Supplementary Materials

A BASELINES DETAILS

In this section, we introduce the details of baselines mentioned in Section 4.1.

A.1 TEXT-ONLY METHODS

- **ARNN** (text only) [4], devises an Attention-RNN to reduce the noise in local contexts and proposes a novel method for sampling informative negative samples.
- **BERT** (text only) [3], is involved to learn the textual features and then learns the matching score between mentions and entity candidates through contrastive learning.
- **BLINK** (text only) [12], proposes a two-stage zero-shot linking algorithm, which first retrieves the entity candidates through two independent bi-encoder and then re-ranked the candidates through a cross-encoder.
- **GENRE** (text only) [2], is an autoregressive end-to-end framework that retrieves entities by generating their unique names, which helps to capture the fine-grained interactions between contexts and entities and also solves the memory storage problem for large entity sets.
- **GHMFC-onlytext** (text only) [10], utilizes the token-level and phrase-level feature extraction to capture the fine-grained textual representations. It also adopts the contrastive learning to minimize the distance between mentions and positive entities, while maximizing the distance between mentions and negative entities.

A.2 TEXT-VISION METHODS

- **JMEL** (text + vision) [1], employs fully connected layers to embed the visual and textual features into the same space and leverages triplet loss to measure the distance between positive and negative entities.
- **DZMNED-BERT** (text + vision) [7], designs a hybrid module with a modality attention to focus on addressing the entity disambiguation problem. For fairness, we also leverage pre-trained BERT and CLIP model to extract the textual and visual representations.
- **HieCoATT-Alter** (text + vision) [6], proposes an alternating co-attention mechanism to fuse the multimodal information and helps the model to capture the fine-grained joint representations.
- **GHMFC** (text + vision) [10], proposes a hierarchical multimodal co-attention module with a gated fusion to learn the fine-grained inter-modal correlations and a multimodal contrastive loss to reduce the noise.
- **LXMERT** (text + vision) [11], adopts two independent multimodal encoders to learn the representation of contexts and entity candidates. Then leverage the pre-trained LXMERT [9] to fuse the multiple modalities.

B IMPLEMENTATION DETAILS

In this section, we provide the implementation details of our method. Our MMEL framework is implemented with PyTorch on NVIDIA RTX A6000. We leverage the pre-trained base-uncased BERT model [3] as the textual encoder and CLIP model [8] as the visual encoder. We set the dimensions of textual and visual features, d_t and d_v, to 512 and 768. The number of stacked modules q is 2 and the new size of visual features k is 4. The learning rate is selected as 5e-5 and the dropout rate is set 0.2 to avoid overfitting. We leverage the AdamW [5] to optimize the whole parameters with the batch size 32. Following [10], we employ the longest common subsequence algorithm, common prefix and normalized edit distance between contexts and entities to obtain $|E| = 100$ candidate entities for each mention . The Top-k metrics are adopted to measure the performance of models and all the hyper-parameters are manually adjusted based on the top-5 result on the validation set.

REFERENCES

[1] Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. Multimodal entity linking for tweets. In *ECIR*, pages 463–478. Springer, 2020.

[2] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. In *ICLR*. OpenReview.net, 2021.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

[4] Yotam Eshel, Noam Cohen, Kira Radinsky, Shaul Markovitch, Ikuya Yamada, and Omer Levy. Named entity disambiguation for noisy text. In *CoNLL*, pages 58–68. Association for Computational Linguistics, 2017.

[5] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017.

[6] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. *Advances in neural information processing systems*, 29, 2016.

[7] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. Multimodal named entity disambiguation for noisy social media posts. In *ACL*, pages 2000–2008, 2018.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[9] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In *EMNLP/IJCNLP (1)*, pages 5099–5110. Association for Computational Linguistics, 2019.

[10] Peng Wang, Jiangheng Wu, and Xiaohang Chen. Multimodal entity linking with gated hierarchical fusion and contrastive training. In *SIGIR*, pages 938–948. ACM, 2022.

[11] Xuwu Wang, Junfeng Tian, Min Gui, Zhixu Li, Rui Wang, Ming Yan, Lihan Chen, and Yanghua Xiao. Wikidiverse: A multimodal entity linking dataset with diversified contextual topics and entity types. In *ACL*, pages 4785–4797. Association for Computational Linguistics, 2022.

[12] Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. Scalable zero-shot entity linking with dense entity retrieval. In *EMNLP*, pages 6397–6407. Association for Computational Linguistics, 2020.