M^3EL: A Multi-task Multi-topic Dataset for Multi-modal Entity Linking

Fang Wang[1], Shenglin Yin[1], Xiaoying Bai[2*], Minghao Hu[2], Tianwei Yan[3], Yi Liang[4]

[1]Peking University
[2]Advanced Institute of Big Data,Beijing
[3]National University of Defense Technology
[4]Xinjiang University
{fangwang, yinsl}@stu.pku.edu.cn, baixy@aibd.ac.cn, huminghao16@gmail.com, augusyan@hotmail.com, 107556522207@stu.xju.edu.cn

Abstract

Multi-modal Entity Linking (MEL) is a fundamental component for various downstream tasks. However, existing MEL datasets suffer from small scale, scarcity of topic types and limited coverage of tasks, making them incapable of effectively enhancing the entity linking capabilities of multi-modal models. To address these obstacles, we propose a dataset construction pipeline and publish M^3EL, a large-scale dataset for MEL. M^3EL includes 79,625 instances, covering 9 diverse multi-modal tasks, and 5 different topics. In addition, to further improve the model's adaptability to multi-modal tasks, We propose a modality-augmented training strategy. Utilizing M^3EL as a corpus, train the *CLIP$_{ND}$* model based on *CLIP* (*ViT-B-32*), and conduct a comparative analysis with an existing multi-modal baselines. Experimental results show that the existing models perform far below expectations (ACC of 49.4%-75.8%), After analysis, it was obtained that small dataset sizes, insufficient modality task coverage, and limited topic diversity resulted in poor generalization of multi-modal models. Our dataset effectively addresses these issues, and the *CLIP$_{ND}$* model fine-tuned with M^3EL shows a significant improvement in accuracy, with an average improvement of 9.3% to 25% across various tasks. Our dataset publicly available to facilitate future research.

Introduction

Multi-modal Entity Linking is a crucial research direction in Natural Language Processing (NLP) and Computer Vision (CV) (Sevgili et al. 2022). It aims to achieve cross-modal information integration and understanding by associating entities in different modalities (e.g., text, images, etc.), and to improve the accuracy of linking to entities in the Knowledge Graph (KG). It is an important foundation for NLP downstream tasks, e.g., Information Retrieval (Chang et al. 2006; Martinez-Rodriguez, Hogan, and Lopez-Arevalo 2020) and Question-Answer systems (Allam and Haggag 2012; Mollá, Van Zaanen, and Smith 2006). Despite recent research progress, existing MEL methods face the following challenges existing in training data, thus far from meeting the requirements of real-world applications.

There are many limitations of current MEL datasets, which are summarised in Table 1. Firstly, existing datasets are generally small in scale; for instance, WIKIPerson (Sun et al. 2022) involves 13K entities and WIKIDiverse (Wang et al. 2022a) covers 40K entities, which are insufficient for complex tasks. Secondly, the scope of modal tasks is restricted (Hoffart et al. 2011; Moon, Neves, and Carvalho 2018; Gan et al. 2021), primarily to *Text-Text*, *Image-Text*, or *Image-*(*Image+Text*), while neglecting other potentially valuable tasks such as *Text-Image* and *Text-*(*Image+Text*). Additionally, most datasets are confined to single topic (Guo and Barbosa 2018; Zhang, Li, and Yang 2021), such as person or news, and lack the ability to generalise across topics. These constraints significantly impede the development of MEL tasks.

In this paper, we address these challenges by releasing the large-scale multi-task, multi-topic, multi-modal entity linking dataset, termed M^3EL, which has three unique key properties. First, the dataset contains 79K instances and 318.5K images, which is nearly 10 times the instances in WIKIPerson (Sun et al. 2022) and 6.37 times the images in WIKIDiverse (Wang et al. 2022a). Second, M^3EL covers the widest range of multi-modal tasks, including: *Text-Text*, *Text-Image*, *Text-*(*Image+Text*), *Image-Text*, *Image-Image*, *Image-*(*Image+Text*), (*Image+Text*)-*Text*, (*Image+Text*)-*Image*, (*Image+Text*)-(*Image+Text*). Finally, the M^3EL dataset provides more extensive topics (movies, common, person, books, sports), while others provide only a single topic. This implies that our dataset is a more challenging setting.

Furthermore, we benchmark many existing models in our dataset, the experimental results reveal that existing models exhibit sub-optimal performance, with accuracy (ACC) ranging from 49.4% to 75.8%. To improve the adaptability of MEL methods, we introduce a modality-augmented training strategy to fine-tune the *CLIP* (*ViT-B-32*) (Radford et al. 2021) model. This strategy, by expanding the expressiveness of modal features, augments the training process, resulting in the model *CLIP$_{ND}$*. Following experimentation and analysis, the *CLIP$_{ND}$* model obtained after training demonstrates significant performance improvements in ACC, with an average increase of 9.3% to 25%. This illustrates the considerable potential of the modal enhancement strategy in handling multi-modal data and emphasizes the research value of the M^3EL dataset.

*Correspond author.

Task	Dataset	Source	Topic	Size	Modality
EL	AIDA (Hoffart et al. 2011)	Wikipedia	Sports	1K docs	$T_m \rightarrow T_e$
	MSNBC(Cucerzan 2007)	Wikipedia	News	20 docs	$T_m \rightarrow T_e$
	AQUA(Milne and Witten 2008)	Wikipedia	News	50 docs	$T_m \rightarrow T_e$
	ACE2004(Ratinov et al. 2011)	Wikipedia	News	57 docs	$T_m \rightarrow T_e$
	CWEB(Guo and Barbosa 2018)	Wikipedia	Web	320 docs	$T_m \rightarrow T_e$
	WIKI(Guo and Barbosa 2018)	Wikipedia	Common	320 docs	$T_m \rightarrow T_e$
	Zeshel(Logeswaran et al. 2019)	Wikia	Common	-	$T_m \rightarrow T_e$
MEL	Snap(Moon, Neves, and Carvalho 2018)	Freebase	Social Media	12K captions	$T_m, I_m \rightarrow T_e$
	Twitter(Adjali et al. 2020a)	Twitter users	Social Media	4M tweets	$T_m, I_m \rightarrow T_e, I_e$
	Movie(Gan et al. 2021)	Wikipedia	Movie Reviews	1K reviews	$T_m, I_m \rightarrow T_e, I_e$
	Weibo(Zhang, Li, and Yang 2021)	Baidu Baike	Social Media	25K posts	$T_m, I_m \rightarrow T_e, I_e$
	WIKIDiverse(Wang et al. 2022a)	Wikipedia	News	8K captions	$T_m, I_m \rightarrow T_e, I_e$
	WIKIPerson(Sun et al. 2022)	Wikipedia	Person	50K images	$T_m, I_m \rightarrow T_e, I_e$
	M^3EL	Wikipedia, Wikidata, Dbpedia, Goodreads	Sports, Movies, Books, Person, Common	79K instances 318.5K images	$T_m, I_m, (I+T)_m \rightarrow T_e, I_e, (I+T)_e$

Table 1: Overview of EL and MEL datasets. T_m (T_e) and I_m (I_e) are abbreviations for $Text_m$ ($Text_e$) and $Image_m$ ($Image_e$), which represent the textual and visual contexts of mentions m (or entities e) respectively.

To summarize, the main contributions of this work are as follows:

1. We present a large-scale manually labelled high-quality dataset, M^3EL, which covers a diverse range of topics and supports various multi-modal tasks.
2. We propose a modality-augmented training strategy. Based on this strategy, we use the M^3EL dataset as a corpus to train a *CLIP* (*ViT-B-32*) model, resulting in *CLIP*$_{ND}$, which further improves the model's performance in multi-modal entity linking.
3. Experimental results indicate that our proposed large-scale, multi-task, multi-topic, multi-modal dataset serves as a high-quality pre-training corpus. While effectively enhancing the model's generalization performance, our dataset still presents a challenging benchmark.

Related Work

Textual Entity Linking Entity Linking (EL) is a fundamental and classical task in Natural Language Processing (NLP) that has been extensively studied. Early research introduced a variety of datasets to evaluate the performance of EL models, including high-quality manually annotated datasets such as AIDA (Hoffart et al. 2011), and large-scale automatically annotated datasets such as CWEB (Guo and Barbosa 2018). In addition, zero-shot entity linking datasets such as Zeshel (Logeswaran et al. 2019) have also been used for relevant studies. However, the performance of many EL methods on traditional datasets such as AIDA-test, MSNBC (Cucerzan 2007), and AQUAINT (Milne and Witten 2008) has stabilised in recent years, approaching the upper limit of performance for the task, which may indicate that performance on these datasets is nearing saturation. With the development of large-scale pre-trained language models, the latest deep learning methods have achieved over 90% accuracy on the AIDA dataset, which also implies that current methods may be approaching a performance bottleneck (De Cao et al. 2020). Consequently, to advance the field, researchers have designed more challenging tasks such as zero-shot entity linking (Logeswaran et al. 2019; Wu et al. 2019), global coherence exploitation (Chen et al. 2020), NIL prediction (Rao, McNamee, and Dredze 2013), and end-to-end solutions for emerging entities (Broscheit 2020). These tasks are intended to drive further development and breakthroughs in the field of EL.

Multi-modal Entity Linking In recent years, the task of MEL has gained significant attention in the field of NLP. This task aims to utilize both textual and visual information to map ambiguous mentions to entities in a Knowledge Graph (also known as Knowledge Base). Moon et al. first demonstrated the effectiveness of image information in dealing with ambiguous and short textual mentions in social media posts, proposing a zero-shot framework to integrate textual, visual and lexical information for entity linking (Moon, Neves, and Carvalho 2018). However, their proposed dataset is unavailable due to GDPR regulations. Adjali et al. developed a framework for automatically constructing an MEL dataset from Twitter, with small data scale, scarcity of entity types, and ambiguous mentions restricting the usefulness of the dataset (Adjali et al. 2020b). Zhang et al. construct a Chinese multi-source social media multi-modal dataset based on the social media platform Weibo, focusing solely on person entities (Zhang, Li, and Yang 2021). Gan et al. published a MEL dataset focusing on roles and characters in the movie domain based on movie reviews (Gan et al. 2021). Both of these two datasets have a single entity type and the generalisation of the training model is weak. Wang et al. constructed a multi-modal entity linking dataset with diverse contextual themes and entity types, including multiple modal tasks (Wang et al. 2022a). However, all those works still face issues such as small data size, limited entity types, and incomplete multi-modal tasks, therefore, the diversity of entity types and modal tasks in datasets needs further exploration and optimization.

Entity Linking Dataset Our M^3EL dataset also encompasses multi-modal image-text datasets. While existing re-

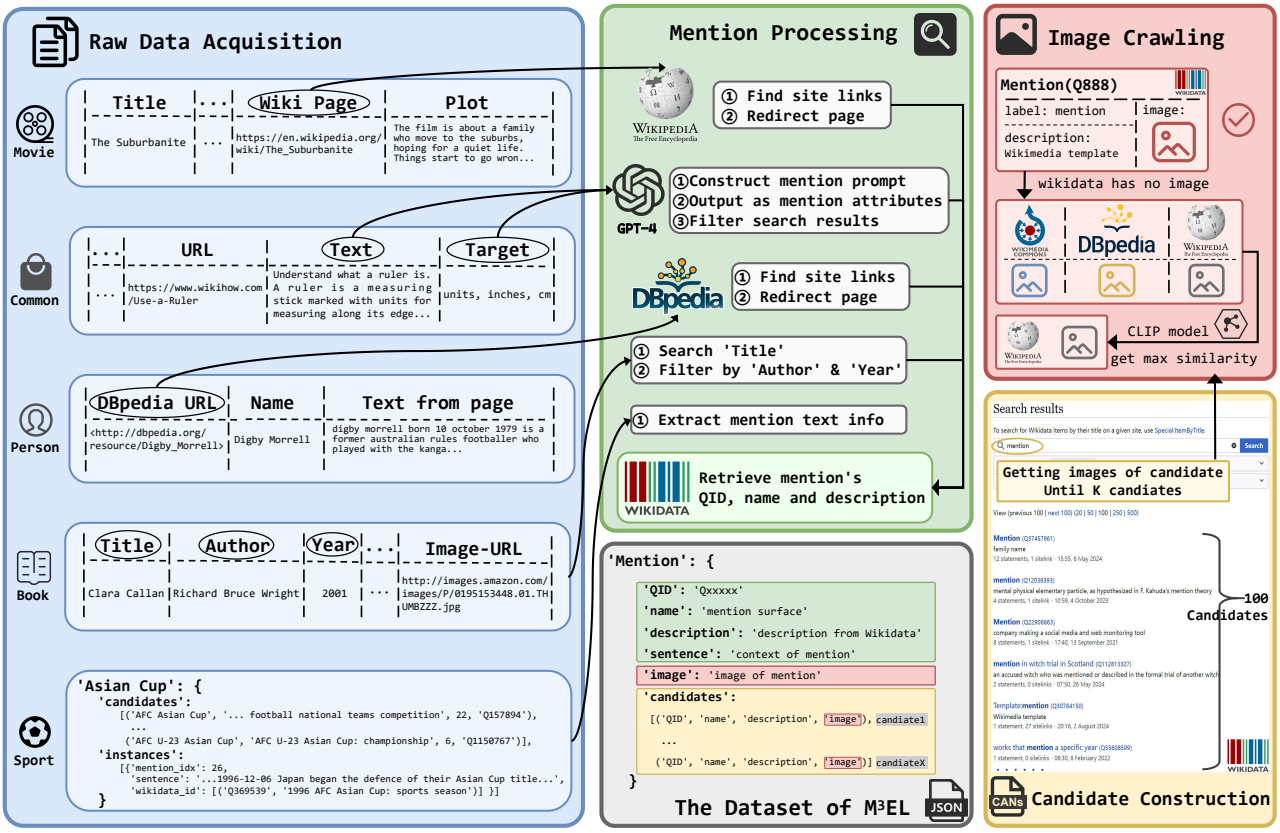


Figure 1: Overview of the M^3EL data construction pipeline.

lated studies ((Biten et al. 2019; Tran, Mathews, and Xie 2020; Liu et al. 2020; Zhang et al. 2018; Lu et al. 2018)) have proposed larger-scale image-text datasets, the entities in these datasets lack detailed explanations and are not linked to a unified knowledge graph. The Flickr 30k (Young et al. 2014) and MSCOCO (Chen et al. 2015) datasets made strides in size and annotate entities with descriptive sentences. However, the mention information in these sentences tends to be more ambiguous. Although the WIKIPerson (Sun et al. 2022) and WIKIDiverse (Wang et al. 2022a) datasets address some of these issues, they contain only person entities and focus mainly on specific multi-modal tasks, such as *Text-Text* and *Image-Image*. The comprehensiveness of the task types still needs improvement.

Problem Formulation

Multi-modal entity linking maps mentions in multi-modal contexts to corresponding entities in a target Knowledge Graph (*KG*). Since the large number of entities in *KG*, traversing all entities for linking significantly increases time costs, thus candidate sets of entities are usually predefined. On this basis, we construct a candidate set for each mention, which may or may not contain the correct corresponding entity, accurately reflecting the challenges inherent in multi-modal entity linking tasks.

Formally, let E represent the set of entities in the *KG*, typically comprising millions of entities. Let C denote the candidate set in the *KG* for each mention m, where $C \subseteq E$. Each mention m or entity $e \in C$ is associated with its corresponding visual context V_m, V_e, and textual context T_m, T_e. Here, T_m and T_e denote the textual context surrounding m and e, respectively. V_m refers to the images related to m, while V_e refers to the images associated with e in the *KG*. Therefore, the multi-modal task of identifying the corresponding entity for mention m can be defined as follows:

$$Sim(M_m, M_e) = \arg\max_{e_i \in C} \Psi(En(M_m), En(M_e)), \quad (1)$$

where M_m and M_e represents the modal information of the target mention m and the candidate entity e, respectively, as detailed in Table 2. En denotes *Encoder*. The function Ψ denotes the similarity score between the m and the e.

M_m	M_e	Multi-modal Tasks
T_m	T_e	$T_m - T_e \mid T_m - I_e \mid T_m - (I+T)_e$
I_m	I_e	$I_m - T_e \mid I_m - I_e \mid I_m - (I+T)_e$
$(I+T)_m$	$(I+T)_e$	$(I+T)_m - T_e \mid (I+T)_m - I_e \mid (I+T)_m - (I+T)_e$

Table 2: Introduction to different modal information and modal tasks. T_m (T_e) and I_m (I_e) refer to the textual and visual information of the mention (entity), respectively.

Data Setups for M^3EL

The framework of the data construction pipeline of M^3EL is illustrated in Figure 1. In the following, we detail the construction of M^3EL[1].

Data Collection

Raw data Acquisition. To construct the foundational dataset for M^3EL, we collected raw data from Kaggle[2] that includes URL information related to DBpedia[3], Wikipedia[4], or Wikidata[5]. These data files are diverse in format, including JSON, CSV, Pickle, and more. Since DBpedia, Wikipedia and Wikidata pages can interlink, where Wikidata's QID attributes can be used as entity labels, these datasets are well-suited as raw data sources for research in multi-modal entity linking. In addition, these Kaggle datasets are derived from real-world data science scenarios and cover a wide range of topics, reflecting the complexity and diversity of data in the real world. Specifically, we utilized *the Wikipedia Movie Plots* [6] dataset as the raw data for the movies topic, the *Wiki-Wiki* [7] dataset for common knowledge topic, the *People Wikipedia Data*[8] for person topic, the *Books Datasets* [9] for book topic, and the *AIDA-CoNLL-Test*[10] for the sports topic.

Data Filter and clean. For the raw dataset, we performed the following data cleaning steps: 1) Removed non-English expressions and single-character mentions. 2) Deleted mentions without associated images or whose images could not be downloaded. Following these pre-processing steps, we finally obtained a total of 82K mentions, each comprising a text-image pair, which serves as the base dataset for M^3EL.

Data Curation

Mention Processing. Due to the varying formats of raw data across different topics, we describe in detail the process of obtaining relevant mentions for each topic separately.

For the raw data on the movie topic, we utilized the "Title" attribute from the *wiki_movie_plots_deduped.csv* file to extract movie titles as mentions. The "Plot" attribute was employed to provide contextual information related to these mentions. Additionally, we accessed the mention-related Wikipedia pages through the URLs provided in the "Wiki Page" attribute. From these Wikipedia pages, we further navigated to the corresponding Wikidata pages to obtain fundamental information, including the QID and descriptions.

For the raw data on commons topic, we utilized the instances provided by the "Target" attribute in the *wikihow.csv* file as mentions, and paragraphs provided by the "Text" attribute as contextual information for mentions. By constructing a mention semantic understanding prompt, we input it into the large language model GPT-4 to obtain semantic information related to the mentions. Subsequently, each mention was searched in Wikidata, and the Wikidata entry that best matches the mention is filtered based on the semantic information obtained from GPT-4, and relevant information such as QID and description is obtained.

For the raw data on person topic, we utilized the *people_wiki.csv* file, specifically the "name" attribute, as the reference for mentions, and the "text" attribute, which provides the context for these mentions. The associated DBpedia pages were accessed via the "URL" attribute. Subsequently, we navigated to the corresponding Wikidata pages to obtain foundational information, including the QID and descriptions.

For the raw data on book topics, we utilized the "Book-Title" attribute of the *books.csv* file as a mention for searching the corresponding Wikidata pages to retrieve basic information, including the QID and description. To ensure the accuracy of the search results, we cross-validated them using the Book-Author, Year of Publication, and Publisher information provided in the file. Mentions that could not be found on the Wikidata page were subsequently removed.

For the raw data on sport topic, the *AIDA_B_dict_raw.pickle* file provides mentions, contextual information, and associated QID information in Wikidata in the form of a dictionary. We directly utilized the textual information from the raw data without any processing.

Candidate Construction. In most entity linking tasks, Wikidata is used as the target Knowledge Graph, linking mentions to their corresponding entities. However, due to the large scale of Wikidata, which contains approximately 95 million entries, matching every mention by traversing all entities would consume substantial computational resources and time. Therefore, to enhance the quality and usability of the dataset, we employ an optimization strategy: for each mention, a refined subset is filtered from Wikidata to serve as the candidate set. This set contains both possible and incorrect corresponding candidates, which truly reflects the challenges of the multi-modal entity linking task.

The specific steps are as follows: 1) We search for the mention surface on the Wikidata page and retrieve the top 100 relevant search results. For each search result, we access its Wikidata detail page and systematically extract the key information, including the entity's QID, entity name, and description, among other textual data. 2) We enter the stage of obtaining candidate entity images, and when the number of candidates satisfying complete textual and visual information reaches a predetermined threshold of k, the candidate set construction is completed. This process ensures the completeness of the information and the high quality of the dataset, which provides a solid foundation for the subsequent data analysis and model training.

[1] https://github.com/ww-ffff/M3EL

[2] https://www.kaggle.com

[3] https://www.dbpedia.org

[4] https://www.wikipedia.org

[5] https://www.wikidata.org/wiki/Wikidata:Main_Page

[6] https://www.kaggle.com/datasets/jrobischon/wikipedia-movie-plots?select=wiki_movie_plots_deduped.csv

[7] https://drive.google.com/file/d/1Oebe1sbbixX7FWHX813diCdqReh1IyWH/view

[8] https://www.kaggle.com/datasets/sameersmahajan/people-wikipedia-data

[9] https://www.kaggle.com/datasets/saurabhbagchi/books-dataset

[10] https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/ambiverse-nlu/aida/downloads

Image Crawling. After obtaining the mention-related textual information of different topics, in order to obtain the image information corresponding to the mentions, we developed a comprehensive image crawling frameworks based on Wikidata, Wikipedia and Wikimedia[11] pages. Furthermore, we utilized a pre-trained *CLIP* (*ViT-B-32*) model to effectively filter the most relevant images by leveraging the joint representation of visual and textual information. Especially for the book topic, we directly use the URLs of book covers provided in the raw data to obtain relevant images. For invalid URLs, we implemented a supplementary crawler that automatically searches for and retrieves relevant images using book titles as keywords on Goodreads[12] and Amazon[13]. These steps ensure that we are able to obtain the appropriate image information for each mention, enriching the multi-modal characteristics of the dataset.

Data Analysis

Size and Difficulty Measure. The statistics of M^3EL dataset in detail are shown in Table 3. This dataset was constructed from 66,342 articles and contains a total of 79K entities, covering 318.5K images. Each entity is mapped to a specific entity in Wikidata. Notably, many entities appear multiple times in different sentences of the articles, ensuring that the entities can be fully learned. In addition, Figure 2(a) reports the distribution of entity topics. Unlike existing MEL datasets, which are typically constructed for specific scenarios or tasks, M^3EL covers 5 distinct topics and involves 9 types of multi-modal linking tasks.

Firstly, we compare the surface form similarity between mentions and ground-truth entities. It is observed that 41.3% of the mentions differ from the surface forms of the ground-truth entities. This significant discrepancy in surface forms presents a challenge for multi-modal entity linking tasks. Secondly, we report the candidates for each mention in Figure 2(b). We observe the following: 1) 99.8% of mentions have three candidates, with 55.8% of these mentions having the correct entity ranked first in the candidate set, and 29.4% having the correct entity ranked elsewhere within the candidate set. 2) 14.8% of mentions have candidate sets that do not include the correct entity. This indicates that the candidate set reflects the real-world scenarios of entity linking to a certain extent.

Diversity of multi-modal tasks. Compared to existing datasets, M^3EL offers a more comprehensive range of multi-modal tasks. Traditional multi-modal entity linking datasets, such as Snap (Moon, Neves, and Carvalho 2018), Twitter (Adjali et al. 2020a), and Weibo (Zhang, Li, and Yang 2021), are designed for specific tasks. Although datasets like WIKIPerson (Sun et al. 2022) and WIKIDiverse (Wang et al. 2022a) introduce various forms of multi-modal entity linking tasks, including I_m-T_e, I_m-I_e, and $(I{+}T)_m$-I_e, they ignored other potential linking tasks. In contrast, M^3EL significantly expands the scope of tasks, covering 9 distinct types, specifically, T_m-T_e, T_m-I_e, T_m-$(I{+}T)_e$, I_m-T_e, I_m-I_e, I_m-$(I{+}T)_e$, $(I{+}T)_m$-T_e, $(I{+}T)_m$-I_e, $(I{+}T)_m$-$(I{+}T)_e$. This task diversity greatly enhances MEL's ability to generalise across a wide range of multi-modal related tasks, enabling it to better accommodate the diverse needs of users in the real world.

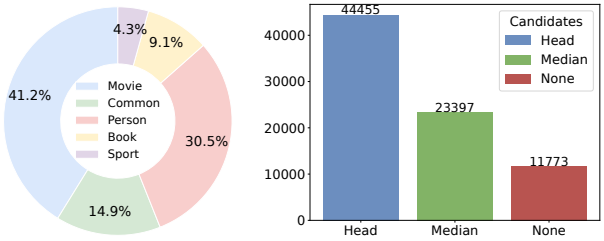
[11]https://commons.wikimedia.org/wiki/Main_Page

[12]https://www.goodreads.com

[13]https://www.amazon.com

Topic	Documents	Mention	Candidate	Image	Tasks
Movie	32131	32800	98400	131200	
Common	1699	11826	35478	47304	
Person	26873	24278	72834	97112	9
Book	5408	7276	21828	29104	
Sport	231	3445	10335	13780	
ALL	66342	79625	238875	318500	9

Table 3: Statistics of M^3EL. Documents represent the raw textual contexts of mentions, Tasks refers to 9 different



(a) Entity topic distribution. (b) Different type candiates distribution.

Figure 2: More statistics of M^3EL. (a) Entity topic type distribution. (b) Distribution of correct entity positions within the candidates.

Coverage of topic types. As shown in Table 1, most of the existing training datasets focus on a limited range of entity topic types, such as person and news. WIKIDiverse (Wang et al. 2022a) extends the types of topic to multiple categories by developing crawling code to extract multi-modal information related to text and images from Wikinews, making it state-of-the-art, but all categories belong to news topic. However, our constructed M^3EL includes 5 types of topics, specifically, movies, common knowledge, person, books, and sports. Moreover, Figure 2(a) illustrates the distribution of M^3EL across different topics.

Model with Augmented Strategy

***CLIP* Definition** The Contrastive Language–Image Pre-training (*CLIP*) method has demonstrated significant effectiveness in training visual models using language supervision. In each training step, a large batch of N pairs of images and texts $\{x_I, x_T\}$ is randomly sampled from the training dataset. Subsequently, data augmentation is applied to the images to simulate various real-world visual variations. The augmented images and their corresponding texts are then processed by dedicated encoders and normalization functions f_I and f_T to extract image and text features. These features are utilized to compute the InfoNCE loss, a type of contrastive loss that distinguishes between matched (posi-

tive) and mismatched (negative) image-text pairs. The training loss is represented by the following equation:

$$L = \sum_{i=1}^{N} \log \frac{exp(sim(f_I(aug_I(x_I^i)), f_T(x_T^i))/\tau)}{\sum_{j=1}^{N} exp(sim(f_I(aug_I(x_I^j)), f_T(x_T^j))/\tau)}, \tag{2}$$

where (x_I^i, x_T^i) represents the i-th image-text pair, and $aug_I()$ denotes the image augmentation function. The $sim(-,-)$ function measures similarity using the dot product, while the temperature parameter τ is a learnable variable used to scale the logits.

Augmented Strategy In Equation 2, while the standard *CLIP* loss applies augmentation to images to enhance the model's adaptability and robustness to visual variations, the textual input remains constant throughout the training process. This approach ignores the potential diversity and ambiguity inherent in textual data. To address this issue, we propose a strategy to augment textual data with entity description information, denoted as aug_T, where x_T is replaced by $aug_T(x_T)$ as input to f_T. By applying text augmentation, more detailed and precise textual information is input into the model, enhancing its ability to handle semantic ambiguities and ultimately improving its performance on multi-modal tasks.

Experiments

Experimental Setups

Datasets To address the limited scale, coverage of modal tasks, and scarcity of entity topics in existing datasets, we constructed a new dataset named M^3EL, which consists 79,625 instances. In total, we collected 318,500 images accompanied by textual captions. The M^3EL dataset has been divided into training, validation, and test sets in the ratio of 8:1:1.

To further verify the quality of the constructed dataset, we also investigated other existing multi-modal datasets:

- **DiffusionDB** (Wang et al. 2022b) is the first large-scale text-to-image cue dataset. We chose the subset "2m_random_1k", which contains 1000 image-text pairs, primarily aimed at the task of I_m-T_e.
- **WIKIPerson** (Sun et al. 2022) is a high-quality visual person linking dataset designed for Visual Named Entity Linking tasks. We selected the test set containing 6,142 images, along with their labels (Wikidata QIDs) and descriptions, primarily aimed at the task of I_m-T_e, I_m-I_e, and I_m-$(I{+}T)_e$.
- **WIKIDiverse** (Wang et al. 2022a) is a high-quality, manually annotated MEL dataset consisting of diverse contextual topics and entity types derived from Wikinews. We selected the test set comprising 1,570 image-caption pairs, primarily designed to implement T_m-T_e, I_m-I_e, $(I{+}T)_m$-I_e and $(I{+}T)_m$-$(I{+}T)_e$ tasks.

Baselines Given the diverse range of modality tasks encompassed by our proposed multi-modal dataset, the number of comparable models is notably limited. In order to establish a valid benchmark for evaluation, we report on the performance of several state-of-the-art methods for entity linking and visual entity recognition, including *ALIGN* (Jia et al. 2021), *BLIP-2* (Li et al. 2023), *CLIP* (Radford et al. 2021), *FLAVA* (Singh et al. 2022), *OWL-ViT* (Minderer et al. 2022), and *SigLIP* (Zhai et al. 2023). Additionally, we present our own optimized approach, *CLIP$_{ND}$*.

Implementation Details We employed the *CLIP* (*ViT-B-32*) architecture, which comprises both text and image encoders, each with a hidden state dimension of 768 and 12 multi-head attention mechanisms. We trained on an A100 GPU with 40GB of memory with 35 epochs and a batch of 256. The AdamW optimizer was applied with an initial learning rate of 1e-5, betas=(0.9, 0.98), eps=1e-6, and weight_decay=0.001.

Evaluation Metric The primary metric we evaluate is the accuracy (ACC) of linking entities to KG. Accuracy is defined by the following formula:

$$ACC = \frac{\sum_{i=1}^{N_{correct}} M_m - M_e}{\sum_{i=1}^{N} M_m - M_e}, \tag{3}$$

where $N_{correct}$ represent the number of correctly linked entity mentions, N represent the total number of mentions, M_m-M_e denotes different modalities of mention-entity linking task.

Experimental Results

Main Results In Table 4, we provide a comprehensive analysis of multi-modal task performance on the M^3EL. The experimental results are shown: Firstly, *CLIP$_{ND}$* achieves significant performance improvements over the baseline model on different versions of the multi-modal dataset. Specifically, on the *M-S* form, *CLIP$_{ND}$* demonstrated a 13.9% increase in accuracy compared to the baseline, with an average improvement of 7% across other datasets. Notably, *CLIP$_{ND}$* and *CLIP* share identical parameters and computational costs during training. Secondly, when auxiliary text information (e.g., entity names and descriptions) is added to the multi-modal task, the linking accuracy of the baseline model increases significantly, by an average of 1.2%. This suggests that the level of detail and clarity of the auxiliary text plays a key role in enhancing the model ability. Thirdly, all models consistently performed better when utilizing the *S$_1$-M-D* format for multi-modal tasks compared to the *S-M-D* format. This is mainly due to the *S$_1$-M-D* format's ability to convey more precise information about the entities. In contrast, the *S-M-D* format may truncate key entity and description information due to text length limitations, thereby negatively impacting model performance. Additionally, while most models perform better on multi-modal tasks that contain richer textual information (*S*format), for models focusing on visual modality such as *FLAVA* and *OWL-ViT*, more textual information may introduce noise and degrade link performance.

Further Analysis In Table 5, we report the accuracy of the model in detail. Based on the experimental results, we observe that our trained multi-modal approach, *CLIP$_{ND}$*, outperforms all baseline methods that rely solely on the

	S_1	S	M-S_1	M-S	S_1-M	S-M	S_1-M-D	S-M-D
ALIGN	0.674	0.675	0.683	0.683	0.680	0.681	0.701	0.697
BLIP-2	0.738	0.739	0.745	0.746	0.743	0.744	0.759	0.753
CLIP	0.742	0.748	0.754	0.758	0.749	0.753	0.771	0.763
FLAVA	0.654	0.649	0.661	0.655	0.659	0.653	0.681	0.699
OWL-ViT	0.487	0.487	0.495	0.494	0.489	0.488	0.490	0.492
SigLIP	0.711	0.714	0.722	0.724	0.715	0.717	0.736	0.726
$CLIP_{ND}$	**0.885**	**0.882**	**0.899**	**0.897**	**0.890**	**0.888**	**0.902**	**0.898**

Table 4: Experimental results of the baseline model on various M^3EL versions.The M^3EL dataset comes in various versions, each made up of composite text and images. *M* indicates the mention's surface, short for *Mention*. *S* its full context, short for *Sentence*. S_1 a single sentence with the mention, short for *Sentence*$_1$, and *D* the mention's description, short for *Description*. Formats include *M-S* (mention: sentence), *S-M* (sentence mention), and *S-M-D* (sentence mention: description), among others.

	Diffusion	WIKIPerson	WIKIDiverse	M^3EL
ALIGN	0.468	0.826	0.707	0.683
BLIP-2	0.552	0.652	0.677	0.746
CLIP	0.517	0.873	0.762	0.758
FLAVA	0.582	0.529	0.718	0.655
OWL-ViT	0.421	0.348	0.609	0.494
SigLIP	0.539	0.771	0.743	0.724
$CLIP_{ND}$	**0.724**	**0.916**	**0.796**	**0.897**

Table 5: Experimental results of the baseline models on each dataset. the best-performing model on each dataset is highlighted in bold, while the best performance of each model across different datasets is indicated with an underline.

raw modal information, especially in the dataset WIKIPerson, where the accuracy is improved by 4.3% to 56.8%. This significant performance boost is primarily attributed to the augmented feature strategy implemented during the training process. Secondly, except for the *BLIP-2* model, other models did not achieve the highest link accuracy on the M^3EL. Compared to the best-performing model, these models exhibited an average accuracy deficit of 7.6%. Lastly, the model performed the worst on the Diffusion dataset, due to the ambiguous mentions in the textual information paired with images, making linking more hard. Models performed most consistently in the WIKIDiverse, as the surface of mentions in the text-image pairs were more explicit. Most models performed best in the dataset WIKIPerson, due to the availability of bounding boxes within images where the mentions were located, allowing the models to access more specific and accurate image features. Overall, the M^3EL dataset combines the strengths of various datasets by providing clear mention and diverse text-image features, which helps enhance the generalization ability of multi-modal models.

Ablation Analysis We conducted an ablation study, and the experimental results are presented in Figure 3. It is evident that the model achieves optimal performance when it utilizes both mentions' name and description information as multi-modal context. In contrast, models trained using only a single source of information (e.g., only name or description) performed poorly. In addition, an interesting phenomenon was observed in the experiments: when the model used only sentences S_1 involving mentions as context, the performance was better than using complete sentences *S*. This suggests that models trained with text augmentation strategies have better comprehension in capturing semantic information related to mentions, whereas the use of longer texts may introduce noise, which reduces model performance.

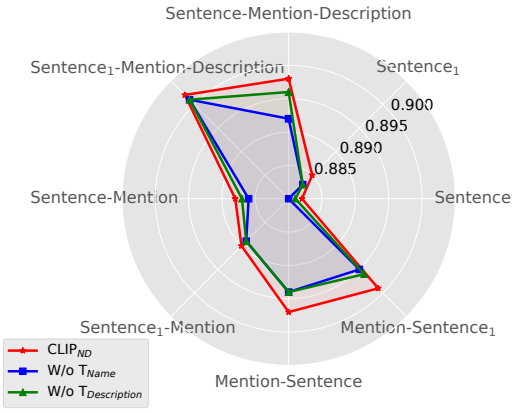


Figure 3: Ablation study to analyze modality absence of mention's different textual information. W/o T_{name} or $T_{description}$ stands for *CLIP*$_{ND}$ trained without the corresponding inputs.

Conclusion

We propose M^3EL, a large-scale multi-task and multi-topic dataset for multi-modal entity linking, encompassing 79,625 mentions across 5 different topics. This dataset provides detailed mention names, descriptions, contextual information, and 318.5K image-related data, covering 9 diverse multi-modal tasks. Compared to existing datasets, M^3EL has advantages in terms of entity type coverage, multi-modal task diversity, and good scalability, ultimately contributing to enhancing the performance of linking models in MEL tasks. Leveraging the rich textual information of M^3EL, we propose augmented modality strategies to train a model *CLIP*$_{ND}$ and effectively improve its performance in MEL. In future work, we will continue to update M^3EL to better support research in MEL. We will explore more complex real-world tasks, such as dynamic EL that requires complex linking in audio or video, and investigate the capabilities of M^3EL in large language models.

Acknowledgments

This work was supported by the National Key $R\&D$ Program of China (No.2022YFB3103600) and National Natural Science Foundation of China (No. 62476283, No. 62376284).

References

Adjali, O.; Besançon, R.; Ferret, O.; Le Borgne, H.; and Grau, B. 2020a. Building a Multimodal Entity Linking Dataset From Tweets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, 4285–4292. Marseille, France: European Language Resources Association. ISBN 979-10-95546-34-4.

Adjali, O.; Besançon, R.; Ferret, O.; Le Borgne, H.; and Grau, B. 2020b. Building a multimodal entity linking dataset from tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4285–4292.

Allam, A. M. N.; and Haggag, M. H. 2012. The question answering systems: A survey. *International Journal of Research and Reviews in Information Sciences (IJRRIS)*, 2(3).

Biten, A. F.; Gomez, L.; Rusinol, M.; and Karatzas, D. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12466–12475.

Broscheit, S. 2020. Investigating entity knowledge in BERT with simple neural end-to-end entity linking. *arXiv preprint arXiv:2003.05473*.

Chang, C.-H.; Kayed, M.; Girgis, M. R.; and Shaalan, K. F. 2006. A survey of web information extraction systems. *IEEE transactions on knowledge and data engineering*, 18(10): 1411–1428.

Chen, S.; Wang, J.; Jiang, F.; and Lin, C. 2020. Improving Entity Linking by Modeling Latent Entity Type Information. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, 7529–7537. AAAI Press.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Cucerzan, S. 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)*, 708–716. Prague, Czech Republic: Association for Computational Linguistics.

De Cao, N.; Izacard, G.; Riedel, S.; and Petroni, F. 2020. Autoregressive entity retrieval. *arXiv preprint arXiv:2010.00904*.

Gan, J.; Luo, J.; Wang, H.; Wang, S.; He, W.; and Huang, Q. 2021. Multimodal Entity Linking: a New Dataset and a Baseline. *Multimedia*.

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

Hoffart, J.; Yosef, M. A.; Bordino, I.; Fürstenau, H.; Pinkal, M.; Spaniol, M.; Taneva, B.; Thater, S.; and Weikum, G. 2011. Robust Disambiguation of Named Entities in Text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 782–792. Edinburgh, Scotland, UK.: Association for Computational Linguistics.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, 4904–4916. PMLR.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Liu, F.; Wang, Y.; Wang, T.; and Ordonez, V. 2020. Visual news: Benchmark and challenges in news image captioning. *arXiv preprint arXiv:2010.03743*.

Logeswaran, L.; Chang, M.-W.; Lee, K.; Toutanova, K.; Devlin, J.; and Lee, H. 2019. Zero-Shot Entity Linking by Reading Entity Descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 3449–3460. Florence, Italy: Association for Computational Linguistics.

Lu, D.; Neves, L.; Carvalho, V.; Zhang, N.; and Ji, H. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1990–1999.

Martinez-Rodriguez, J. L.; Hogan, A.; and Lopez-Arevalo, I. 2020. Information extraction meets the semantic web: a survey. *Semantic Web*, 11(2): 255–335.

Milne, D.; and Witten, I. H. 2008. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, 509–518.

Minderer, M.; Gritsenko, A.; Stone, A.; Neumann, M.; Weissenborn, D.; Dosovitskiy, A.; Mahendran, A.; Arnab, A.; Dehghani, M.; Shen, Z.; Wang, X.; Zhai, X.; Kipf, T.; and Houlsby, N. 2022. Simple Open-Vocabulary Object Detection with Vision Transformers. arXiv:2205.06230.

Mollá, D.; Van Zaanen, M.; and Smith, D. 2006. Named entity recognition for question answering. In *Australasian Language Technology Association Workshop*, 51–58. Australasian Language Technology Association.

Moon, S.; Neves, L.; and Carvalho, V. 2018. Multimodal Named Entity Disambiguation for Noisy Social Media Posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2000–2008. Melbourne, Australia: Association for Computational Linguistics.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rao, D.; McNamee, P.; and Dredze, M. 2013. Entity linking: Finding extracted entities in a knowledge base. *Multisource, multilingual information extraction and summarization*, 93–115.

Ratinov, L.; Roth, D.; Downey, D.; and Anderson, M. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 1375–1384. Portland, Oregon, USA: Association for Computational Linguistics.

Sevgili, Ö.; Shelmanov, A.; Arkhipov, M.; Panchenko, A.; and Biemann, C. 2022. Neural entity linking: A survey of models based on deep learning. *Semantic Web*, 13(3): 527–570.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15638–15650.

Sun, W.; Fan, Y.; Guo, J.; Zhang, R.; and Cheng, X. 2022. Visual named entity linking: A new dataset and a baseline. *arXiv preprint arXiv:2211.04872*.

Tran, A.; Mathews, A.; and Xie, L. 2020. Transform and tell: Entity-aware news image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13035–13045.

Wang, X.; Tian, J.; Gui, M.; Li, Z.; Wang, R.; Yan, M.; Chen, L.; and Xiao, Y. 2022a. WikiDiverse: a multimodal entity linking dataset with diversified contextual topics and entity types. *arXiv preprint arXiv:2204.06347*.

Wang, Z. J.; Montoya, E.; Munechika, D.; Yang, H.; Hoover, B.; and Chau, D. H. 2022b. DiffusionDB: A Large-Scale Prompt Gallery Dataset for Text-to-Image Generative Models. *arXiv:2210.14896 [cs]*.

Wu, L.; Petroni, F.; Josifoski, M.; Riedel, S.; and Zettlemoyer, L. 2019. Scalable zero-shot entity linking with dense entity retrieval. *arXiv preprint arXiv:1911.03814*.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2: 67–78.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11975–11986.

Zhang, L.; Li, Z.; and Yang, Q. 2021. Attention-Based Multimodal Entity Linking with High-Quality Images. In *International Conference on Database Systems for Advanced Applications*, 533–548. Springer.

Zhang, Q.; Fu, J.; Liu, X.; and Huang, X. 2018. Adaptive Co-attention Network for Named Entity Recognition in Tweets. In McIlraith, S. A.; and Weinberger, K. Q., eds., *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 5674–5681. AAAI Press.