Attribute-Aware Multimodal Entity Linking

Barry Menglong Yao
Virginia Tech
barryyao@vt.edu

Meghana Holla
Virginia Tech
mmeghana@vt.edu

Abstract

We propose attribute-aware multimodal entity linking, where the input is a mention and a large corpus of entities, including text, images and attribute information, and the goal is to predict the corresponding target entity of the input mention. To support this research, we construct *Anonymized*[1], a large-scale dataset consisting of 35,002 reviews and 37,111 products. To establish baseline performance on *Anonymized*, we experiment with the current multimodal entity linking model and our enhanced attribute-aware model and demonstrate the importance of incorporating the attribute information into the entity linking process. To be best of our knowledge, we are the first to build a benchmark dataset and solutions for the attribute-aware multimodal entity linking task. Datasets and codes will be made publicly available.

1 Introduction

Entity Linking is a critical task in many natural language processing applications such as Question Answering and Recommendation Systems. A host of approaches have been proposed (Onoe and Durrett, 2020; Zhang et al., 2022; Sun et al.; Tang et al., 2021a; Yan; Ganea and Hofmann; Prabhakar et al.; Ayoola et al., a,b) to solve this task.

Recently, researchers have started to work on multimodal entity linking and show promising results after incorporating the visual information into the entity linking tasks (Tang et al., 2021b; Moon et al.; Qin et al., 2021; Li and Wang, 2021; Venkitasubramanian et al., 2017; Zheng et al., 2022; Dost et al., 2020; Wang et al., 2022; Adjali et al.). Most works exploit the relationship between text and image while ignoring the metadata of entities. For example, for a Wikipedia entity, except for the article and the corresponding image(s), there are also abundant structural attributes like the location and date of the mentioned event, e.g. 2022 North Kosovo crisis[2]. On the other hand, in the e-commerce domain, each product has many structural attributes like color, shape, brand, system memory size, and processor model for a laptop product. We believe that these pieces of information can prove to be important sources of cross-modal understanding and can help enhance the entity-linking process. With this motivation, we decide to explore the effect of incorporating attribute information into the multimodal entity linking task.

We start by creating an attribute-aware multimodal entity linking dataset. We noticed that the review-product pairs within the e-commerce websites, like Amazon, and BestBuy, are natural sources for our attribute-aware multimodal entity linking task since we can get the review image(s), product image(s), and product attributes from these websites. The process of finding the target product of one specific review can also be considered a natural entity-linking process. As a result, we form our dataset based on the review-product pairs from the e-commerce website.

We then design our attribute-aware multimodal entity-linking model to take advantage of this information and show its benefits with extended experiments. Overall, the contributions of our work are as follows:

- To the best of our knowledge, this is the first study that integrates attribute information into the multimodal entity linking task.
- We create the first benchmark dataset called *Anonymized* for attribute-aware multimodal entity linking.
- We illustrate the effectiveness of our method, and the benefits of incorporating the attribute information at large, through experiments that compare

[1] Dataset not made public yet due to funding restrictions.

[2] https://en.wikipedia.org/wiki/2022_North_Kosovo_crisis

our attribute-aware model to recent multimodal entity linking models.

2 Related Work

Multimodal entity linking has been explored in various contexts such as social media (Tang et al., 2021b; Moon et al.; Qin et al., 2021; Li and Wang, 2021) and domain-specific videos (Venkitasubramanian et al., 2017) as well approaches covering a diverse set of domains (Wang et al., 2022). Works in the social media context either focus on the reduction of noise in the abundant social media visual inputs (Tang et al., 2021b; Li and Wang, 2021). Specifically, Tang et al. (2021b) proposes utilizing multiple attention mechanisms to overcome the influence of noise through irrelevant images. Li and Wang (2021) utilize topic clustering mechanisms to coalesce similar concepts from texts and images to filter out noise from irrelevant images. On another note, Moon et al. throw light on the limited context that the textual counterparts of the social media posts tend to be owing to very short captions associated with their image posts. Given the diverse amounts of multimodal data that need parsing, recent works focus on proposing zero-shot multimodal entity linking (Zheng et al., 2022; Moon et al.). Venkitasubramanian et al. (2017) propose entity disambiguation in a video-textual context, by employing probabilistic graphical models by visualizing the text-video pair as a bipartite graph. With the emergence of the need for entity linking in a multimodal context, a lot of works propose datasets to remedy a host of issues in this space. VTKEL (Dost et al., 2020) is proposed to effectively incorporate the alignment of background knowledge with visual and textual information. Adjali et al. craft a dataset with the entities grounded in the Twitter Knowledge base. Additionally, they propose a method to jointly model image and text to model representation for entities and mentions. (Qin et al., 2021) propose Weibo-MEL, Wikidata-MEL, and Richpedia-MEL that encompass a diverse range of sources. In a similar vein, Zheng et al. (2022) propose ZEMELD that focuses on zero-shot entity linking capacities. Gan et al., M3EL along with a bipartite graph matching multimodal entity linking benchmark. Finally, Wang et al. (2022), in an attempt to counter limited coverage and simplified mention ambiguity, introduce WikiDiverse, which is grounded in Wikipedia.

Our approach, in contrast to prior works in Multimodal entity linking and disambiguation, takes into consideration unique attribute information along with the visual and textual inputs.

3 Dataset Construction

Anonymized dataset details redacted due to funding restrictions.

4 Approach

Our method for entity linking comprises two steps - Entity Text/Image Retrieval and Entity Disambiguation, inspired by our baseline work (Wang et al., 2022). The following sections cover both these steps in detail:

4.1 Entity Text/Image Retrieval

Prior to the disambiguation step, we conduct a text-image retrieval mechanism for obtaining the entity candidate set to condense our search space to the most relevant entities for consideration in the entity disambiguation step. Specifically, we consider the mention and the global entity texts for retrieving entities that have similar properties. We encode the text using BLINK (Wu et al., 2019) embedding. A top-J ($J = 1000$) retrieval using cosine similarity ranking is employed. Then, we employ a neural similarity ranking using BLINK to retrieve top-K ($K = 10$) candidates. To obtain visually similar entities, we perform a similar encode-and-retrieve mechanism on entity images by employing CLIP (Radford et al., 2021) for obtaining the image representations followed by a top-K retrieval step using cosine similarity as previously mentioned. The final multimodal entity candidates set is generated by coalescing the two candidate sets retrieved. Figure 2 illustrates the entity retrieval process employed in this work.

4.2 Entity Disambiguation

The entity disambiguation step is where our major contribution lies. As previously mentioned, prior work only employs descriptive text and image representations for disambiguation. In contrast, our approach considers the structured entity attribute text for our mention/entity representation in addition to the image and description text representations. In other words, we have a third stream of data encoding the attribute information. Section 4.3 goes through our proposed approach in detail. Figure 1 illustrates our approach.

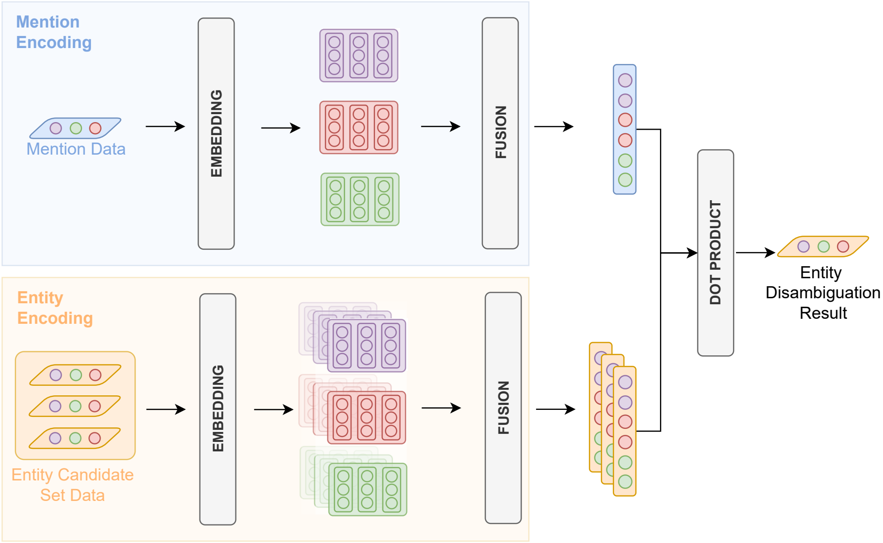

Figure 1: Proposed approach overview: Entities and mentions have image, description, and attribute information. The encoding of entities and mentions happens separately, but with identical steps of embedding followed by fusion and dot product operation for entity disambiguation.

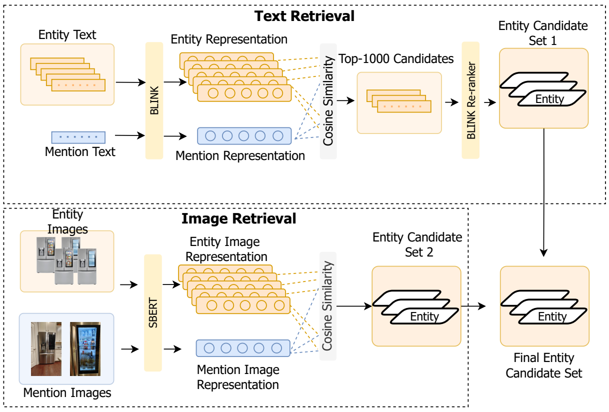

Figure 2: Our Entity candidate retrieval pipeline. We retrieve the most relevant candidate using cosine similarity with regard to both the textual entity information relevance and image contextual relevance and fuse the retrieved candidate to obtain a final candidate set.

4.3 Model Overview

Baseline: We design our baseline to follow Wang et al. (2022)'s architecture. Specifically, our baseline considers entity and mention representation by using the image and the description text information. A simple fusion mechanism that concatenates the two representations follows this step. A dot product operation is employed to find the entity representation most similar to the mention.

Attribute Integration: The attribute information is encoded by a simple concatenation of the key-value pairs. A typical attribute set can be represented as key-value pairs. e.g., $\{name : ABC, brand : XYZ\}$. We perform a simple concatenation of the keys and values, each pair separated by a ".", to form the single sentence representation. In the above example, our attribute sentence representation would look like "$name\ ABC.\ brand\ XYZ$". This textual representation is passed through the same encoding and tokenization pipeline as with the description text.

4.3.1 Model Setup

Our approach is composed of two identical branches - one each for the mention and entities. For a given mention/entity, we encode the three data streams image (I), textual description (T), and attributes (A). Specifically, we obtain the representations for description and attribute using BERT embeddings, i.e., $\{t_1, t_2, \ldots, t_n\} = Bert(T)$ and attribute text $\{a_1, a_2, \ldots, a_n\} = Bert(A)$. As for the image representation, we employ CLIP to obtain the representation: $\{i_1, i_2, \ldots, i_m\} = Clip(I)$. n is the padded length of the tokens and m is the number of visual features extracted. Following this, we employ a fusion operation between the three streams of data to obtain the joint representation of the entity, as explained below:

Fusion by Concatenation This configuration fuses the encoding of image I, description text D, and attribute text A using a simple concatenation operation to get the fused representation $X_{fused} = [I, D, A]$. This representation is further encoded using a transformer encoder to obtain the final mention/entity representation(X).

$$X = Transformer(X_{fused})$$

As mentioned earlier, there are separate branches of the above pipeline for encoding mention and entities. We receive the fused mention representation X_m and entity representations for d candidate entities $\{X_e^1, X_e^2, \ldots, X_e^d\}$. Finally, a scoring mechanism is employed using a dot product to obtain the disambiguated result X_e^*.

5 Experiments and Analysis

5.1 Entity Disambiguation

Given one review and its product candidates, our model predicts the target product based on the coreference between their text and images. We evaluate the entity disambiguation performance based on F1-score. We compare our methods with three baseline. (1)*Wikidiverse* (Wang et al., 2022), which is a multimodal entity linking model without incorporating the attribute information. (2) *Random Label*, which chooses the target product randomly. We also set up a human performance by sampling 50 reviews and asking one annotator to choose the target product from 21 product candidates of each review. Table 1 displays the results of our approach compared to the above-mentioned settings. We see that our approach surpasses the baseline *Wikidiverse*. the performance of our model also significantly beats *Random Label* configuration. On the other hand, the *Human* results are still considerably higher. . An increase in the F1-score from out baseline (*Wikidiverse*) indicates a positive influence of adding attribute information. We believe this to be the case because the attribute information potentially contains traits of the entities that effectively bridges the gap between the textual (i.e., description) and visual inputs (i.e., product image), since the attribute contains visual attributes as well as descriptive attribute information. However, there still is a considerable gap between our model and human annotations. Perhaps, exploring fusion methods of higher sophistication could help push the performance further. At the same time, we want to clarify that the human performance is based on the top-21 setting while our current model is evaluated on the top-6 setting, which means there is still a huge room to increase the performance.

Setting	# Candidate	F1-score (%)
Human	21	56
Ours	6	45.71
Wikidiverse	6	43.47
Random	6	16.67

Table 1: Performance of entity disambiguation. Candidate Number specifies the number of candidates presented (to the human/model) for each review

5.2 Sensitivity Analysis: Top-K for Query Product Candidates

In our dataset, we query the top-K product candidates whose text/images are most similar to the text/images of the target product. In this section, we explore the sensitivity of variable K. In detail, we sample 50 reviews and then sample the top-20 product candidates, top-10 product candidates, respectively, and then ask an annotator to do the entity-linking task. We can find that the performance of these two settings is close to each other, which means most of the confusing product candidates are in the top-10 so the task does not become harder after increasing candidates from top-10 to top-20.

Top-K	Modality	Data Size	Accuracy
20	Image, Text	50	62%
10	Image, Text	50	56%

Table 2: Human Performance. Giving the data samples, we use the top-K similar products based on text/image to form the candidate set and ask the annotator to choose the target product from this set. The low accuracy shows that the dataset can be further improved.

6 Conclusion

We propose an attribute-enhanced multimodal entity linking approach, which utilizes three streams of data - image, descriptive text, and structured attribute information - for multimodal entity linking/disambiguation. Our experimental analysis shows that adding the attribute information indeed enhances model performance by a considerable amount. Additionally, our curated dataset *Anonymized* encapsulates image and text annotations with attribute annotations, which can be utilized to explore further in this line of research. Future work could explore more complex modes of fusion between the three streams of data, thereby enforcing improved cross-modal representation and understanding.

7 On-going Experimentation

We are also simultaneously exploring more high orders of fusion between the three data streams. Given below is one such variant we are currently experimenting on:

Two-Step Fusion This configuration fuses the encoding of image I, description text D, and attribute text A in two steps. Firstly, the textual representation D and T are fused using a cross-attention mechanism.

$$T_{fused} = CA(D, A)$$

Following this, the fused text representation and the Image representation I are fused using a vision-language module. Specifically, we employ LXMERT(Tan and Bansal, 2019) to obtain a joint representation.

$$X = LXMERT(I, T_{fused})$$

Finally, we take the hidden state representations corresponding to the first timestep for the final fused representation.

References

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. Multimodal entity linking for tweets. Multiple features, then cancat.

Tom Ayoola, Joseph Fisher, and Andrea Pierleoni. a. Improving entity disambiguation by reasoning over a knowledge base.

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. b. Refined: An efficient zero-shot-capable approach to end-to-end entity linking.

Shahi Dost, Luciano Serafini, Marco Rospocher, Lamberto Ballan, and Alessandro Sperduti. 2020. Vtkel: A resource for visual-textual-knowledge entity linking. *Proceedings of the ACM Symposium on Applied Computing*, pages 2021–2028.

Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, Qingming Huang, and Qing-Ming Huang. Multimodal entity linking: A new dataset and a baseline dataset; multimodal alignment; multimodal entity linking; optimal transportation acm reference format.

Octavian-Eugen Ganea and Thomas Hofmann. Deep joint entity disambiguation with local neural attention.

Pengyuan Li and Yongli Wang. 2021. A multimodal entity linking approach incorporating topic concepts. *Proceedings - 2021 International Conference on Computer Information Science and Artificial Intelligence, CISAI 2021*, pages 491–494.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. Zeroshot multimodal named entity disambiguation for noisy social media posts. pages 2000–2008. Purely lexical embedding.

Yasumasa Onoe and Greg Durrett. 2020. Fine-grained entity typing for domain independent entity linking. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, pages 8576–8583.

Manoj Prabhakar, Kannan Ravi, Kuldeep Singh, Isaiah Onando Mulang', Saeedeh Shekarpour, Johannes Hoffart, and Jens Lehmann. Cholan: A modular approach for neural entity linking on wikipedia and wikidata.

Bing Qin, Zhi Jin, Haofen Wang, Jeff Pan, and Yongbin Liu. 2021. *Weibo-MEL, Wikidata-MEL and Richpedia-MEL: Multimodal Entity Linking Benchmark Datasets*.

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. 2021. Learning transferable visual models from natural language supervision.

Yaming Sun, Lei Lin, Duyu Tang, Nan Yang, Zhenzhou Ji, and Xiaolong Wang. Modeling mention, context and entity with neural networks for entity disambiguation.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Hongyin Tang, Xingwu Sun, Beihong Jin, and Fuzheng Zhang. 2021a. A bidirectional multi-paragraph reading model for zero-shot entity linking. *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, 15:13889–13897.

Xiu Tang, Sai Wu, Gang Chen, Ke Chen, and Lidan Shou. 2021b. *Attention-Based Multimodal Entity Linking with High-Quality Images*, volume 12682 LNCS.

Aparna Nurani Venkitasubramanian, Tinne Tuytelaars, and Marie Francine Moens. 2017. Entity linking across vision and language. *Multimedia Tools and Applications*, 76:22599–22622.

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 *Proceedings of the 60th Annual Meeting of* *the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4785–4797, Dublin, Ireland. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2019. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.

Wenzheng Zhang, Wenyue Hua, and Karl Stratos. 2022. Entqa: Entity linking as question answering.

Qiushuo Zheng, Hao Wen, Meng Wang, Guilin Qi, and Chaoyu Bai. 2022. Faster zero-shot multi-modal entity linking via visual-linguistic representation under a creative commons attribution 4.0 international (cc by 4.0) license.