Towards Better Entity Linking with Multi-View Enhanced Distillation

Yi Liu[1,2,*] **Yuan Tian**[3], **Jianxun Lian**[3], **Xinlong Wang**[3], **Yanan Cao**[1,2,†]
Fang Fang[1,2], **Wen Zhang**[3], **Haizhen Huang**[3], **Denvy Deng**[3], **Qi Zhang**[3]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]School of Cyber Security, University of Chinese Academy of Sciences
[3]Microsoft

{liuyi1999,caoyanan,fangfang0703}@iie.ac.cn
{yuantian,jialia,xinlongwang,zhangw,hhuang,dedeng,qizhang}@microsoft.com

Abstract

Dense retrieval is widely used for entity linking to retrieve entities from large-scale knowledge bases. Mainstream techniques are based on a dual-encoder framework, which encodes mentions and entities independently and calculates their relevances via rough interaction metrics, resulting in difficulty in explicitly modeling multiple mention-relevant parts within entities to match divergent mentions. Aiming at learning entity representations that can match divergent mentions, this paper proposes a **M**ulti-**V**iew Enhanced **D**istillation (MVD) framework, which can effectively transfer knowledge of multiple fine-grained and mention-relevant parts within entities from cross-encoders to dual-encoders. Each entity is split into multiple views to avoid irrelevant information being over-squashed into the mention-relevant view. We further design cross-alignment and self-alignment mechanisms for this framework to facilitate fine-grained knowledge distillation from the teacher model to the student model. Meanwhile, we reserve a global-view that embeds the entity as a whole to prevent dispersal of uniform information. Experiments show our method achieves state-of-the-art performance on several entity linking benchmarks[1].

1 Introduction

Entity Linking (EL) serves as a fundamental task in Natural Language Processing (NLP), connecting mentions within unstructured contexts to their corresponding entities in a Knowledge Base (KB). EL usually provides the entity-related data foundation for various tasks, such as KBQA (Ye et al., 2022), Knowledge-based Language Models (Liu et al., 2020) and Information Retrieval (Li et al., 2022). Most EL systems consist of two stages: entity retrieval (candidate generation), which retrieves

---

Figure 1: The illustration of two types of entities. Mentions in contexts are in **bold**, key information in entities is highlighted in color. The information in the first type of entity is relatively consistent and can be matched with a corresponding mention. In contrast, the second type of entity contains diverse and sparsely distributed information, can match with divergent mentions.

a small set of candidate entities corresponding to mentions from a large-scale KB with low latency, and entity ranking (entity disambiguation), which ranks those candidates using a more accurate model to select the best match as the target entity. This paper focuses on the entity retrieval task, which poses a significant challenge due to the need to retrieve targets from a large-scale KB. Moreover, the performance of entity retrieval is crucial for EL systems, as any recall errors in the initial stage can have a significant impact on the performance of the latter ranking stage (Luan et al., 2021).

Recent advancements in pre-trained language models (PLMs) (Kenton and Toutanova, 2019) have led to the widespread use of dense retrieval technology for large-scale entity retrieval (Gillick et al., 2019; Wu et al., 2020). This approach typically adopts a dual-encoder architecture that embeds the textual content of mentions and entities independently into fixed-dimensional vectors (Karpukhin et al., 2020) to calculate their relevance

scores using a lightweight interaction metric (e.g., dot-product). This allows for pre-computing the entity embeddings, enabling entities to be retrieved through various fast nearest neighbor search techniques (Johnson et al., 2019; Jayaram Subramanya et al., 2019).

The primary challenge in modeling relevance between an entity and its corresponding mentions lies in explicitly capturing the mention-relevant parts within the entity. By analyzing the diversity of intra-information within the textual contents of entities, we identify two distinct types of entities, as illustrated in Figure 1. Entities with uniform information can be effectively represented by the dual-encoder; however, due to its single-vector representation and coarse-grained interaction metric, this framework may struggle with entities containing divergent and sparsely distributed information. To alleviate the issue, existing methods construct multi-vector entity representations from different perspectives (Ma et al., 2021; Zhang and Stratos, 2021; Tang et al., 2021). Despite these efforts, all these methods rely on coarse-grained entity-level labels for training and lack the necessary supervised signals to select the most relevant representation for a specific mention from multiple entity vectors. As a result, their capability to effectively capture multiple fine-grained aspects of an entity and accurately match mentions with varying contexts is significantly hampered, ultimately leading to suboptimal performance in dense entity retrieval.

In order to obtain fine-grained entity representations capable of matching divergent mentions, we propose a novel Multi-View Enhanced Distillation (MVD) framework. MVD effectively transfers knowledge of multiple fine-grained and mention-relevant parts within entities from cross-encoders to dual-encoders. By jointly encoding the entity and its corresponding mentions, cross-encoders enable the explicit capture of mention-relevant components within the entity, thereby facilitating the learning of fine-grained elements of the entity through more accurate soft-labels. To achieve this, our framework constructs the same multi-view representation for both modules by splitting the textual information of entities into multiple fine-grained views. This approach prevents irrelevant information from being over-squashed into the mention-relevant view, which is selected based on the results of cross-encoders.

We further design cross-alignment and self-alignment mechanisms for our framework to separately align the original entity-level and fine-grained view-level scoring distributions, thereby facilitating fine-grained knowledge transfer from the teacher model to the student model. Motivated by prior works (Xiong et al., 2020; Zhan et al., 2021; Qu et al., 2021; Ren et al., 2021), MVD jointly optimizes both modules and employs an effective hard negative mining technique to facilitate transferring of hard-to-train knowledge in distillation. Meanwhile, we reserve a global-view that embeds the entity as a whole to prevent dispersal of uniform information and better represent the first type of entities in Figure 1.

Through extensive experiments on several entity linking benchmarks, including ZESHEL, AIDA-B, MSNBC, and WNED-CWEB, our method demonstrates superior performance over existing approaches. The results highlight the effectiveness of MVD in capturing fine-grained entity representations and matching divergent mentions, which significantly improves entity retrieval performance and facilitates overall EL performance by retrieving high-quality candidates for the ranking stage.

2 Related Work

To accurately and efficiently acquire target entities from large-scale KBs, the majority of EL systems are designed in two stages: entity retrieval and entity ranking. For entity retrieval, prior approaches typically rely on simple methods like frequency information (Yamada et al., 2016), alias tables (Fang et al., 2019) and sparse-based models (Robertson et al., 2009) to retrieve a small set of candidate entities with low latency. For the ranking stage, neural networks had been widely used for calculating the relevance score between mentions and entities (Yamada et al., 2016; Ganea and Hofmann, 2017; Fang et al., 2019; Kolitsas et al., 2018).

Recently, with the development of PLMs (Kenton and Toutanova, 2019; Lewis et al., 2020), PLM-based models have been widely used for both stages of EL. Logeswaran et al. (2019) and Yao et al. (2020) utilize the cross-encoder architecture that jointly encodes mentions and entities to rank candidates, Gillick et al. (2019) employs the dual-encoder architecture for separately encoding mentions and entities into high-dimensional vectors for entity retrieval. BLINK (Wu et al., 2020) improves overall EL performance by incorporating both architectures in its retrieve-then-rank pipeline,

making it a strong baseline for the task. GERENE (De Cao et al., 2020) directly generates entity names through an auto-regressive approach.

To further improve the retrieval performance, various methods have been proposed. Zhang and Stratos (2021) and Sun et al. (2022) demonstrate the effectiveness of hard negatives in enhancing retrieval performance. Agarwal et al. (2022) and GER (Wu et al., 2023) construct mention/entity centralized graph to learn the fine-grained entity representations. However, being limited to the single vector representation, these methods may struggle with entities that have multiple and sparsely distributed information. Although Tang et al. (2021) and MuVER (Ma et al., 2021) construct multi-view entity representations and select the optimal view to calculate the relevance score with the mention, they still rely on the same entity-level supervised signal to optimize the scores of different views within the entity, which limit the capacity of matching with divergent mentions.

In contrast to existing methods, MVD is primarily built upon the knowledge distillation technique (Hinton et al., 2015), aiming to acquire fine-grained entity representations from cross-encoders to handle diverse mentions. To facilitate fine-grained knowledge transfer of multiple mention-relevant parts, MVD splits the entity into multiple views to avoid irrelevant information being squashed into the mention-relevant view, which is selected by the more accurate teacher model. This Framework further incorporates cross-alignment and self-alignment mechanisms to learn mention-relevant view representation from both original entity-level and fine-grained view-level scoring distributions, these distributions are derived from the soft-labels generated by the cross-encoders.

3 Methodology

3.1 Task Formulation

We first describe the task of entity linking as follows. Give a mention m in a context sentence $s = <c_l, m, c_r>$, where c_l and c_r are words to the left/right of the mention, our goal is to efficiently obtain the entity corresponding to m from a large-scale entity collection $\varepsilon = \{e_1, e_2, ..., e_N\}$, each entity $e \in \varepsilon$ is defined by its title t and description d as a generic setting in neural entity linking (Ganea and Hofmann, 2017). Here we follow the two-stage paradigm proposed by (Wu et al., 2020): 1) retrieving a small set of candidate enti-

ties $\{e_1, e_2, ..., e_K\}$ corresponding to mention m from ε, where $K \ll N$; 2) ranking those candidates to obtain the best match as the target entity. In this work, we mainly focus on the first-stage retrieval.

3.2 Encoder Architecture

In this section, we describe the model architectures used for dense retrieval. Dual-encoder is the most adopted architecture for large-scale retrieval as it separately embeds mentions and entities into high-dimensional vectors, enabling offline entity embeddings and efficient nearest neighbor search. In contrast, the cross-encoder architecture performs better by computing deeply-contextualized representations of mention tokens and entity tokens, but is computationally expensive and impractical for first-stage large-scale retrieval (Reimers and Gurevych, 2019; Humeau et al., 2019). Therefore, in this work, we use the cross-encoder only during training, as the teacher model, to enhance the performance of the dual-encoder through the distillation of relevance scores.

3.2.1 Dual-Encoder Architecture

Similar to the work of (Wu et al., 2020) for entity retrieval, the retriever contains two-tower PLM-based encoders $\text{Enc}_m(\cdot)$ and $\text{Enc}_e(\cdot)$ that encode mention and entity into single fixed-dimension vectors independently, which can be formulated as:

$$E(m) = \text{Enc}_m([\text{CLS}] \, c_l \, [\text{M}_s] \, m \, [\text{M}_e] \, c_r \, [\text{SEP}])$$
$$E(e) = \text{Enc}_e([\text{CLS}] \, t \, [\text{ENT}] \, d \, [\text{SEP}])$$
$$(1)$$

where m, c_l, c_r, t, and d are the word-piece tokens of the mention, the context before and after the mention, the entity title, and the entity description. The special tokens $[\text{M}_s]$ and $[\text{M}_e]$ are separators to identify the mention, and $[\text{ENT}]$ serves as the delimiter of titles and descriptions. $[\text{CLS}]$ and $[\text{SEP}]$ are special tokens in BERT. For simplicity, we directly take the $[\text{CLS}]$ embeddings $E(m)$ and $E(e)$ as the representations for mention m and entity e, then the relevance score $s_{de}(m, e)$ can be calculated by a dot product $s_{de}(m, e) = E(m) \cdot E(e)$.

3.2.2 Cross-Encoder Architecture

Cross-encoder is built upon a PLM-based encoder $\text{Enc}_{ce}(\cdot)$, which concatenates and jointly encodes mention m and entity e (remove the $[\text{CLS}]$ token in the entity tokens), then gets the $[\text{CLS}]$ vectors as their relevance representation $\text{E}(m, e)$, finally fed

it into a multi-layer perceptron (MLP) to compute the relevance score $s_{ce}(m, e)$.

3.2.3 Multi-View Based Architecture

With the aim to prevent irrelevant information being over-squashed into the entity representation and better represent the second type of entities in Figure 1, we construct multi-view entity representations for the entity-encoder $\text{Enc}_e(\cdot)$. The textual information of the entity is split into multiple fine-grained **local-views** to explicitly capture the key information in the entity and match mentions with divergent contexts. Following the settings of MuVER (Ma et al., 2021), for each entity e, we segment its description d into several sentences $d^t(t = 1, 2, .., n)$ with NLTK toolkit [2], and then concatenate with its title t as the t-th view $e^t(t = 1, 2, .., n)$:

$$E(e^t) = \text{Enc}_e([\text{CLS}]\, t\, [\text{ENT}]\, d^t\, [\text{SEP}]) \quad (2)$$

Meanwhile, we retain the original entity representation $E(e)$ defined in Eq. (1) as the **global-view e^0 in inference**, to avoid the uniform information being dispersed into different views and better represent the first type of entities in Figure 1. Finally, the relevance score $s(m, e_i)$ of mention m and entity e_i can be calculated with their multiple embeddings. Here we adopt a max-pooler to select the view with the highest relevant score as the **mention-relevant view**:

$$
\begin{aligned}
s(m, e_i) &= \max_t \{s(m, e_i^t)\} \\
&= \max_t \{E(m) \cdot E(e^t)\}
\end{aligned}
\quad (3)
$$

3.3 Multi-View Enhanced Distillation

The basic intuition of MVD is to accurately transfer knowledge of multiple fine-grained views from a more powerful cross-encoder to the dual-encoder to obtain mention-relevant entity representations. First, in order to provide more accurate relevance between mention m and each view $e^t(t = 1, 2, ..., n)$ of the entity e as a supervised signal for distillation, we introduce a multi-view based cross-encoder following the formulation in Sec 3.2.3:

$$E(m, e^t) = \text{Enc}_{ce}([\text{CLS}]\, m_{\text{enc}}\, [\text{SEP}]\, e_{\text{enc}}^t\, [\text{SEP}]) \quad (4)$$

where m_{enc} and $e_{\text{enc}}^t(t = 1, 2, .., n)$ are the word-piece tokens of the mention and entity representations defined as in Eq. (1) and (2), respectively.

We further design cross-alignment and self-alignment mechanisms to separately align the original entity-level scoring distribution and fine-grained view-level scoring distribution, in order to facilitate the fine-grained knowledge distillation from the teacher model to the student model.

Cross-alignment In order to learn entity-level scoring distribution among candidate entities at the multi-view scenario, we calculate the relevance score $s(m, e_i)$ for mention m and candidate entity e_i in candidates $\{e_1, e_2, ..., e_K\}$ by all its views $\{e_i^1, e_i^2, ..., e_i^n\}$, the indexes of relevant views i_{de} and i_{ce} for dual-encoder and cross-encoder are as follows:

$$
\begin{aligned}
i_{de} &= \arg\max_t \{s_{de}(m, e_i^t)\} \\
i_{ce} &= \arg\max_t \{s_{ce}(m, e_i^t)\}
\end{aligned}
\quad (5)
$$

here to avoid the mismatch of relevant views (i.e., $i_{de} \neq i_{ce}$), we **align their relevant views** based on the index i_{ce} of max-score view in cross-encoder, the loss can be measured by KL-divergence as

$$\mathcal{L}_{cross} = \sum_{i=1}^{K} \tilde{s}_{ce}(m, e_i) \cdot log \frac{\tilde{s}_{ce}(m, e_i)}{\tilde{s}_{de}(m, e_i)} \quad (6)$$

where

$$
\begin{aligned}
\tilde{s}_{de}(m, e_i) &= \frac{e^{s_{de}(m, e_i^{ice})}}{e^{s_{de}(m, e_i^{ice})} + \sum_{j \neq i} e^{s_{de}(m, e_j^{jce})}} \\
\tilde{s}_{ce}(m, e_i) &= \frac{e^{s_{ce}(m, e_i^{ice})}}{e^{s_{ce}(m, e_i^{ice})} + \sum_{j \neq i} e^{s_{ce}(m, e_j^{jce})}}
\end{aligned}
\quad (7)
$$

here $\tilde{s}_{de}(m, e_i)$ and $\tilde{s}_{ce}(m, e_i)$ denote the probability distributions of the entity-level scores which are represented by the i_{ce}-th view over all candidate entities.

Self-alignment Aiming to learn the view-level scoring distribution within each entity for better distinguishing relevant view from other irrelevant views, we calculate the relevance score $s(m, e^t)$ for mention m and each view $e_i^t(t = 1, 2, ..., n)$ of entity e_i, the loss can be measured by KL-divergence as:

$$\mathcal{L}_{self} = \sum_{i=1}^{K} \sum_{t=1}^{n} \tilde{s}_{ce}(m, e_i^t) \cdot log \frac{\tilde{s}_{ce}(m, e_i^t)}{\tilde{s}_{de}(m, e_i^t)} \quad (8)$$
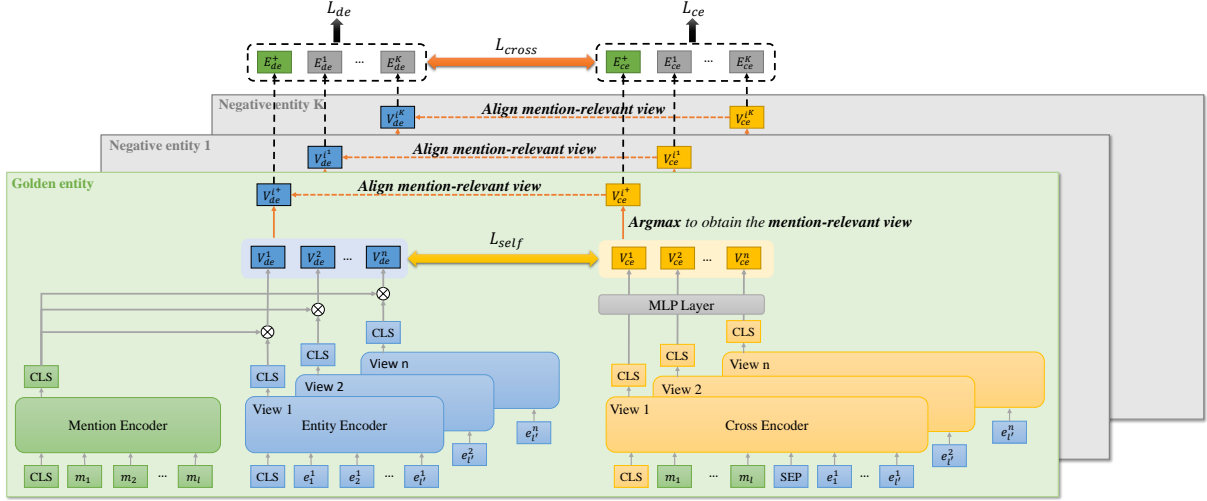
Figure 2: The general framework of Multi-View Enhanced Distillation (MVD). V_{de}^i and V_{ce}^i are the relevance scores between m and e^i separately calculated by dual-encoder and cross-encoder, E_{de} and E_{ce} are the entity relevance scores represented by V_{de}^i and V_{ce}^i, base on the max-score view's index **i** in cross-encoder.

where

$$\tilde{s}_{de}(m, e_i^t) = \frac{e^{s_{de}(m,e_i^t)}}{e^{s_{de}(m,e_i^t)} + \sum_{j \neq t} e^{s_{de}(m,e_i^j)}}$$

$$\tilde{s}_{ce}(m, e_i^t) = \frac{e^{s_{ce}(m,e_i^t)}}{e^{s_{ce}(m,e_i^t)} + \sum_{j \neq t} e^{s_{ce}(m,e_i^j)}} \quad (9)$$

here $\tilde{s}_{de}(m, e_i^t)$ and $\tilde{s}_{ce}(m, e_i^t)$ denote the probability distributions of the view-level scores over all views within each entity.

Joint training The overall joint training framework can be found in Figure 2. The final loss function is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{de} + \mathcal{L}_{ce} + \alpha \mathcal{L}_{cross} + \beta \mathcal{L}_{self} \quad (10)$$

Here, $\mathcal{L}_{cross}$ and $\mathcal{L}_{self}$ are the knowledge distillation loss with the cross-encoder and defined as in Eq. (6) and (8) respectively, α and β are coefficients for them. Besides, $\mathcal{L}_{de}$ and $\mathcal{L}_{ce}$ are the supervised training loss of the dual-encoder and cross-encoder on the labeled data to maximize the $s(m, e_k)$ for the golden entity e_k in the set of candidates $\{e_1, e_2, ..., e_K\}$, the loss can be defined as:

$$\mathcal{L}_{de} = -s_{de}(m, e_k) + log \sum_{j=1}^{K} \exp(s_{de}(m, e_j))$$

$$\mathcal{L}_{ce} = -s_{ce}(m, e_k) + log \sum_{j=1}^{K} \exp(s_{ce}(m, e_j))$$

$$\quad (11)$$

Inference we only apply the mention-encoder to obtain the mention embeddings, and then retrieve targets directly from pre-computing view embeddings via efficient nearest neighbor search. These view embeddings encompass both global and local views and are generated by the entity-encoder following joint training. Although the size of the entity index may increase due to the number of views, the time complexity can remain sub-linear with the index size due to mature nearest neighbor search techniques (Zhang et al., 2022).

3.4 Hard Negative Sampling

Hard negatives are effective information carriers for difficult knowledge in distillation. Mainstream techniques for generating hard negatives include utilizing static samples (Wu et al., 2020) or top-K dynamic samples retrieved from a recent iteration of the retriever (Xiong et al., 2020; Zhan et al., 2021), but these hard negatives may not be suitable for the current model or are pseudo-negatives (i.e., unlabeled positives) (Qu et al., 2021). Aiming to mitigate this issue, we adopt a simple negative sampling method that first retrieves top-N candidates, then randomly samples K negatives from them, which reduces the probability of pseudo-negatives and improves the generalization of the retriever.

4 Experiments

4.1 Datasets

We evaluate MVD under two distinct types of datasets: three standard EL datasets AIDA-CoNLL

Method	R@1	R@2	R@4	R@8	R@16	R@32	R@50	R@64
BM25	-	-	-	-	-	-	-	69.26
BLINK (Wu et al., 2020)	-	-	-	-	-	-	-	82.06
Partalidou et al. (2022)	-	-	-	-	-	-	84.28	-
BLINK*	45.59	57.55	66.10	72.47	77.65	81.69	84.31	85.56
SOM (Zhang and Stratos, 2021)	-	-	-	-	-	-	-	89.62
MuVER (Ma et al., 2021)	43.49	58.56	68.78	75.87	81.33	85.86	88.35	89.52
Agarwal et al. (2022)	50.31	61.04	68.34	74.26	78.40	82.02	-	85.11
GER (Wu et al., 2023)	42.86	-	66.48	73.00	78.11	82.15	84.41	85.65
MVD (ours)	**52.51**	**64.77**	**73.43**	**79.74**	**84.35**	**88.17**	**90.43**	**91.55**

Table 1: **Recall@K(R@K)** on test set of ZESHEL, **R@K** measures the percentage of mentions for which the top-K retrieved entities include the golden entities. The best results are shown in **bold** and the results unavailable are left blank. * is reproduced by Ma et al. (2021) that expands context length to 512.

Method	AIDA-b			MSNBC			WNED-CWEB		
	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	82.23	86.09	88.68
MuVER	94.53	95.25	98.11	95.02	96.62	97.75	<u>79.31</u>	<u>83.94</u>	<u>88.15</u>
MVD (ours)	**97.05**	**98.15**	**98.80**	**96.74**	**97.71**	**98.04**	**85.01**	**88.18**	**91.11**

Table 2: **Recall@K(R@K)** on test set of Wikipedia datasets, best results are shown in **bold**. Underline notes for the results we reproduce.

(Hoffart et al., 2011), WNED-CWEB (Guo and Barbosa, 2018) and MSNBC (Cucerzan, 2007), these datasets are all constructed based on a uniform Wikipedia KB; and a more challenging Wikia-based dataset ZESHEL (Logeswaran et al., 2019), adopts a unique setup where the train, valid, and test sets correspond to different KBs. Statistics of these datasets are listed in Appendix A.1.

4.2 Training Procedure

The training pipeline of MVD consists of two stages: Warmup training and MVD training. In the Warmup training stage, we separately train dual-encoder and cross-encoder by in-batch negatives and static negatives. Then we initialize the student model and the teacher model with the well-trained dual-encoder and cross-encoder, and perform multi-view enhanced distillation to jointly optimize the two modules following Section 3.3. Implementation details are listed in Appendix A.2.

4.3 Main Results

Compared Methods We compare MVD with previous state-of-the-art methods. These methods can be divided into several categories according to the representations of entities: BM25 (Robertson et al.,

2009) is a sparse retrieval model based on exact term matching. BLINK (Wu et al., 2020) adopts a typical dual-encoder architecture that embeds the entity independently into a single fixed-size vector. SOM (Zhang and Stratos, 2021) represents entities by its tokens and computes relevance scores via the sum-of-max operation (Khattab and Zaharia, 2020). Similar to our work, MuVER (Ma et al., 2021) constructs multi-view entity representations to match divergent mentions and achieved the best results, so we select MuVER as the main compared baseline. Besides, ARBORESCENCE (Agarwal et al., 2022) and GER (Wu et al., 2023) construct mention/entity centralized graphs to learn fine-grained entity representations.

For Zeshel dataset we compare MVD with all the above models. As shown in Table 1, MVD performs better than all the existing methods. Compared to the previously best performing method MuVER, MVD significantly surpasses in all metrics, particularly in **R@1**, which indicates the ability to directly obtain the target entity. This demonstrates the effectiveness of MVD, which uses hard negatives as information carriers to explicitly transfer knowledge of multiple fine-grained views from the cross-encoder to better represent entities for

Model	R@1	R@64
MVD	**51.69**	**89.78**
- w/o multi-view cross-encoder	50.85	89.24
- w/o relevant-view alignment	51.02	89.55
- w/o self-alignment	51.21	89.43
- w/o cross-alignment	50.82	88.71
- w/o all components	51.40	84.16

Table 3: Ablation for fine-grained components in MVD on test set of ZESHEL. Results on Wikipedia-based datasets are similar and omitted due to limited space.

Method	R@1	R@64
MVD	**51.69**	**89.78**
- w/o dynamic distillation	51.11	88.50
- w/o dynamic negatives	50.26	88.46
- w/o all strategies	50.16	87.54

Table 4: Ablation for training strategies in MVD on test set of ZESHEL.

matching multiple mentions, resulting in higher-quality candidates for the ranking stage.

For Wikipedia datasets we compare MVD with BLINK [3] and MuVER (Ma et al., 2021). As shown in Table 2, our MVD framework also outperforms other methods and achieves state-of-the-art performance on AIDA-b, MSNBC, and WNED-CWEB datasets, which verifies the effectiveness of our method again in standard EL datasets.

4.4 Ablation and Comparative Studies

4.4.1 Ablation Study

For conducting fair ablation studies and clearly evaluating the contributions of each fine-grained component and training strategy in MVD, we exclude the coarse-grained global-view to evaluate the capability of transferring knowledge of multiple fine-grained views, and utilize Top-K dynamic hard negatives without random sampling to mitigate the effects of randomness on training.

Fine-grained components ablation results are presented in Table 3. When we replace the multi-view representations in the cross-encoder with the original single vector or remove the relevant view selection based on the results of the cross-encoder, the retrieval performance drops, indicat-

---

[3]BLINK performance is reported in `https://github.com/facebookresearch/BLINK`

Method	View Type	R@1	R@64
BLINK	global	46.04	**87.46**
MuVER	global	36.90	80.65
MVD (ours)	global	**47.11**	87.04
BLINK	local	37.20	86.38
MuVER	local	41.99	89.25
MVD (ours)	local	**51.27**	**90.25**
MVD (ours)	global+local	**52.51**	**91.55**

Table 5: Comparison for representing entities from multi-grained views on test set of ZESHEL. Results of BLINK and MuVER are reproduced by us.

ing the importance of providing accurate supervised signals for each view of the entity during distillation. Additionally, the removal of cross-alignment and self-alignment results in a decrease in performance, highlighting the importance of these alignment mechanisms. Finally, when we exclude all fine-grained components in MVD and employ the traditional distillation paradigm based on single-vector entity representation and entity-level soft-labels, there is a significant decrease in performance, which further emphasizes the effectiveness of learning knowledge of multiple fine-grained and mention-relevant views during distillation.

Training strategies we further explore the effectiveness of joint training and hard negative sampling in distillation, Table 4 shows the results. First, we examine the effect of joint training by freezing the teacher model's parameters to do a static distillation, the retrieval performance drops due to the teacher model's limitation. Similarly, the performance drops a lot when we replace the dynamic hard negatives with static negatives, which demonstrates the importance of dynamic hard negatives for making the learning task more challenging. Furthermore, when both training strategies are excluded and the student model is independently trained using static negatives, a substantial decrease in retrieval performance is observed, which validates the effectiveness of both training strategies in enhancing retrieval performance.

4.4.2 Comparative Study on Entity Representation

To demonstrate the capability of representing entities from multi-grained views, we carry out comparative analyses between MVD and BLINK (Wu et al., 2020), as well as MuVER (Ma et al., 2021).

Candidate Retriever	U.Acc.
Base Version Ranker	
BM25 (Logeswaran et al., 2019)	55.08
BLINK (Wu et al., 2020)	61.34
SOM (Zhang and Stratos, 2021)	65.39
Agarwal et al. (2022)	62.53
MVD (ours)	**66.85**
Large Version Ranker	
BLINK (Wu et al., 2020)	63.03
SOM (Zhang and Stratos, 2021)	67.14
MVD (ours)	**67.84**

Table 6: Performance of ranker based on different candidate retrievers on the test set of ZESHEL. **U.Acc.** means the unnormalized macro accuracy.

These systems are founded on the principles of coarse-grained global-views and fine-grained local-views, respectively.

We evaluate the retrieval performance of both entity representations and present the results in Table 5. The results clearly indicate that MVD surpasses both BLINK and MuVER in terms of entity representation performance, even exceeding BLINK's global-view performance in **R@1**, despite being a fine-grained training framework. Unsurprisingly, the optimal retrieval performance is attained when MVD employs both entity representations concurrently during the inference process.

5   Further Analysis

5.1   Facilitating Ranker's Performance

To evaluate the impact of the quality of candidate entities on overall performance, we consider two aspects: candidates generated by different retrievers and the number of candidate entities used in inference. First, we separately train BERT-base and BERT-large based cross-encoders to rank the top-64 candidate entities retrieved by MVD. As shown in Table 6, the ranker based on our framework achieves the best results in the two-stage performance compared to other candidate retrievers, demonstrating its ability to generate high-quality candidate entities for the ranking stage.

Additionally, we study the impact of the number of candidate entities on overall performance, as shown in Figure 3, with the increase of candidates number k, the retrieval performance grows steadily while the overall performance is likely to
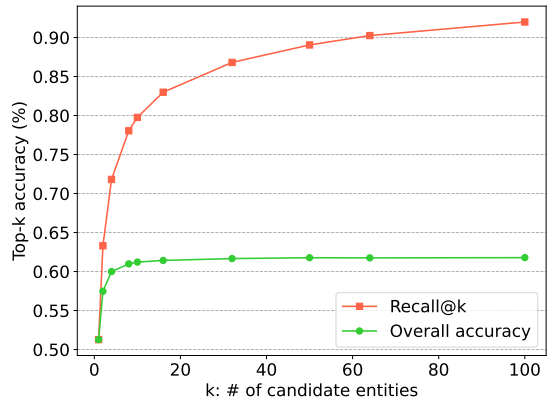

Figure 3: Recall and overall micro accuracy based on different number of candidates k.

be stagnant. This indicates that it's ideal to choose an appropriate k to balance the efficiency and efficacy, we observe that $k = 16$ is optimal on most of the existing EL benchmarks.

5.2   Qualitative Analysis

To better understand the practical implications of fine-grained knowledge transfer and global-view entity representation in MVD, as shown in Table 7, we conduct comparative analysis between our method and MuVER (Ma et al., 2021) using retrieval examples from the test set of ZESHEL for qualitative analysis.

In the first example, MVD clearly demonstrates its ability to accurately capture the mention-relevant information *Rekelen were members of this movement* and *professor Natima Lang* in the golden entity "Cardassian dissident movement". In contrast, MuVER exhibits limited discriminatory ability in distinguishing between the golden entity and the hard negative entity "Romulan underground movement". In the second example, Unlike MuVER which solely focuses on local information within the entity, MVD can holistically model multiple mention-relevant parts within the golden entity "Greater ironguard" through a global-view entity representation, enabling matching with the corresponding mention "improved version of lesser ironguard".

6   Conclusion

In this paper, we propose a novel Multi-View Enhanced Distillation framework for dense entity retrieval. Our framework enables better representation of entities through multi-grained views, and by using hard negatives as information car-

Mention and Context	Entity retrieved by MVD	Entity retrieved by MuVER
Rekelen was a member of the **underground movement** and a student under professor Natima Lang. In 2370, Rekelen was forced to flee Cardassia prime because of her political views.	Title: Cardassian dissident movement The Cardassian dissident movement was a resistance movement formed to resist and oppose the Cardassian Central Command and restore the authority of the Detapa Council. They believed this change was critical for the future of their people. **Professor Natima Lang, Hogue, and Rekelen were members of this movement in the late 2360s and 2370s.** ...	Title: Romulan underground movement The Romulan underground movement was formed sometime prior to the late 24th century on the planet Romulus by a group of Romulan citizens who opposed the Romulan High Command and who supported a Romulan - Vulcan reunification. **Its methods and principles were similar to those of the Cardassian dissident movement which emerged in the Cardassian Union around the same time.** ...
Known as the **improved version of lesser ironguard**, this spell granted the complete immunity from all common, un-enchanted metals to the caster or one creature touched by the caster.	Title: Greater ironguard Greater ironguard was an arcane abjuration spell that temporarily **granted one creature immunity from all non-magical metals and some enchanted metals. It was an improved version of ironguard**. The effects of this spell were the same as for "lesser ironguard" except that it also granted immunity and transparency to metals that had been enchanted up to a certain degree. ...	Title: Lesser ironguard ... **after an improved version was developed, this spell became known as lesser ironguard**. Upon casting this spell, the caster or one creature touched by the caster became completely immune to common, unenchanted metal. metal weapons would pass through the individual without causing harm. likewise, the target of this spell could pass through metal barriers such as iron bars, grates, or portcullises. ...

Table 7: Examples of entities retrieved by MVD and MuVER, mentions in contexts and mention-relevant information in entities are in **bold**.

riers to effectively transfer knowledge of multiple fine-grained and mention-relevant views from the more powerful cross-encoder to the dual-encoder. We also design cross-alignment and self-alignment mechanisms for this framework to facilitate the fine-grained knowledge distillation from the teacher model to the student model. Our experiments on several entity linking benchmarks show that our approach achieves state-of-the-art entity linking performance.

Limitations

The limitations of our method are as follows:

- We find that utilizing multi-view representations in the cross-encoder is an effective method for MVD, however, the ranking performance of the cross-encoder may slightly decrease. Therefore, it is sub-optimal to directly use the cross-encoder model for entity ranking.

- Mention detection is the predecessor task of our retrieval model, so our retrieval model will be affected by the error of the mention detection. Therefore, designing a joint model of mention detection and entity retrieval is an improvement direction of our method.

Acknowledgements

References

Dhruv Agarwal, Rico Angell, Nicholas Monath, and Andrew McCallum. 2022. Entity linking via explicit mention-mention coreference modeling. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4644–4658.

Silviu Cucerzan. 2007. Large-scale named entity disambiguation based on wikipedia data. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 708–716.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

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*, pages 438–447.

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*, pages 2619–2629.

Dan Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 528–537.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

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*, pages 782–792.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Suhas Jayaram Subramanya, Fnu Devvrit, Harsha Vardhan Simhadri, Ravishankar Krishnawamy, and Rohan Kadekodi. 2019. Diskann: Fast accurate billion-point nearest neighbor search on a single node. *Advances in Neural Information Processing Systems*, 32.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

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*, pages 39–48.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Xiangsheng Li, Jiaxin Mao, Weizhi Ma, Zhijing Wu, Yiqun Liu, Min Zhang, Shaoping Ma, Zhaowei Wang, and Xiuqiang He. 2022. A cooperative neural information retrieval pipeline with knowledge enhanced automatic query reformulation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 553–561.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

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.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Xinyin Ma, Yong Jiang, Nguyen Bach, Tao Wang, Zhongqiang Huang, Fei Huang, and Weiming Lu. 2021. Muver: Improving first-stage entity retrieval with multi-view entity representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2617–2624.

Eleni Partalidou, Despina Christou, and Grigorios Tsoumakas. 2022. Improving zero-shot entity retrieval through effective dense representations. In *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*, pages 1–5.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 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*, pages 2523–2544.

Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang. 2021. Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5835–5847.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Kai Sun, Richong Zhang, Samuel Mensah, Yongyi Mao, and Xudong Liu. 2022. A transformational biencoder with in-domain negative sampling for zero-shot entity linking. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1449–1458.

Hongyin Tang, Xingwu Sun, Beihong Jin, and Fuzheng Zhang. 2021. A bidirectional multi-paragraph reading model for zero-shot entity linking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13889–13897.

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.

Taiqiang Wu, Xingyu Bai, Weigang Guo, Weijie Liu, Siheng Li, and Yujiu Yang. 2023. Modeling fine-grained information via knowledge-aware hierarchical graph for zero-shot entity retrieval. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1021–1029.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

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*, pages 250–259.

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*, pages 2517–2522.

Xi Ye, Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, and Caiming Xiong. 2022. RNG-KBQA: Generation augmented iterative ranking for knowledge base question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6032–6043, Dublin, Ireland. Association for Computational Linguistics.

Jingtao Zhan, Jiaxin Mao, Yiqun Liu, Jiafeng Guo, Min Zhang, and Shaoping Ma. 2021. Optimizing dense retrieval model training with hard negatives. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1503–1512.

Shunyu Zhang, Yaobo Liang, Ming Gong, Daxin Jiang, and Nan Duan. 2022. Multi-view document representation learning for open-domain dense retrieval. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5990–6000.

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*, pages 1090–1101.

A Appendix

A.1 Statistics of Datasets

Table 8 shows statistics for ZESHEL dataset, which was constructed based on documents in Wikia from 16 domains, 8 for train, 4 for valid, and 4 for test.

Domain	#Entity	#Mention
Training		
American Football	31929	3898
Doctor Who	40281	8334
Fallout	16992	3286
Final Fantasy	14044	6041
Military	104520	13063
Pro Wrestling	10133	1392
Star Wars	87056	11824
World of Warcraft	27677	1437
Training	**332632**	**49275**
Validation		
Coronation Street	17809	1464
Muppets	21344	2028
Ice Hockey	28684	2233
Elder Scrolls	21712	4275
Validation	**89549**	**10000**
Testing		
Forgotten Realms	15603	1200
Lego	10076	1199
Star Trek	34430	4227
YuGiOh	10031	3374
Testing	**70140**	**10000**

Table 8: Statistics of ZESHEL dataset.

Table 9 shows statistics for three Wikipedia-based datasets: AIDA, MSNBC, and WNED-CWEB. MSNBC and WNED-CWEB are two out-of-domain test sets, which are evaluated on the model trained on AIDA-train, and we test them on the version of Wikipedia dump provided in KILT (Petroni et al., 2021), which contains 5.9M entities.

Dataset	#Mention	#Entity
AIDA-train	18448	
AIDA-valid	4791	
AIDA-test	4485	5903530
MSNBC	678	
WNED-CWEB	10392	

Table 9: Statistics of three Wikipedia-based datasets.

A.2 Implementation Details

For ZESHEL, we use the BERT-base to initialize both the student dual-encoder and the teacher cross-encoder. For Wikipedia-based datasets, we finetune our model based on the model released by BLINK, which is pre-trained on 9M annotated mention-entity pairs with BERT-large. All experiments are performed on 4 A6000 GPUs and the results are the average of 5 runs with different random seeds.

Warmup training We initially train a dual-encoder using in-batch negatives, followed by training a cross-encoder as the teacher model via the top-k static hard negatives generated by the dual-encoder. Both models utilize multi-view entity representations and are optimized using the loss defined in Eq. (11), training details are listed in Table 10.

Hyperparameter	ZESHEL	Wikipedia
Dual-encoder		
Max mention length	128	32
Max view num	10	5
Max view length	40	40
Learning rate	1e-5	1e-5
Negative num	63	63
Batch size	64	64
Training epoch	40	40
Training time	4h	2h
Cross-encoder		
Max input length	168	72
Learning rate	2e-5	2e-5
Negative num	15	15
Batch size	1	1
Training epoch	3	3
Training time	7h	5h

Table 10: Hyperparameters for Warmup training.

MVD training Next, we initialize the student model and the teacher model with the well-trained dual-encoder and cross-encoder obtained from the Warmup training stage. We then employ multi-view enhanced distillation to jointly optimize both modules, as described in Section 3.3. To determine the values of α and β in Eq. (10), we conduct a grid search and find that setting $\alpha = 0.3$ and $\beta = 0.1$ yields the best performance. We further adopt a simple negative sampling method in Sec 3.4 that first retrieves top-N candidates and then samples K as negatives. Based on the analysis in Sec 5.1 that

16 is the optimal candidate number to cover most hard negatives and balance the efficiency, we set it as the value of K; then to ensure high recall rates and sampling high quality negatives, we search from a candidate list [50, 100, 150, 200, 300] and eventually determine N=100 is the most suitable value. The training details are listed in Table 11.

Hyperparameter	ZESHEL	Wikipedia
Max mention length	128	32
Max view num	10	5
Max view length	40	40
Max cross length	168	72
Learning rate	2e-5	2e-5
Negative num	15	15
Batch size	1	1
Training epoch	5	5
Training time	15h	6h

Table 11: Hyperparameters for MVD training.

Inference MVD employs both local-view and global-view entity representations concurrently during the inference process, details are listed in Table 12.

Hyperparameter	ZESHEL	Wikipedia
Local-view length	40	40
Global-view length	512	128
Avg view num	16	6

Table 12: Hyperparameters for Inference.