KGMEL: Knowledge Graph-Enhanced Multimodal Entity Linking (Online Appendix)

Anonymous Author(s)

A PROMPT TEMPLATES

This section provides the prompt templates used in KGMEL for generating triples and re-ranking candidate entities.

A.1 Triple Generation Prompt

Prompt template for triple generation of mentions P_{triple}:

> [IMAGE] Given the image and text [mention sentence], please generate triples for the entities list of mention word. following the steps below:
> **### Step 1: Entity Type**
> For each entity in [list of mention words], identify its type, following the format:
> - "entity_name": type of entity
>
> Type of entity can be : person, nationality, religious group, political group, organization, country, city, state, building, airport, highway, bridge, company, agency, institution, product, event, work of art, law, language, etc.
> **### Step 2: Entity Description**
> Provide a description for each entity, following the format:
> - "entity_name": entity description
>
> Focus on factual information that can be inferred from the image and text to describe the entity.
> **### Step 3: Triples**
> Using the type and information from steps 1 and 2, generate all possible triples for each entity in.
> Convert the entity types and information into triples using the format, with each triple on a new line:
> - "entity_name" | relation1 | entity1
> - "entity_name" | relation2 | entity2
>
> Based on the entity type and information provided in the image and text, choose the most relevant relations from the following list to generate triples:
> "instance of", "subclass of", "part of", "has characteristic", "field of work", "occupation", "sex or gender", "country of citizenship", "position held", "religion or worldview", "member of", "owner of", "country", "capital", "continent", "located in", "industry", "participant", "genre", "named after"

The triple generation prompt guides the VLM through a step-by-step process. First, for entity type identification, we reference the NER (Named Entity Recognition) categories from [17]. This identified type is further incorporated into a triple with the relation "instance of." Next, rather than generating triples directly, we use entity descriptions as a reasoning step. To better align with existing KG triples $\mathcal{T}_e$, we select 20 relation types based on their frequency within $\mathcal{T}_e$ and their semantic relevance to mention contexts, providing them as references. We generate triples for each mention sentence, processing all the mention words in the sentence one at a time. This approach considers the contextual relationships between mentions and is also efficient.

A.2 Re-ranking Prompt

Prompt template for zero-shot re-ranking P_{rerank} :

> Given the context below, please identify the most corresponding entity from the list of candidates.
> **Context:** [mention sentence]
> **Candidate Entities:**
> [K^{th} entity name] ([QID]) : [description]
> - Triple: {list of KG triples}
>
> . . .
> [1^{st} entity name] ([QID]) : [description]
> - Triple: [list of KG triples]
>
> **Context:** [mention sentence]
> **Target Entity:** [mention words]: [generated description]
> - Triple: [list of generated triples]
>
> First, read the context and the target entity. Then, review the candidate entities and their descriptions.
> From the candidate entities, select the *supporting triples* that align with the context and the target entity. (Note that triples may contain inconsistent information.)
> Based on the selected supporting triples, identify the most relevant entity that best matches the given sentence context.

The entity re-ranking prompt instructs LLMs to identify *supporting triples* from filtered entity triples $\mathcal{T}_e^{(filt)}$ and filtered mention triples $\mathcal{T}_m^{(filt)}$ to determine the most relevant entity match from candidates $C(m)$. The candidates are presented following the order from the candidate retrieval stage, but in reverse order, which we empirically found to improve performance. The prompt also leverages entity description and the generated mention descriptions from step 2 of the triple generation process as additional context.

B DATASET STATISTICS

We evaluate KGMEL on three MEL datasets: WikiDiverse [16], RichpediaMEL [15] and WikiMEL [15]. We use a subset of Wikidata as KB, following [15] and retrieve KG triples via SPARQL queries. Table 1 summarizes dataset statistics and retrieved KG triples.

C EXPERIMENTAL SETUP

This section presents the experimental setup, including the evaluation metrics for baselines, hyperparameter configurations for each dataset, and additional experimental results

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Statistics of three MEL datasets.

Datasets	WikiDiverse	RichpeidaMEL	WikiMEL
# sentences	7,405	17,724	22,070
# mentions	15,093	17,805	25,846
# KG triples	60,842,321	32,761,864	65,131,860
# candidate entities[†]	132,460	160,935	109,976
# total entities[‡]	776,407	831,737	761,343
# relation	1,322	1,288	1,289

[†] All entities in a subset of Wikidata KB used as candidates.
[‡] Includes candidate entities and all tail entities from retrieved KG triples.

C.1 Evaluation Metrics

We evaluate KGMEL using HITS@k and MRR, defined as

$$HITS@k = \frac{1}{N}\sum_{i} I(rank(i) < k), \tag{1}$$

$$MRR = \frac{1}{N}\sum_{i} \frac{1}{rank(i)}, \tag{2}$$

where N is the total number of test instances, $rank(i)$ denotes the rank of the correct entity for the i-th instance, and $I(\cdot)$ is an indicator function.

C.2 Baselines

We compare the performance of KGMEL with several baseline methods, which are grouped into two categories:
Retrieval-based Methods:

- **CLIP** [12] aligns visual and textual inputs using two transformer-based encoders trained on extensive image-text pairs with a contrastive loss.
- **ViLT** [4] employs shallow embeddings for text and images, emphasizing deep modality interactions through transformer layers.
- **ALBEF** [5] integrates visual and textual features via a multimodal transformer encoder, utilizing image-text contrastive loss and momentum distillation for improved learning from noisy data.
- **METER** [2] explores semantic relationships between modalities using a co-attention mechanism comprising self-attention, cross-attention, and feed-forward networks.
- **DZMNED** [10] is the first method for MEL, integrates visual features with word-level and character-level textual features using an attention mechanism.
- **JMEL** [1] extracts and fuses unigram and bigram textual embeddings, jointly learning mention and entity representations from both textual and visual contexts.
- **VELML** [19] utilizes VGG-16 for object-level visual features and a pre-trained BERT for text, combining them via an attention mechanism.
- **GHMFC** [15] employs hierarchical cross-attention to capture fine-grained correlations between text and images, optimized through contrastive learning.
- **MIMIC** [8] proposes a multi-grained multimodal interaction network that captures both global and local features from text and images, enhancing entity disambiguation through comprehensive intra- and inter-modal interactions.
- **OT-MEL** [18] addresses multimodal fusion and fine-grained matching by formulating correlation assignments between multimodal features and mentions as an optimal transport problem, with knowledge distillation.
- **MELOV** [14] optimizes visual features in a latent space by combining inter-modality and intra-modality enhancements, improving consistency between mentions and entities.
- **M3EL** [3] introduces a multi-level matching network for multimodal feature extraction, intra-modal matching, and bidirectional cross-modal matching, enabling comprehensive interactions within and between modalities.
- **FBMEL** [9] enhances the significance of visual prompts and fully leverages information in image-text pairs to improve entity linking accuracy.

Generative-based Methods:

- **GPT-3.5-turbo** [11] is a large language model (LLM), and we utilize the results reported by GEMEL.
- **LLaVA-13B** [6] is a vision-language model (VLM), and we utilize the results reported by GELR.
- **GEMEL** [13] leverages large language model (LLM) to directly generate target entity names, aligning visual features with textual embeddings through a feature mapper.
- **GELR** [7] enhances the generation process by incorporating knowledge retriever, improving accuracy through the retrieval of relevant context from a knowledge base.

C.3 Hyperparameter Settings

Table 2 shows the hyperparameter settings used for each dataset in our experiments.

Table 2: Hyperparameter settings.

Hyperparameter	WikiDiverse	RichpediaMEL	WikiMEL
β	0.5	0.5	0.5
$\tau_{\text{att}}, \tau_{\text{cl}}$	0.1	0.1	0.1
$\lambda_{MM}, \lambda_{EE}$	0.1	0.1	0.1
K	16	16	16
p	3	3	5
n	15	10	15

D EXTENDED EXPERIMENTAL RESULTS

Table 3 shows extended experimental results, including HITS@{1,3,5} and MRR. The results are reported as the mean ± standard deviation across three experimental runs. For the re-ranking results, the top-1 retrieved entity is replaced by the entity selected during the re-ranking stage, while the ranking order of the remaining candidates is preserved. We obtained the baseline results from [8], while the results for [3, 9, 14, 18] were sourced from the original papers. Results for [7, 13] were also taken from the original papers, along with those for [6, 11]. Additionally, we excluded the WikiDiverse results for these methods, as they use different data compared to the other baselines, following [3].

REFERENCES

[1] Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020. Multimodal entity linking for tweets. In *ECIR*.

Table 3: Evaluation results on three MEL datasets. H@k denotes Hits@k, MRR denotes Mean Reciprocal Rank. The best results are in bold; the second-best are underlined.

	WikiDiverse				RichpediaMEL				WikiMEL			
	H@1	H@3	H@5	MRR	H@1	H@3	H@5	MRR	H@1	H@3	H@5	MRR
CLIP [12]	61.21	79.63	85.18	71.69	67.78	85.22	90.04	77.57	83.23	92.10	94.51	88.23
ViLT [4]	34.39	51.07	57.83	45.22	45.85	62.96	69.80	56.63	72.64	84.51	87.86	79.46
ALBEF [5]	60.59	75.59	81.30	69.93	65.17	82.84	88.28	75.29	78.64	88.93	91.75	84.56
METER [2]	53.14	70.93	77.59	63.71	63.96	82.24	87.08	74.15	72.46	84.41	88.17	79.49
DZMNED [10]	56.90	75.34	81.41	67.59	68.16	82.94	87.33	76.63	78.82	90.02	92.62	84.97
JMEL [1]	37.38	54.23	61.00	48.19	48.82	66.77	73.99	60.06	64.65	79.99	84.34	73.39
VELML [19]	54.56	74.43	81.15	66.13	67.71	84.57	89.17	77.19	76.62	88.75	91.96	83.42
GHMFC [15]	60.27	79.40	84.74	70.99	72.92	86.85	90.60	80.76	76.55	88.40	92.01	83.36
MIMIC [8]	63.51	81.04	86.43	73.44	81.02	91.77	94.38	86.95	87.98	95.07	96.37	91.82
OT-MEL [18]	66.07	82.82	87.39	75.43	83.30	92.39	94.83	88.27	88.97	**95.63**	**96.96**	92.59
MELOV [14]	67.32	83.69	87.54	76.57	84.14	**92.81**	94.89	**88.80**	88.91	95.61	96.58	92.32
M3EL [3]	74.06	86.57	90.04	81.29	82.82	92.73	**95.34**	88.26	88.84	95.20	96.71	92.30
FBMEL [9]	63.24	79.76	86.36	-	78.96	91.63	94.36	-	81.15	92.38	93.44	-
GPT-3.5-turbo [11]	-	-	-	-	-	-	-	-	73.80	-	-	-
LLaVA-13B [6]	-	-	-	-	-	-	-	-	76.10	-	-	-
GEMEL [13]†	-	-	-	-	-	-	-	-	82.60	-	-	-
GELR [7]†	-	-	-	-	-	-	-	-	84.80	-	-	-
KGMEL (retrieval)	82.12±0.21	90.28±0.17	92.07±0.05	86.00±0.16	76.40±0.30	85.92±0.28	88.82±0.15	80.94±0.45	87.29±0.08	92.47±0.34	93.94±0.27	89.99±0.25
KGMEL (+re-rank)	**88.23**±0.29	**92.82**±0.06	**93.61**±0.17	**90.84**±0.13	**85.21**±0.24	89.85±0.20	91.32±0.19	88.08±0.21	**90.58**±0.25	95.18±0.29	95.87±0.24	**93.04**±0.26

†: Those that fine-tune LLMs.

[2] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*.

[3] Zhiwei Hu, Víctor Gutiérrez-Basulto, Ru Li, and Jeff Z Pan. 2024. Multi-level Matching Network for Multimodal Entity Linking. In *KDD*.

[4] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*.

[5] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*.

[6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. In *NeurIPS*.

[7] Xinwei Long, Jiali Zeng, Fandong Meng, Jie Zhou, and Bowen Zhou. 2024. Trust in internal or external knowledge? generative multi-modal entity linking with knowledge retriever. In *ACL Findings*.

[8] Pengfei Luo, Tong Xu, Shiwei Wu, Chen Zhu, Linli Xu, and Enhong Chen. 2023. Multi-grained multimodal interaction network for entity linking. In *KDD*.

[9] Hongze Mi, Jinyuan Li, Xuying Zhang, Haoran Cheng, Jiahao Wang, Di Sun, and Gang Pan. 2024. VP-MEL: Visual Prompts Guided Multimodal Entity Linking.

[10] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity recognition for short social media posts. In *ACL*.

[11] OpenAI. 2023. GPT-3.5 Turbo. https://openai.com/index/gpt-3-5-turbo-fine-tuning-and-api-updates/

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

[13] Senbao Shi, Zhenran Xu, Baotian Hu, and Min Zhang. 2024. Generative multimodal entity linking. In *LREC-COLING*.

[14] Xuhui Sui, Ying Zhang, Yu Zhao, Kehui Song, Baohang Zhou, and Xiaojie Yuan. 2024. MELOV: Multimodal entity linking with optimized visual features in latent space. In *Findings of ACL*.

[15] Peng Wang, Jiangheng Wu, and Xiaohang Chen. 2022. Multimodal entity linking with gated hierarchical fusion and contrastive training. In *SIGIR*.

[16] Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. 2022. WikiDiverse: a multimodal entity linking dataset with diversified contextual topics and entity types. In *ACL*.

[17] Ralph Weischedel, Sameer Pradhan, Lance Ramshaw, Martha Palmer, Nianwen Xue, Mitchell Marcus, Ann Taylor, Craig Greenberg, Eduard Hovy, Robert Belvin, et al. 2011. Ontonotes release 4.0. In *LDC2011T03, Philadelphia, Penn.: Linguistic Data Consortium*, Vol. 17.

[18] Zefeng Zhang, Jiawei Sheng, Chuang Zhang, Yunzhi Liang, Wenyuan Zhang, Siqi Wang, and Tingwen Liu. 2024. Optimal Transport Guided Correlation Assignment for Multimodal Entity Linking. In *Findings of ACL*.

[19] Qiushuo Zheng, Hao Wen, Meng Wang, and Guilin Qi. 2022. Visual entity linking via multi-modal learning. *Data Intelligence* 4, 1 (2022), 1–19.