61 ACL 2023

ACL 2023投稿分享

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, Beijing, China

[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

[3]Microsoft Corporation, Beijing, China

{liuyi1999,caoyanan,fangfang0703}@iie.ac.cn

{yuantian,jianxun.lian,xinlongwang,zhangw,hhuang,dedeng,zhang.qi}@microsoft.com

中国科学院信息工程研究所

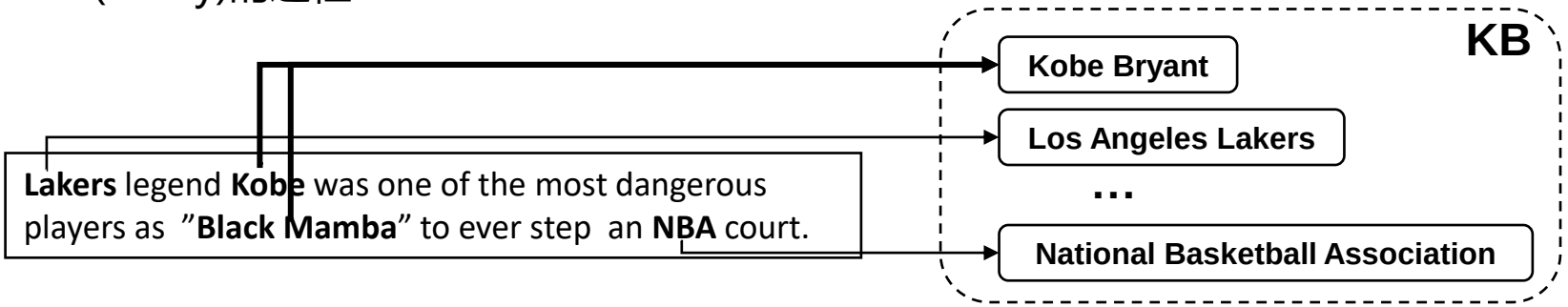
INSTITUTE OF INFORMATION ENGINEERING,CAS

1. 目录

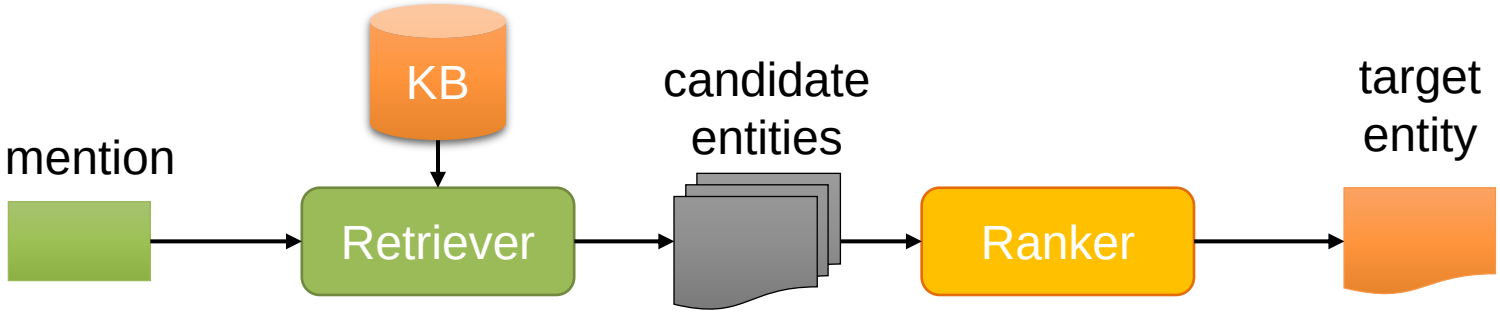
- 介绍（Introduction）
- 方法（Methodology）
- 实验（Experiment）
- Rebuttal

1. 介绍：研究任务

- 实体链接：将非结构化文本中的提及（mention)映射到知识库(KB)中对应的实体(entity)的过程



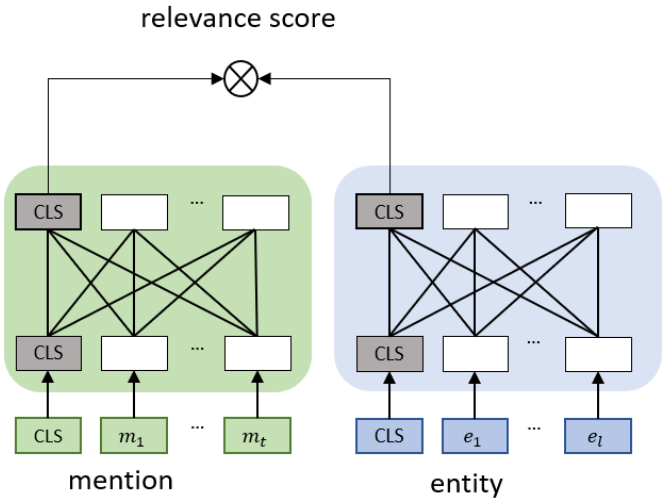
- 主流的实体链接系统遵循召回-排序的两阶段范式



- 本文主要关注实体召回，原因如下：
  - 由于需要从大规模知识库中准确地获取目标实体，召回会比排序更加具有挑战性
  - 召回阶段的性能严格限制着排序阶段的上限

1. 介绍：已有方法

- 主流技术：基于预训练语言模型（PLMs）的稠密检索技术
  - 使用双塔编码（dual-encoder）器结构
  - 将提及和实体的文本内容（提及上下文；实体描述）编码为单个的向量
  - 使用点积等轻量级的交互策略来建模相关性，满足大规模向量检索的需求



Dual-Encoder Architecture

1. 介绍：面临挑战

- 建模提及-实体间相关性的主要挑战：如何准确地捕获实体文本内容中与提及相关的部分
- 根据实体内信息分布，划分为两类：
  - 具有一致性且均匀分布信息的实体（匹配单一的提及）
  - 具有多样化且稀疏分布信息的实体（可以匹配多个不同的提及）
- 当前方法受限于单一向量的实体表征和粗粒度（点积）的交互策略，无法很好地表示第二类的实体

Entity 1: 2014 UEFA Champions League final
Description: Real Madrid won the match 4–1 after extra time, with goals from Cristiano Ronaldo, Gareth Bale, Marcelo and Sergio Ramos. In doing so, Real Madrid secured a record 10th title in the competition. As the winners, Real Madrid earned the right to play against 2013–14 UEFA Europa League winners Sevilla in the 2014 UEFA Super Cup.

Mention: Ronaldo fired home the penalty as Real Madrid won **Europe's biggest prize** for the 10th time in its history.

Entity 2: Cristiano Ronaldo
Description: ... Ronaldo has won five Ballon d'Or awards and four European Golden Shoes, he has won 32 trophies in his career, including seven league titles, five UEFA Champions Leagues, the UEFA European Championship and the UEFA Nations League. ... Ronaldo was cautioned by police for smashing a phone out of a 14-year-old boy's hand following his team's 1-0 Premier League defeat to Everton in April.

Mention: **Ronaldo** has amassed an unrivalled collection of records in the Champions League and EURO finals.

Mention: **Ronaldo has been banned** with improper conduct by the FA for smashing a teenage fan's phone.

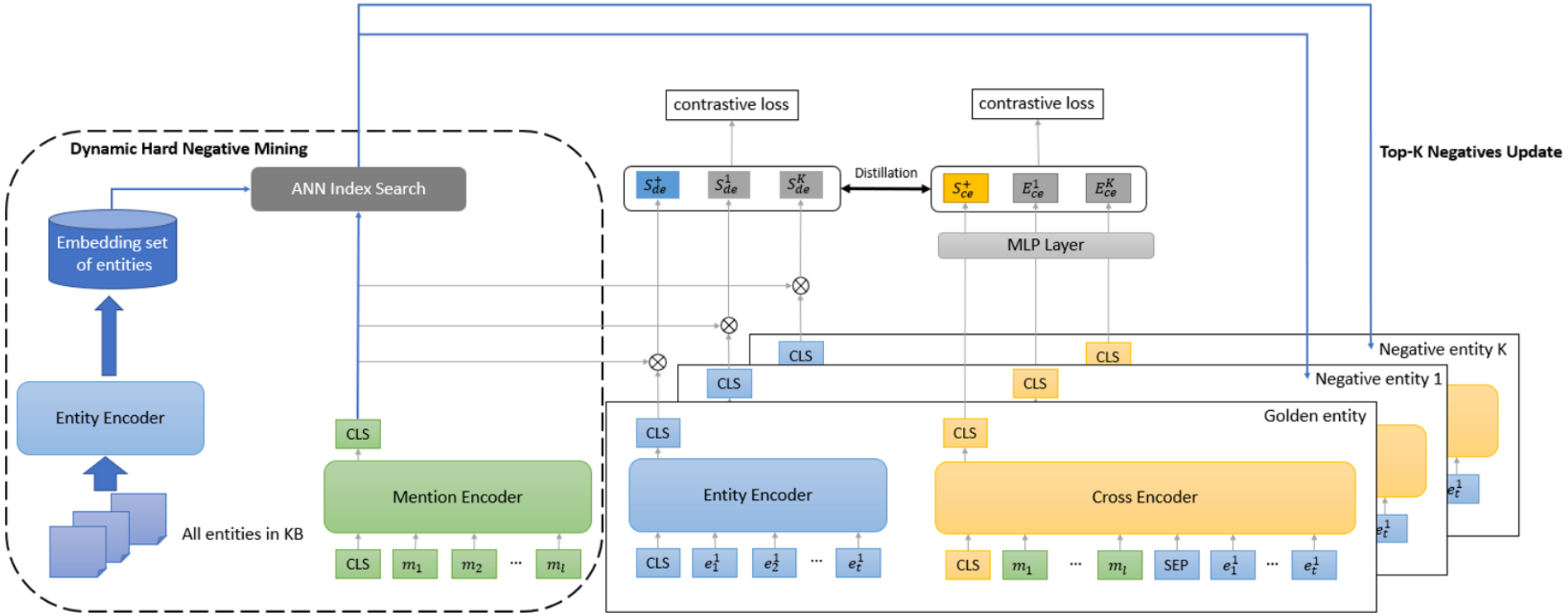
Figure 1: Illustration of two types of entities. Mentions in contexts are in bold, key information in entities is highlighted in color. The information in the first type of entities is relatively consistent and can be matched with a corresponding mention. In contrast, the second type of entities contains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

1. 介绍：创新思路

- 针对上述问题，为获得细粒度的，能匹配不同提及的实体表征（目的），我们提出了多视图增强的蒸馏框架MVD
- 多视图增强的蒸馏框架 (multi-View Enhanced Distillation): 将实体中**细粒度的、和提及相关的知识，**通过知识蒸馏有效地从更准确的交叉编码器（教师模型）转移到双编码器（学生模型）中
  - 多视图实体表征（multi-view entity representation）
  - 自对齐和交叉对齐机制（cross-alignment and self-alignment mechanisms）
    - 正负例实体之间原始的实体级分数分布
    - 实体内部细粒度的视图级分数分布
  - 训练策略
    - 联合优化学生模型和教师模型（joint training）
    - 动态强负例挖掘（dynamic hard negative mining）

2. 方法：训练策略

- 知识蒸馏：交叉编码器 → 双编码器，联合优化教师-学生模型
- 强负例挖掘：每轮迭代后，使用学生模型在KB中召回得到前K个候选实体（排除正例），作为新一轮次的负例
- 损失函数：正例-负例之间的对比学习损失



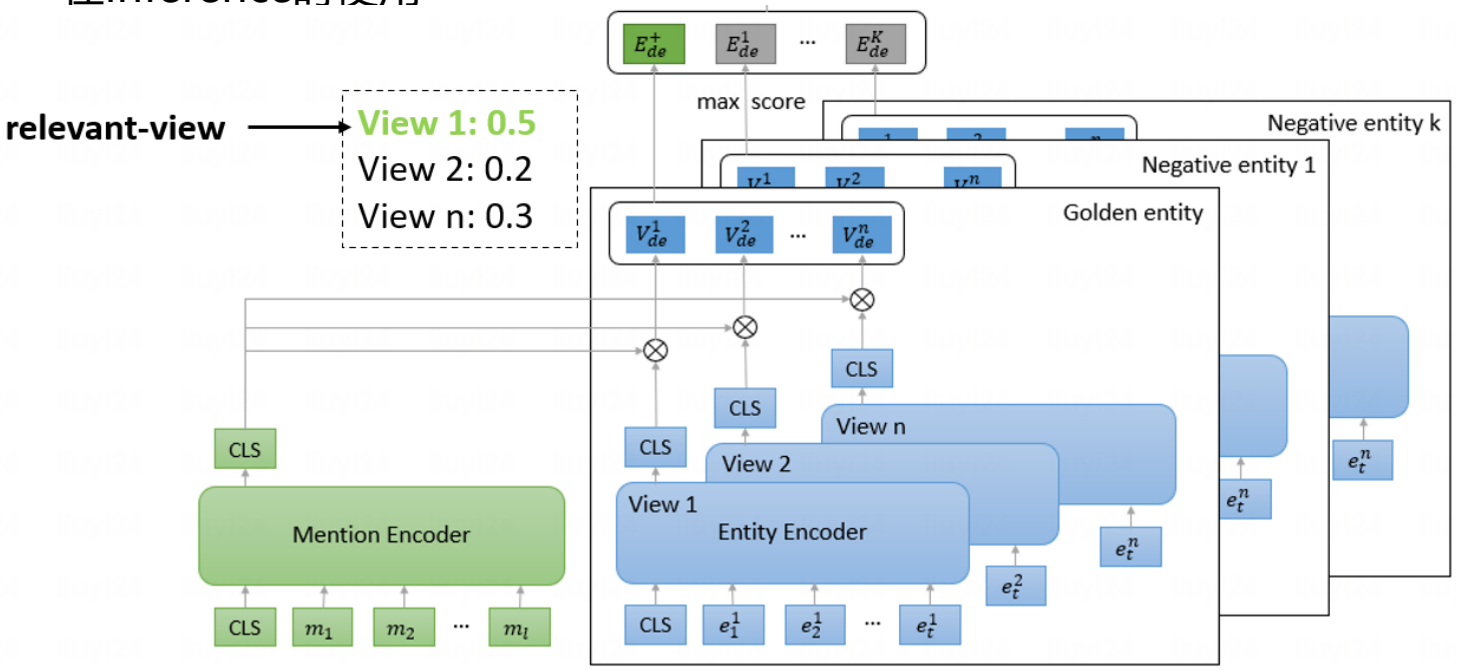
2. 方法：多视图实体表示

- 文档级实体描述 → 句子级的细粒度视图
- 建模提及-实体相关性时，实体由视图表示
- 期望在训练时引入监督信号，使选择的视图能准确地表征实体中提及相关部分的内容

Here we adopt a max-pooler to select the view with the highest relevant score as the **mention-relevant view**:

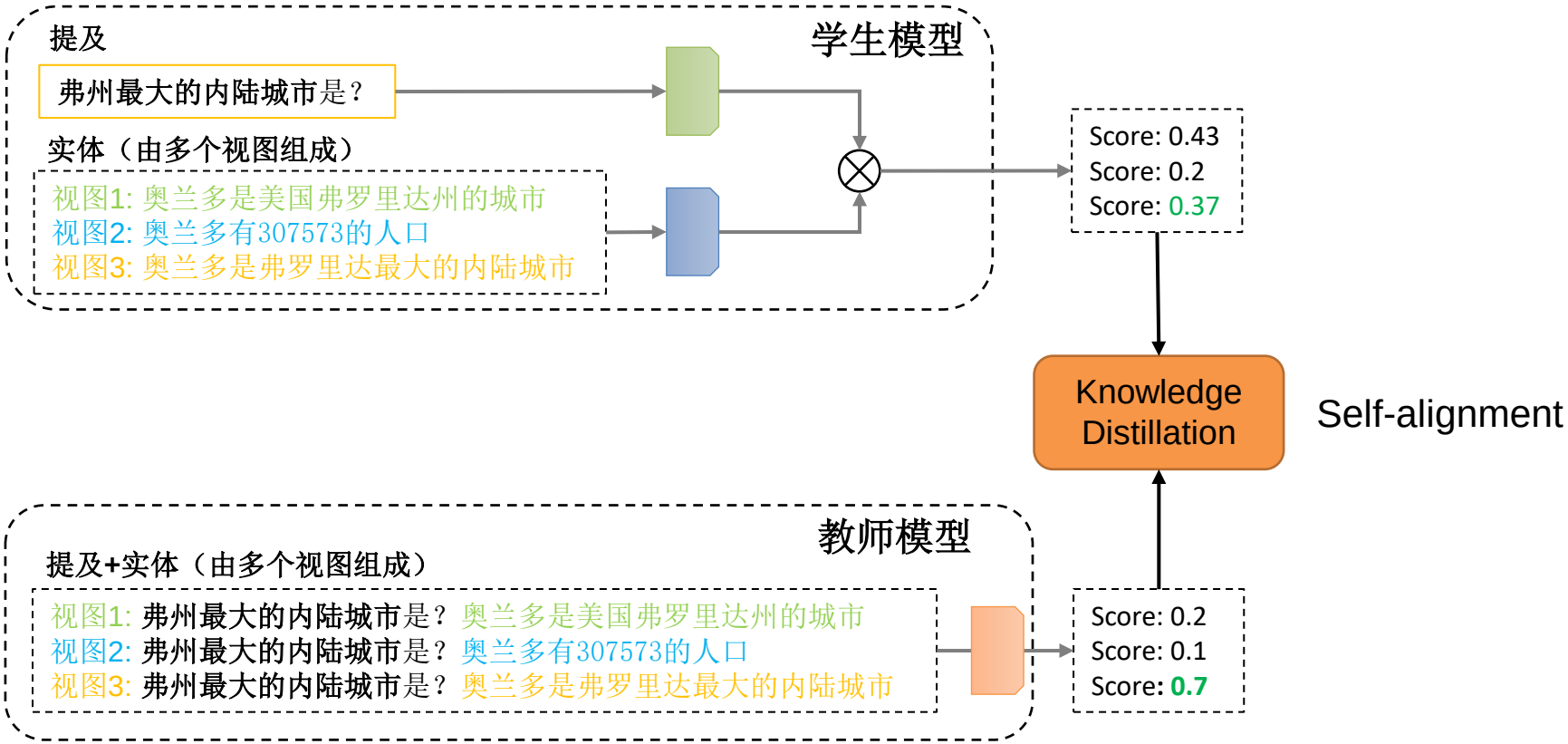
$$s(m, e_i) = \max_t \{s(m, e_i^t)\} = \max_t \{E(m) \cdot E(e^t)\} \quad (3)$$

- 保留一个粗粒度的，对实体全部文本内容建模的全局视图，该视图不参与训练，只在Inference时使用



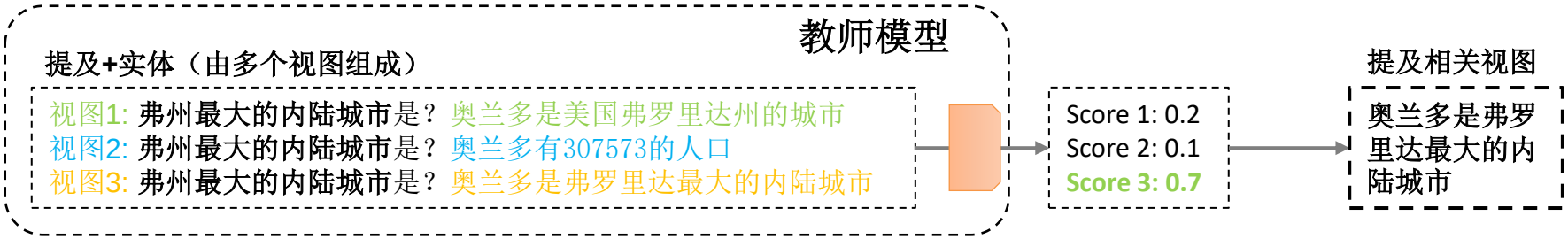
2. 方法：自对齐机制（Self-alignment）

- 基于教师模型产生的软标签监督信号，学习实体内部**细粒度视图**的相关性分数分布



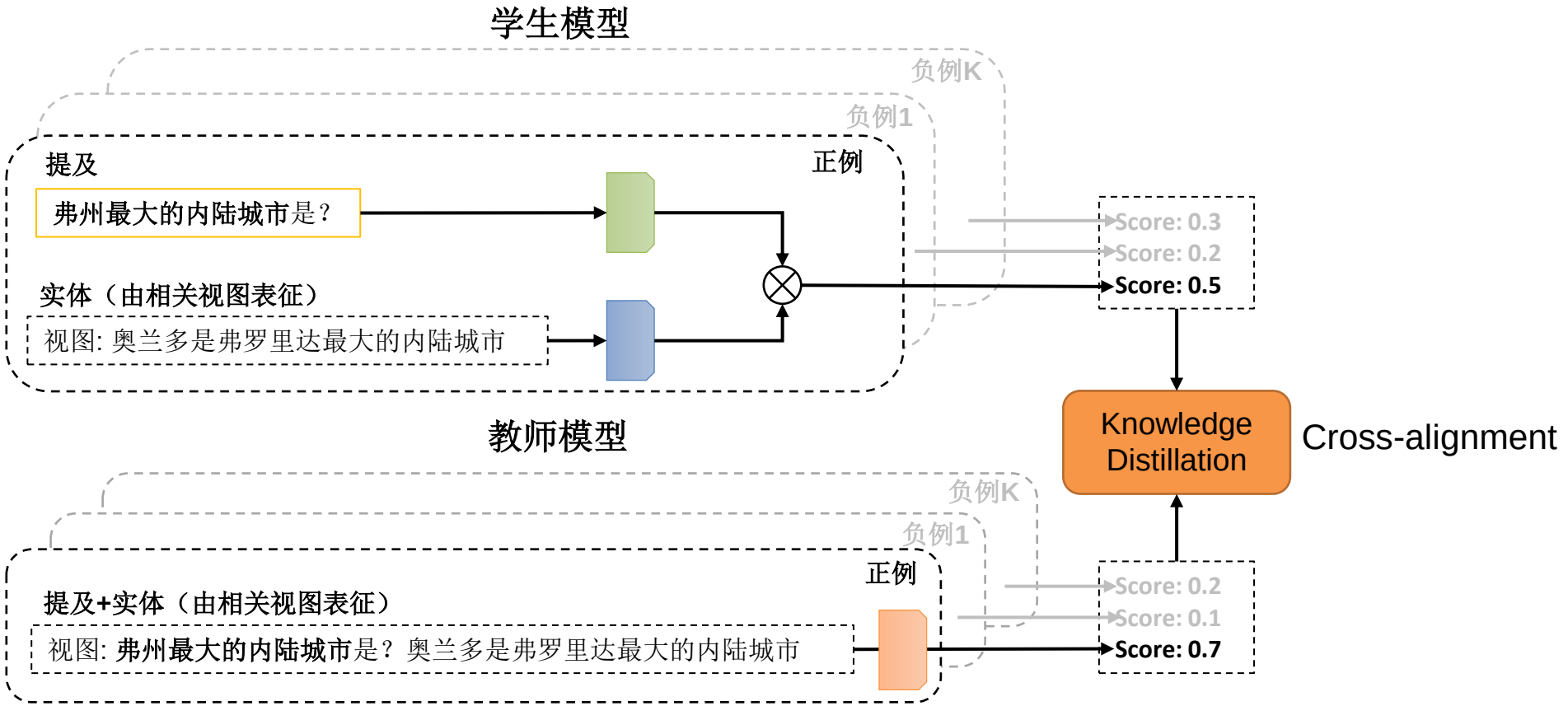
2. 方法：交叉对齐机制（Cross-alignment）

- 实体由提及相关视图（mention-relevant view）表示，该视图由教师模型选择作为监督信号



2. 方法：交叉对齐机制（Cross-alignment）

- 基于教师模型选择的相关视图表示实体，学习**正负例**实体间的相关性分数分布



3．主实验

- 三个基于相同KB（Wikipedia）的数据集：AIDA、MSNBC和WNED-CWEB
- 一个zero-shot设定的EL数据集ZESHEL

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 the test set of ZESHEL, which is the average of 5 runs with different random seeds. 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	79.31	83.94	88.15
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 the test set of Wikipedia datasets, best results are shown in bold. Underline notes for the results we reproduce.

3. 辅助实验：消融

- MVD中的细粒度组件 (view-alignment/cross-alignment/self-alignment)

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 here due to limited space.

- MVD中的训练策略（joint training/hard negative sampling）

Method	R@1	R@64
MVD	**51.69**	**89.78**
- w/o dynamic distillation	51.11	88.50
- w/o dynamic negatives	48.80	88.43
- w/o all strategies	47.36	87.24

Table 5: Ablation for training strategies on test set of ZESHEL.

3．辅助实验：对比

- MVD对实体进行全局建模和细粒度的局部建模的能力 (global-view/local-view)
  - 基于粗粒度的全局视图，进行实体表示学习的BLINK
  - 基于细粒度的局部视图，进行实体表示表示学习的MuVER

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

Table 4: Comparison for representing entities from multi-grained views on test set of ZESHEL. Results of BLINK and MuVER are reproduced by us.

4. Rebuttal

- Soundness：研究问题的深度/广度，实验效果与合理性，评估方法的有效性
- Excitement：研究问题与提出的方法是否新颖
- Score: 4/3.5 (accept), 4/3.5 (accept), 3/3 (borderline)

4. Rebuttal: Reasons to accept

- 新颖的训练框架（novel framework）
- SOTA的实验结果（strong experimental results）
- 可以应用到其他领域的任务上（applicability to other tasks）

4. Rebuttal: Reasons to reject

- W1: 对现有工作的增量贡献（incremental contribution，excitement）
- A1：解释我们方法和Baseline的区别（Motivation/Method）

- W2: 实验不够充分（more extensive analysis，soundness）
- A2：我们做过了，但这部分实验不是重点，且由于篇幅限制所以省去~

- W3: 实验结果不清晰，有的地方未明确说明（reproducibility may be improved， soundness ），Reviewer提出了如下的问题：

Questions for the Author(s)

(A) Is the "global-view" entity representation affected by the MVD training approach, or is it obtained after initial (warm up) training of the dual-encoder and then frozen?

(B) How are the "top-N candidates" (line 354) retrieved specifically? what values of N (and K) were considered in the experiments?

(C) Tables 1 and 4 report MVD R@1=52.51 and R@64=91.55, whereas Tables 3 and 5 report MVD R@1=51.69 and R@64=89.78. Why are these numbers different? Is this related to the inclusion/exclusion of the global view?

- A3：对没有写明白的实验补充说明即可

4. Rebuttal: Answer to W2

========== W2:

more extensive analysis of downstream EL performance.

Answer2: In Section 5.1, we examined the impact of MVD on downstream EL performance from two perspectives: the quality of candidates generated by different retrievers and the number of candidates used in inference. These experiments were conducted under the unnormalized setting, where the gold entity may not be included in the candidates generated by the retriever. To further evaluate EL performance, we also conducted experiments under the normalized setting, where the gold entity is among the candidates retrieved by BM25. However, due to space constraints and our primary focus on entity retrieval, as well as the more common unnormalized setting, we have excluded this section.

4. Rebuttal: Answer to W3

Why Tables 1 and 4 report different R@1/64 from Tables 3 and 5?

Answer C: The difference in the R@1/64 metrics is due to that we exclude both the (1) global-view and (2) random sampling negatives components in the setting of Tables 3 and 5. This exclusion is essential for conducting fair ablation studies and clearly evaluating the contributions of each fine-grained components in the MVD training framework. Specifically, the exclusion/replacement of the two components are for the following reasons:

- Excluded the coarse-grained global-view to evaluate the capability of transferring knowledge of multiple fine-grained views.
- Utilized Top-K dynamic hard negatives without random sampling to mitigate the effects of randomness on training.

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 here due to limited space.

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

Table 4: Comparison for representing entities from multi-grained views on test set of ZESHEL. Results of BLINK and MuVER are reproduced by us.

4. Post Rebuttal

- W1: 对现有工作的增量贡献（incremental contribution，excitement）
- Update：作者在回复中讨论了区别，但基础模型仍然相同只是训练方式不同，促使我对该论文的增量性质进行了确认

- W2: 实验不够充分（more extensive analysis，soundness）
- Update：由作者回复解决，虽然我想看到更全面的实验和分析，但理解篇幅限制的问题

- W3: 实验结果不清晰，有的地方未明确说明（reproducibility may be improved， soundness ）：
- Update：由作者回复解决，通过对提出问题的回复提供了更可靠的信息

- Post-rebuttal：I thank the authors for their informative response, which provides some clarifications w.r.t. the questions and concerns expressed in my original review. Overall, I confirm the evaluation scores in my original review

4. Meta Review

- There was unanimous agreement among the reviewers that the MVD framework is **novel** and **convincingly demonstrated state-of-the-art performance** on multiple benchmarks, making the work quite **sound**.
- Only one reason to reject persisted after the rebuttal period: a concern about the **incrementality of the improvement w/r/t past work (MuVER),** which the reviewer found to be foundationally quite similar to the MVD approach, **which they acknowledged MVD improved on.**
- The reviewers found the paper **moderately exciting**.

谢谢大家！

中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS