by checking all subsets exhaustively. In this work, we consider a subset that contains only one view to approximate it. Specifically, we select the best $v^*(m, e) \triangleq \arg\min_{v \in \mathcal{V}} d(m, \{v\})$ from $\mathcal{V}$ as an alternative to the optimal view Q^*:

$$d(m, Q^*(m, e)) \approx d(m, v^*(m, e))) \quad (4)$$

Note that this approximation can be done in time complexity of $O(N)$, which simply selects a view with minimal distance to the given mention. Using Equation 4, we can rewrite the NCE loss as:

$$\mathcal{L}_{NCE} = \mathop{\mathbb{E}}_{\mathcal{E}'}\left[\log \frac{\exp(d(m, \{v^*(m, e)\}))}{\sum_{e_i \in \mathcal{E}'} \exp(d(m, \{v^*(m, e_i)\}))}\right]$$

2.3 Heuristic Searching for Inference

The approximation in Equation 4 obviously can not reveal the matching distance because $v^*(m, e)$ contains insufficient information for retrieval. We want to search for a better view $Q^{'} \subset \mathcal{V}$ that $d(m, Q^{'}) < d(m, v^*(m, e))$.

Combining views (Q_1, Q_2) that contain complementary information is more likely to incorporate richer information into the newly assembled view. Considering two sets $Q_1 \subset \mathcal{V}$ and $Q_2 \subset \mathcal{V}$ and a distance metric $d(Q1, Q2) = \|f(t, Q_1) - f(t, Q_2)\|$, where t is the title of the entity and f represents the entity encoder, the most distant pair of views (Q_1, Q_2) achieve the largest magnitude on $d(Q_1, Q_2)$ among all pairs and is interpreted as the pair of views with less shared information. For each iteration, We search the top-k distant pairs (Q_1, Q_2) to form a new view $Q^{'} = Q_1 \cup Q_2$ and expand $Q^{'}$ into $\mathcal{V}$ to encode the merged Q' by $f(t, Q^{'})$ to produce a new representation for the involved entity. Searching and merging are performed iteratively until $|\mathcal{V}|$ reaches the maximal allowable value or the number of iterations reaches the preset value. During the inference, we precompute and cache the representations of views and select the view with minimal distance to m.

3 Experiments

3.1 Datasets

We evaluate MuVER under two different knowledge bases: Wikia, which the Zero-shot EL dataset is built upon, and Wikipedia, which contains 5.9M entities. We select one in-domain dataset, AIDA-CoNLL (Hoffart et al., 2011), and two out-of-domain datasets, WNED-WIKI (Guo and Barbosa, 2018) and MSNBC (Cucerzan, 2007), from standard EL datasets to validate MuVER in the full Wikipedia setting. Statistics of datasets are listed in Appendix A.1.

3.2 KB: Wikia

Logeswaran et al. (2019) constructs a zero-shot entity linking dataset (ZESHEL), which places more

Method	R@1	R@2	R@4	R@8	R@16	R@32	R@50	R@64
BM25	-	-	-	-	-	-	-	69.13
BLINK(Wu et al., 2020)	-	-	-	-	-	-	-	82.06
Partalidou et al. (2021)	-	-	-	-	-	-	84.28	-
BLINK (Wu et al., 2020)†	**46.51**	58.22	67.00	72.77	77.29	81.03	83.38	84.78
BLINK (Wu et al., 2020)*	45.59	57.55	66.10	72.47	77.65	81.69	84.31	85.56
SOM (Zhang and Stratos, 2021)	-	-	-	-	-	-	-	87.62
MuVER (w/o Heuristic Search)	43.49	58.56	68.78	75.87	81.33	85.86	88.35	89.52
MuVER	45.40	**60.84**	**71.26**	**78.27**	**83.19**	**87.58**	**89.75**	**90.84**

Table 1: Recall@k (R@k) on the test set of ZESHEL to retrieve entities from Wikia. †We reproduce BLINK and achieve a higher result compared with the result reported in the paper. * expands context length to 512. For SOM, we report the performance using in-batch negatives to have a fair comparison.

	AIDA-b			MSNBC			WNED-WIKI		
	R@10	R@30	R@100	R@10	R@30	R@100	R@10	R@30	R@100
BLINK	92.38	94.87	96.63	93.03	95.46	96.76	93.47	95.46	**97.76**
MuVER	**94.53**	**95.25**	**98.11**	**95.02**	**96.62**	**97.75**	**94.05**	**95.78**	97.34

Table 2: Results on three standard Entity Linking datasets. We test our model under the setting that only descriptions of entities are available. The number of basic views for each entity is 5.

emphasis on understanding the unstructured descriptions of entities to resolve the ambiguity of mentions on four unseen domains.

Concretely, MuVER uses BERT-base for f and g to make a fair comparison with previous works. We adopt an adam optimizer with a small learning rate 1e^{-5} and 10% warmup steps. We use batched random negatives and set the batch size to 128. The max number of context tokens is 128 and the max number of view tokens equals 40. Training 20 epochs takes one hour on 8 Tesla-v100 GPUs.

We compare MuVER with previous baselines in Table 1. Since MuVER is not limited by the length of descriptions, we add another baseline to extend BLINK to have 512 tokens (which is the max number of tokens for BERT-base). As shown in the table, we exceed BLINK by 5.28% and outperform SOM by 3.22% on Recall@64. We observe that Recall@1 of MuVER is lower than BLINK and the heuristic searching method can alleviate this problem. Detailed results on unseen domains are listed in Appendix A.3.

Effect of Heuristic Search We compare two distance-based merging strategies: taking closer or farther pairs of views to merge. We find out that merging views whose sentences are adjacent to each other in the original unstructured descriptions is a computationally efficient way to select the combined views. Table 3 shows that as the number of views increases, MuVER yields higher-quality candidates while the opposite strategy is troubled to provide more valuable views. Besides, our method can be regarded as a generalized form of SOM (Zhang and Stratos, 2021) and BLINK (Wu et al., 2020), which contain 128 views and one view correspondingly. SOM computes the similarity between mentions and tokens in descriptions, which stores 128 embeddings for each entity. Compared with SOM, MuVER reduces the number of views to a smaller size with improved quality, which is more efficient and effective.

Without View Merging		
Methods	# of Views	Recall@64
BLINK	1	85.56
SOM	128	87.62
MuVER	15.33	89.52

With View Merging			
Methods	# of Views	Distant Pairs	Close Pairs
MuVER	21.07	90.15	89.95
	26.18	90.51	89.99
	28.39	90.66	89.98
	30.48	90.79	89.89
	32.48	90.84	89.92

Table 3: Recall@64 on ZESHEL with varying number of views. We shot different merging strategies and "Distant Pairs" refers to our Heuristic Search method.

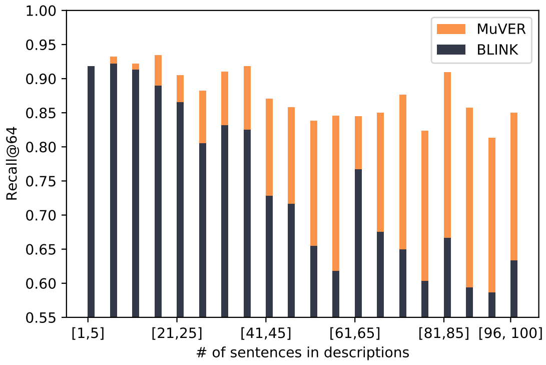

Figure 2: Recall@64 differences between BLINK and MuVER on entities with 1 to 100 sentences in their descriptions. We partition the entities by the number of sentences in entity descriptions and calculate metrics within each bin. The size for each bin is 5.

Effect on entities with long descriptions As shown in Figure 2, existing EL systems (like BLINK) obtain passable performance on entities with short descriptions but fail to manage those well-populated entities as the length of descriptions increases. For instance, the error rate of BLINK is 7.79% for entities with 5-10 sentences but 39.91% for entities with 75-80 sentences, which is more likely to contain various aspects for the entity. MuVER demonstrates its superiority over entities with long descriptions, which significantly reduces the error rate to 17.65% (-22.06%) for entities with 75-80 sentences while maintains the performance on entities with short descriptions, which achieves the error rate of 6.78% (-1.01%) for entities with 5-10 sentences.

3.3 KB: Wikipedia

We test AIDA-B, MSNBC and WNED-WIKI on the version of Wikipedia dump provided in KILT (Petroni et al., 2021), which contains 5.9M entities. Implementation details are listed in Appendix A.2. BLINK performance on these datasets is reported in its official Github repository[2]. We report the In-KB accuracy in Table 2 and observe that MuVER out-performs BLINK on all datasets except the recall@100 on WNED-WIKI.

4 Related Work

Representing each entity with a fixed-sized vector has been a common approach in Entity Linking. Ganea and Hofmann (2017) defines a word-entity conditional distribution and samples positive words from it. The representations of those positive words aim to approximate the entity embeddings compared with random words. Yamada et al. (2016) models the relatedness between entities into entity representations. NTEE (Yamada et al., 2017) trains entity representations by predicting the relevant entities for a given context in DBPedia abstract corpus. Ling et al. (2020) and Yamada et al. (2020) pre-train variants of the transformer-based model by maximizing the consistency between the context of the mentions and the corresponding entities. Those entity representations suffer from a cold-start problem that they cannot link mentions to unseen entities.

Another line of work is to generate entity representations using entity textual information, such as entity descriptions. Logeswaran et al. (2019) introduces an EL dataset in the zero-shot scenario to place more emphasis on reading entity descriptions. BLINK (Wu et al., 2020) proposes a bi-encoder to encode the descriptions and enhance the bi-encoder by distilling the knowledge from the cross-encoder. Yao et al. (2020) repeats the position embedding to solve the long-range modeling problem in entity descriptions. Zhang and Stratos (2021) demonstrates that hard negatives can enhance the contrast when training an EL model.

5 Conclusion

In this work, we propose a novel approach to construct multi-view representations from descriptions, which shows promising results on four EL datasets. Extensive results demonstrate the effectiveness of multi-view representations and the heuristic search strategy. In the future, we will explore more reliable and efficient approaches to construct views.

Acknowledgement

This work was supported by Alibaba Group through Alibaba Innovative Research Program and was also supported by the Key Research and Development Program of Zhejiang Province, China (No. 2021C01013), the National Key Research and Development Project of China (No. 2018AAA0101900), the Chinese Knowledge Center of Engineering Science and Technology (CKCEST) and MOE Engineering Research Center of Digital Library. We thank Gongfan Fang, Yongliang Shen for their valuable comments on writing this paper.

[2] https://github.com/facebookresearch/BLINK

References

Alexandr Andoni and Piotr Indyk. 2008. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. *Commun. ACM*, 51(1):117–122.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, Raffaele Perego, and Salvatore Trani. 2013. Learning relatedness measures for entity linking. In *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*, pages 139–148. ACM.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *EMNLP-CoNLL 2007, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28-30, 2007, Prague, Czech Republic*, pages 708–716. ACL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Zheng Fang, Yanan Cao, Qian Li, Dongjie Zhang, Zhenyu Zhang, and Yanbing Liu. 2019. Joint entity linking with deep reinforcement learning. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 438–447. ACM.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2619–2629. Association for Computational Linguistics.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego García-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 528–537. Association for Computational Linguistics.

Zhaochen Guo and Denilson Barbosa. 2018. Robust named entity disambiguation with random walks. *Semantic Web*, 9(4):459–479.

Xianpei Han, Le Sun, and Jun Zhao. 2011. Collective entity linking in web text: a graph-based method. In *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*, pages 765–774. ACM.

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 782–792. ACL.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, pages 1–1.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1595–1604. Association for Computational Linguistics.

Phong Le and Ivan Titov. 2019. Boosting entity linking performance by leveraging unlabeled documents. In *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 1935–1945. Association for Computational Linguistics.

Jeffrey Ling, Nicholas FitzGerald, Zifei Shan, Livio Baldini Soares, Thibault Févry, David Weiss, and Tom Kwiatkowski. 2020. Learning cross-context entity representations from text. *CoRR*, abs/2001.03765.

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 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.

Eleni Partalidou, Despina Christou, and Grigorios Tsoumakas. 2021. Improving zero-shot entity retrieval through effective dense representations. *CoRR*, abs/2103.04156.

Maria Pershina, Yifan He, and Ralph Grishman. 2015. Personalized page rank for named entity disambiguation. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 238–243. The Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick S. H. Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*. Association for Computational Linguistics.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

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 2020, Online, November 16-20, 2020*, pages 6397–6407. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6442–6454. Association for Computational Linguistics.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016*, pages 250–259. ACL.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2017. Learning distributed representations of texts and entities from knowledge base. *Trans. Assoc. Comput. Linguistics*, 5:397–411.

Yi Yang, Ozan Irsoy, and Kazi Shefaet Rahman. 2018. Collective entity disambiguation with structured gradient tree boosting. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 777–786. Association for Computational Linguistics.

Zonghai Yao, Liangliang Cao, and Huapu Pan. 2020. Zero-shot entity linking with efficient long range sequence modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2517–2522. Association for Computational Linguistics.

Wenzheng Zhang and Karl Stratos. 2021. Understanding hard negatives in noise contrastive estimation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021*. Association for Computational Linguistics.

Stefan Zwicklbauer, Christin Seifert, and Michael Granitzer. 2016. Robust and collective entity disambiguation through semantic embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, SIGIR 2016, Pisa, Italy, July 17-21, 2016*, pages 425–434. ACM.

A Appendix

A.1 Statistics of datasets

Table 5 shows statistics for four entity linking datasets: AIDA, MSNBC, WNED-WIKI and ZESHEL. MSNBC and WNED-WIKI are two out-of-domain test sets, which are evaluated on the model trained or finetuned on AIDA-train.

Dataset		Mention Num	KB	Entity Num
AIDA	Train	18448	Wiki-pedia	5903530
	Valid(A)	4791		
	Test(B)	4485		
MSNBC		656		
WNED-WIKI		6821		
ZESHEL	Train	49275	Wikia	332632
	Valid	10000		89549
	Test	10000		70140

Table 5: Statistics of four EL datasets.

A.2 Implementation Details

ZESHEL We have reported the best-performing hyperparameter in Section 3.2. Here we show the search bounds for the hyperparameters. We perform grid search on learning rate, weight decay, warmup ratio and batch size:

- Learning rate: [5e^{-6}, 1e^{-5}, 2e^{-5}, 5e^{-5}]
- Weight decay: [0.1, 0.01, 0.001]
- Warmup ratio: [0, 0.1]
- Batch size: [32, 64, 128, 196]

AIDA We finetune MuVER based on the EL model released by BLINK, which is pretrained on 9M annotated mention-entity pairs. Unlike the experiments on ZESHEL that adopting in-batch random negatives to train our model, we add hard negatives in batch. Due to the vast size of entities in Wikipedia, randomly sampled negatives are always too simple for the model to extract semantic features, thus degrading performance. We finetune our model on AIDA-CoNLL train set for one epoch. Batch size is set to 8. We add 3 hard negatives for each mention into the random in-batch negatives, which are precomputed using BLINK. The number of views is 5 for each entity and we choose the first 5 paragraphs with first 40 tokens, which are more likely to be summarizations. Other hyper-parameters are consistent with configurations on ZESHEL.

Parameters for MuVER Since MuVER has two BERT encoders, it has twice the number of parameters as BERT, which are listed in Table 6.

Model	Number of parameters
MuVER (base)	220M
MuVER (large)	680M

Table 6: Numbers of parameters for MuVER. MuVER (base) is used in ZESHEL and MuVER (large) is used in datasets under full Wikipedia setting.

A.3 Performance on Unseen Domains

In Table 4, we compare MuVER with BLINK on four unseen domains on ZESHEL. We observe a significant improvement on all four unseen domains, especially on Yugioh, which achieves +11.35 points on Recall@64. Furthermore, MuVER can reach comparative performance with BLINK's top-64 candidates by retrieving around 16-32 candidates, which reduces the computational cost for entity ranking.

Domain	**Method**	**R@1**	**R@2**	**R@4**	**R@8**	**R@16**	**R@32**	**R@50**	**R@64**
Forgotten Realms	BLINK	63.75	74.83	82.17	85.50	89.08	91.17	92.83	93.75
	MuVER	62.5	78.5	86.67	90.92	93.58	96.00	96.75	97.00
Lego	BLINK	50.04	65.39	75.81	81.65	84.82	88.41	90.58	91.83
	MuVER	50.46	68.81	78.32	84.4	88.82	91.91	93.33	93.74
Star Trek	BLINK	49.28	60.07	68.87	74.26	78.94	82.47	84.62	85.88
	MuVER	47.95	62.17	71.28	77.45	82.40	86.87	89.19	90.32
Yugioh	BLINK	35.66	47.45	56.14	63.22	68.35	73.00	75.90	77.71
	MuVER	34.32	50.06	63.25	72.61	78.48	83.94	86.69	88.26

Table 4: Recall@k on four unseen domains: Forgotten Realms, Lego, Star Trek and Yugioh.