WIKIDiverse: A Multimodal Entity Linking Dataset with Diversified Contextual Topics and Entity Types

Xuwu Wang[1], Junfeng Tian[2], Min Gui[3]*, Zhixu Li[1]✉, Rui Wang[4], Ming Yan[2], Lihan Chen[1], Yanghua Xiao[1,5]✉

[1] School of Computer Science, Fudan University, China
[2] Alibaba Group, China [3] Shopee, Singapore [4] Vipshop (China) Co., Ltd., China
[5] Fudan-Aishu Cognitive Intelligence Joint Research Center, China
{xwwang18,zhixuli,shawyh}@fudan.edu.cn,
{tjf141457, ym119608}@alibaba-inc.com,
min.gui@shopee.com, mars198356@hotmail.com, lhc825@gmail.com

Abstract

Multimodal Entity Linking (MEL) which aims at linking mentions with multimodal contexts to the referent entities from a knowledge base (e.g., Wikipedia), is an essential task for many multimodal applications. Although much attention has been paid to MEL, the shortcomings of existing MEL datasets including limited contextual topics and entity types, simplified mention ambiguity, and restricted availability, have caused great obstacles to the research and application of MEL. In this paper, we present WIKIDiverse, a high-quality human-annotated MEL dataset with diversified contextual topics and entity types from Wikinews, which uses Wikipedia as the corresponding knowledge base. A well-tailored annotation procedure is adopted to ensure the quality of the dataset. Based on WIKIDiverse, a sequence of well-designed MEL models with intra-modality and inter-modality attentions are implemented, which utilize the visual information of images more adequately than existing MEL models do. Extensive experimental analyses are conducted to investigate the contributions of different modalities in terms of MEL, facilitating the future research on this task. Our dataset and the baseline model are available at https://github.com/wangxw5/wikiDiverse.

1 Introduction

Entity linking (EL) has attracted increasing attention in the natural language processing community, which aims at linking ambiguous mentions to the referent unambiguous entities in a given knowledge base (KB) (Shen et al., 2014). It has been applied to a lot of downstream tasks such as information extraction (Yaghoobzadeh et al., 2016), question answering (Yih et al., 2015) and semantic search (Blanco et al., 2015).

*This work was conducted when Min Gui worked at Alibaba.

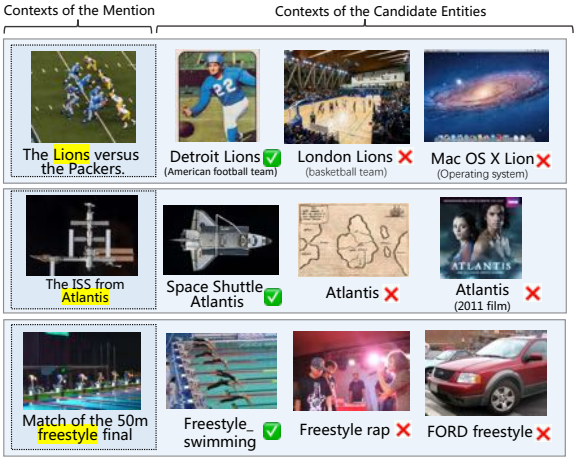


Figure 1: Several MEL examples, with the mention highlighted in the caption and the first entity of each entity list as the gold label.

As named entities (i.e., mentions) with multimodal contexts such as texts and images are ubiquitous in daily life, recent studies (Moon et al., 2018; Adjali et al., 2020a) turn their focus towards improving the performance of EL models through utilizing image information, i.e., **Multimodal Entity linking (MEL)**[1]. Several MEL examples are depicted in Figure 1, where the images could effectively help the disambiguation for entity mentions of different types. Due to its importance to many multimodal understanding tasks including VQA, multimodal retrieval, and the construction of multimodal KBs, much effort has been dedicated to the research of MEL. Moon et al. (2018) first addressed the MEL task under a zero-shot setting. Adjali et al. (2020a) designed a model to combine the visual, textual and statistical information for MEL. Zhang et al. (2021) designed a two-stage mechanism that

[1]In this paper, we focus on mentions coming from text spans and leave the visual mentions (i.e. objects from the images) for the future work.

Task	Dataset	Source	KB	Modality	Topic	Ent. Types	Manual	Open	Lang	Size
EL	AIDA(Hoffart et al., 2011)	News	Wikipedia	$T_m \rightarrow T_e$	Multiple	Multiple	✔	✔	en	1K docs
	MSNBC(Cucerzan, 2007)	News	Wikipedia	$T_m \rightarrow T_e$	Multiple	Multiple	✔	✔	en	20 docs
	AQUA(Milne and Witten, 2008)	News	Wikipedia	$T_m \rightarrow T_e$	Multiple	Multiple	✔	✔	en	50 docs
	ACE2004(Ratinov et al., 2011)	News	Wikipedia	$T_m \rightarrow T_e$	Multiple	Multiple	✔	✔	en	57 docs
	CWEB(Guo and Barbosa, 2018)	Web	Wikipedia	$T_m \rightarrow T_e$	Multiple	Multiple	✗	✔	en	320 docs
	WIKI(Guo and Barbosa, 2018)	Wiki	Wikipedia	$T_m \rightarrow T_e$	Multiple	Multiple	✗	✔	en	320 docs
	Zeshel(Logeswaran et al., 2019)	Wiki	Wikia	$T_m \rightarrow T_e$	Multiple	Multiple	✗	✔	en	-
MEL	Snap(Moon et al., 2018)	Social Media	Freebase	$T_m, V_m \rightarrow T_e$	Multiple	Multiple	✔	✗	en	12K captions
	Twitter(Adjali et al., 2020a)	Social Media	Twitter users	$T_m, V_m \rightarrow T_e, V_e$	Multiple	PER, ORG	✗	✗	en	4M tweets
	Movie(Gan et al., 2021)	Movie Reviews	Wikipedia	$T_m, V_m \rightarrow T_e, V_e$	Movie	PER	✔	✔	en	1K reviews
	Weibo(Zhang et al., 2021)	Social Media	Baidu Baike	$T_m, V_m \rightarrow T_e, V_e$	multiple	PER	✗	✔	cn	25K posts
	WIKIDiverse	News	Wikipedia	$T_m, V_m \rightarrow T_e, V_e$	Multiple	Multiple	✔	✔	en	8K captions

Table 1: Overview of EL and MEL datasets. T_m (T_e) and V_m (V_e) represent the textual and visual contexts of mentions m (or entities e) respectively, "Manual" denotes whether it is manually annotated, and "Open" denotes whether it is an open source.

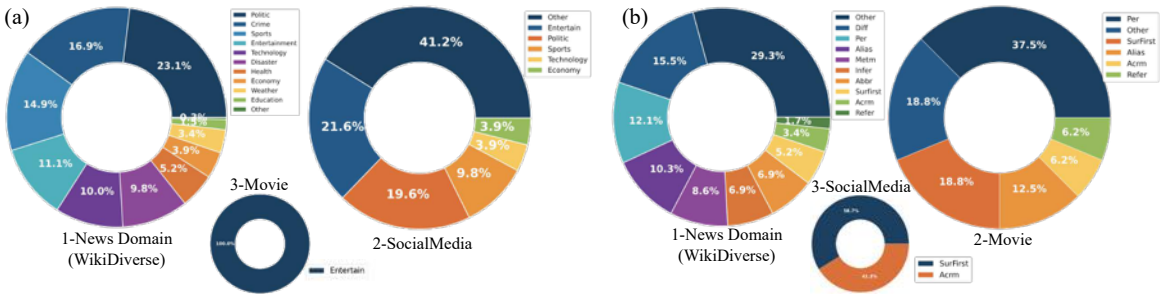


Figure 2: (a) compares the topic distribution of different domains. The statistics of social media are observed on sampled Twitter (Adjali et al., 2020a). The statistics of news domain are observed on WIKIDiverse. (b) compares the ambiguity distribution of different domains, where ten types of ambiguity are observed on our dataset, including different types of objects with the same name (Diff), persons with the same name (Per), alias, metonymy (Metm), inferring (Infer), surname or first name (SurFirst), acronym (Acrm), reference (Refer) and others.

first determines the relations between images and texts to remove negative impacts of noisy images and then performs the disambiguation. Gan et al. (2021) disambiguated visual mentions and textual mentions respectively at first, and then used graph matching to explore possible relations among intermodal mentions.

Although much attention has been paid to MEL, the existing MEL datasets as listed in the middle rows of Table 1 have deficiencies in the following aspects, which hinder the further advancement of research and application for MEL.

- **Limited Contextual Topics.** As shown in Figure 2(a), the existing MEL datasets are mainly collected from social media or movie reviews, where there are only 5 topics in the social media domain and 1 topic in the movie review domain. But as we observed in the news domain, there are more than 10 topics including other popular topics like disaster and education. The lack of topics would limit the generalization ability of the MEL model.

- **Limited Entity Types.** Entities in the existing MEL datasets mainly belong to the types of "Person (PER)" and "organization (ORG)". This restricts the application of the MEL models over other entity types such as locations, events and works, etc., which are also ubiquitous in common application scenarios.

- **Simplified Mention Ambiguity**: Some datasets such as Twitter (Adjali et al., 2020a) create artificial ambiguous mentions by replacing the original entity names with the surnames of persons or acronyms of organizations. Besides, limited entity types also lead to the limited mention ambiguity that only occurs with `PER` and `ORG`. According to our statistics of different domains as depicted in Figure 2(b), there are overall ten kinds of mention ambiguities in news domain such as Wikinews[2], while existing datasets collected from social media or movie reviews only cover a small scope of ambiguity.

[2]https://www.wikinews.org. It is a free-content news wiki.

- **Restricted Availability.** Most of the existing MEL datasets are not publicly available.

To enable more detailed research of MEL, we propose a manually annotated MEL dataset named WIKIDiverse with multiple topics and multiple entity types. It consists of 8K image-caption pairs collected from WikiNews and is based on the KB of Wikipedia with ~16M entities in total. Both the mentions and entities are characterized by multimodal contexts. We design a well-tailored annotation procedure to ensure the quality of WIKIDiverse and analyze the dataset from multiple perspectives (Section 4). Based on WIKIDiverse, we propose a sequence of MEL models with intra-modality and inter-modality attentions, which utilize the visual information of images more adequately than the existing MEL models (Section 5). Furthermore, extensive empirical experiments are conducted to analyze the contributions of different modalities for the MEL task and visual clues provided by the visual contexts (Section 6). In summary, the contributions of our work are as follows:

- We present a new manually annotated high-quality MEL dataset that covers diversified topics and entity types.
- Multiple well-designed MEL models with intra-modal attention and inter-modal attention are given which could utilize the visual information of images more adequately than the existing MEL models.
- Extensive empirical results quantitatively show the role of textual and visual modalities for MEL, and detailed analyses point out promising directions for the future research.

2 Related Work

Textual EL There is vast prior research on textual entity linking. Multiple datasets have been proposed over the years including the manually-annotated high-quality datasets like AIDA (Hoffart et al., 2011), automatically-annotated large-scale datasets like CWEB (Guo and Barbosa, 2018) and zero-shot datasets like Zeshel (Logeswaran et al., 2019). To evaluate the EL models' performance, it is usual to train on the AIDA-train dataset, and test on the datasets of AIDA-test, MSNBC(Cucerzan, 2007), AQUAINT(Milne and Witten, 2008), etc. However, as mentioned in (Cao et al., 2021), many methods have achieved high and similar results within recent three years. One possible explanation is that it may simply be near the ceiling of what can be achieved for these datasets, and it is difficult to conduct further research based on them.

Multimodal EL In recent years, the growing trend towards multimodality requires to extend the research of EL from monomodality to multimodality. Moon et al. (2018) first address the MEL task and build a zero-shot framework, which extracts textual, visual and lexical information for EL in social media posts. However, its proposed dataset is unavailable due to GDPR rules. Adjali et al. (2020a,b) propose a framework of automatically building the MEL dataset from Twitter. The dataset has limited entity types and ambiguity of mentions, thus it is not challenging enough. Zhang et al. (2021) study on a Chinese MEL dataset collected from the Chinese social media platform Weibo, which mainly focuses on the person entities. Gan et al. (2021) release a MEL dataset collected from movie reviews and propose to disambiguate both visual and textual mentions. This dataset mainly focuses on characters and persons of the movie domain. Peng (2021) propose three MEL datasets, which are built from Weibo, Wikipedia, and Richpedia information and use CNDBpedia, Wikidata and Richpedia as the corresponding KBs. However, using Wikipedia as the target dataset may lead to the data leakage problem as many language models are pretrained on it.

Our MEL dataset is also related to other named entity-related multimodal datasets, including entity-aware image caption datasets (Biten et al., 2019; Tran et al., 2020; Liu et al., 2021), multimodal NER datasets (Zhang et al., 2018; Lu et al., 2018), etc. However, the entities in these datasets are not linked to a unified KB. So our research of MEL can enhance the understanding of named entities, thereby enhancing the research in these areas.

3 Problem Formulation

Multimodal entity linking is defined as mapping a mention with multimodal contexts to its referent entity in a pre-defined multimodal KB. Since the boundary and granularity of mentions may be controversial, the mention span is usually pre-specified. Here we assume each mention has a corresponding entity in the KB, which is the *in-KB* evaluation problem.

Formally, let E represent the entity set of the KB,

which usually contains millions of entities. Each mention m or entity $e_i \in E$ is characterized by the corresponding visual context V_m, V_{e_i} and textual context T_m, T_{e_i}. Here T_m and T_{e_i} represent the textual spans around m and e_i respectively. V_m is the image correlated with m and V_{e_i} is the image of e_i in the KB. In real life, entities in KBs may contain more than one image. To simplify it, we select the first image of e_i as V_{e_i} and leave MEL with multiple images per entity as the future work. So the referent entity of mention m is predicted through:

$$e^*(m) = \underset{e_i \in E}{\arg\max} \ \Psi\left(m\left(T_m, V_m\right); e_i\left(T_{e_i}, V_{e_i}\right)\right).$$

where $\Psi(\cdot)$ represents the similarity score between the mention and entity.

4 Dataset Construction

In this section, we present the dataset construction procedure. Many factors including annotation quality, coverage of topics, diversity of entity types, coverage of ambiguity are taken into consideration to ensure the research value of WIKIDiverse.

4.1 Data Collection

Data Source Selection 1) For the source of image-text pairs, considering news articles are widely-studied in traditional EL (Hoffart et al., 2011; Cucerzan, 2007) and usually cover a wide range of topics and entity types, we decide to use news articles. Wikinews and BBC are two popular sources of news articles. So we compared them from two aspects. As shown in Table 2, Wikinews has advantages in terms of alignment degree between image-text pairs and MEL difficulty. So we select the image-caption pairs of Wikinews to build the corpus. 2) For the source of KB, we use the commonly-used Wikipedia (Hoffart et al., 2011; Ratinov et al., 2011; Guo and Barbosa, 2018). We also provide the annotation of the corresponding Wikidata entity for flexible studies.

Data Acquisition 1) For the image-caption pairs, we collect all the English news from the year 2007 to 2020 from Wikinews with multiple topics including sports, politics, entertainment, disaster, technology, crime, economy, education, health and weather. The data covers most of the common topics in the real world. Finally, we obtain a raw corpus with 14k image-caption pairs. 2) For the KB, we use the Wikipedia dump of 20210101. The entity set consists of all the entities in the main namespace with the size of ~16M.

Source	Alignment Degree with Image			MEL Difficulty		
	Caption	Headline	First Sent.	No	Easy	Hard
Wikinews	99%	30%	23%	1%	5%	94%
BBC	82%	53%	53%	2%	30%	68%

Table 2: Comparing the alignment degrees and corresponding MEL difficulty of image-caption, image-news headline, and image-first sentence between Wikinews and BBC, where the MEL difficulty is measured through the surface form similarity between mentions and entities.

Data Cleaning For the image-caption pairs, we remove the cases that 1) contain pornographic, profane, and violent content; 2) the text is shorter than 3 words. Finally, we get a corpus with 8K image-caption pairs.

4.2 Annotation

Annotation Design The primary goal of WIKIDiverse is to link mentions with multimodal contexts to the corresponding Wikipedia entity. Therefore, given an image-text pair, annotators need to 1) detect mentions from the text (Mention Detection, MD) and 2) label each detected mention with the corresponding entity in the form of a Wikipedia URL (Entity Linking, EL). For mentions that do not have corresponding entities in Wikipedia, they are labeled with "NIL". Seven common entity types (i.e., `Person`, `Organization`, `Location`, `Country`, `Event`, `Works`, `Misc`) are required to be annotated. To avoid subjective errors, we design detailed annotation guidelines with multiple samples to avoid the controversy of mention boundary, mention granularity, entity URL, etc. Details can be found in the Appendix. We also hold regular communications to discuss some emerging annotations problems.

Annotation Procedure The annotators include 13 annotators and 2 experienced experts. All annotators have linguistic knowledge and are instructed with detailed annotation principles. Each image-caption pair is independently annotated by two annotators. Then an experienced expert goes over the controversial annotations, and makes the final decision. Following Ding et al. (2021), we calculate the Cohen's Kappa to measure the agreements between two annotators. The Kappas of MD and

Figure 3: An example of WIKIDiverse. GT denotes the ground truth entity. The red words and blue words indicate the annotated entity type and Wikipedia entity respectively.

	Train	Dev.	Test	Total
# pairs	6311	755	757	7823
# ment. per pair	2.09	2.06	2.07	2.09
# words per pair	10.16	10.30	10.03	10.16

Table 3: Statistics of WIKIDiverse.

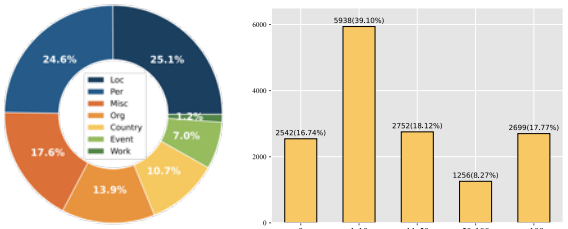
EL are 88.98% and 83.75% respectively, indicating a high degree of consistency.

4.3 Analysis of WIKIDiverse

Size and Distribution of WIKIDiverse We divide WIKIDiverse into training set, validation set, and test set with the ratio of 8:1:1. The statistics of WIKIDiverse are shown in Table 3. The collected Wikipedia KB has ~16M entities in total (i.e. $|E|$ ≈16M). Besides, we report the entity type distribution in Figure 4(a) and report the topic distribution in Figure 2(a).

Difficulty Measure Firstly, we compare surface form similarity of mentions and ground-truth entities. 51.31% of the mentions have different surface forms compared with ground-truth entities. Specifically, 16.05% of the mentions are totally different from the ground-truth entities. The large difference of the surface form brings challenges for MEL.

Secondly, we report the #candidate entities for each mention in Figure 4(b). Intuitively, the more entities a mention may refer to, the more ambiguous the mention is, and the more difficult the EL/MEL is. Specifically, we generate a $m \rightarrow e$ hash list based on the (m, e) co-occurrence statistics from Wikipedia (See Section 5.1 for details). As shown in Figure 4(b), we can see that 1) 44.2% mentions have more than 10 candidate entities. 2) 16.7% mentions are not contained in the hash list, which means their candidates are the entire entity



(a) Entity type distribution. (b) Distribution of # candidates per mention.

Figure 4: Other statistics of WIKIDiverse. (a) Entity type distribution. (b) Distribution of # cands per mention

set of the KB.

Thirdly, we randomly sample 200 image-caption pairs from WIKIDiverse to evaluate the diversity of ambiguity. As shown in Figure 2(b), WIKIDiverse covers a wide range of ambiguity.

5 Methods

It is challenging to directly predict the entity from a large-scale KB because it consumes large amounts of time and space resources. Therefore, following previous work (Yamada et al., 2016; Ganea and Hofmann, 2017; Cao et al., 2021), we split MEL into two steps: 1) candidate retrieval (CR) is first used to guarantee the recall and obtain a candidate entity set consisting of the TopK entities that are most similar to the mention; 2) entity disambiguation (ED) is then conducted to guarantee the precision and predict the entity with the highest matching score.

5.1 Candidate Retrieval

Existing methods (Yamada et al., 2016; Ganea and Hofmann, 2017; Le and Titov, 2018) mainly utilize two types of clues to generate the candidate entity set E_m: (I) the $m \rightarrow e$ hash list recording prior probabilities from mentions to entities: $P(e|m)$. (II) the similarity between the contexts of mention m and entity e.

Following these works, we implement a series of baselines as follows: (I) $\mathbf{P(e|m)}$ (Ganea and Hofmann, 2017): $P(e|m)$ is calculated based on 1) mention entity hyperlink count statistics from Wikipedia; 2) Wikipedia redirect pages; 3) Wikipedia disambiguation pages. (II) **Baselines of textual modality**: we retrieve the TopK candidate entities with the most similar textual context of the mention based on BM25 (Robertson and Zaragoza, 2009) (denoted as BM25), pretrained embeddings of words and entities obtained from (Yamada et al.,

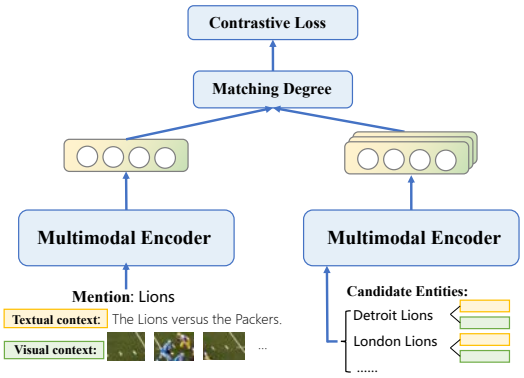


Figure 5: Framework of the introduced baselines.

2020) (denoted as WikiVec) and BLINK (Wu et al., 2020) (denoted as BLINK). (III) **Baseline of visual modality:** we retrieve the TopK candidate entities with the most similar visual contexts of the mention based on **CLIP** (Radford et al., 2021).

5.2 Contrastive Entity Disambiguation

The interaction between multimodal contexts of mentions and entities is complicated, which may bring noises to the model without careful handling. So we also introduce several baselines to explore the fusion of multimodal information.

The key component of ED is to design the function $\Psi(m; e_i)$ that quantifies the matching score between the mention m and every entity $e_i \in E_m$. As shown in Figure 5, the backbone of $\Psi(m; e_i)$ includes different multimodal encoders of m and e_i respectively, followed by dot-production to evaluate the matching degree between them. Specially, a multi-layer perceptron (MLP) is then used to combine the $P(e|m)$. Formally, e^* of m is predicted through:

$$\begin{aligned} \mathbf{m} =& \text{Encoder}_{\text{m}}(T_m, V_m); \mathbf{e_i} = \text{Encoder}_{\text{e}}(T_{e_i}, V_{e_i}) \\ e^* =& \arg\max_{e_i \in E_m} \text{MLP}\left(\mathbf{m} \odot \mathbf{e_i}, P(e_i|m)\right) \end{aligned} \tag{1}$$

So the multimodal encoders of mentions and entities are the most significant parts of MEL. They use the same structure but training with different parameters.

Multimodal Encoder Firstly, we get the textual context's embeddings. For the mention's textual context $T_m = \{w_1, \ldots, w_{L_1}\}$, we directly embed it with the word embedding layer of BERT (Devlin et al., 2019). While for e_i, we embed it as the pre-trained embeddings from Yamada et al. (2020), which have compressed the semantics of e_i's entire contexts from Wikipedia.

$$\{\hat{\mathbf{w}}_1, ..., \hat{\mathbf{w}}_{L_1}\} = \text{BERT}_{EMB}(T_m) \tag{2}$$

Secondly, we get the visual context embeddings. Instead of the widely used region-based visual features, we adopt grid features following (Huang et al., 2020), which has the advantage of end-to-end. Specifically, the visual features are represented with the grid features from :

$$\{\hat{\mathbf{v}}_1, ..., \hat{\mathbf{v}}_{L_2}\} = \text{Flat}(\text{ResNet}(V)) \tag{3}$$

where $\text{Flat}(\cdot)$ represents flatting the feature along the spatial dimension and L_2 indicates the number of grid features.

Finally, taking the embeddings of the two modalities as inputs, we capture the interaction between them. We adopt several backbones to fuse multiple modalities. 1) **UNITER** (Chen et al., 2020): the two modalities are concatenated and then fed into self-attention transformers to fuse them together. 2) **UNITER***: we apply separate self-attention transformers to the two modalities before UNITER for better feature extraction of each modality. 3) **LXMERT** (Tan and Bansal, 2019): the two modalities are fed into separate self-attention transformers at first and then interact with cross-modal attention. The design of intra-modal and inter-modal attention helps better alignment and interaction of multiple modalities.

After multiple layers of the fusion operation: Fuse $(\{\hat{\mathbf{w}}_1, ..., \hat{\mathbf{w}}_{L_1}\}, \{\hat{\mathbf{v}}_1, ..., \hat{\mathbf{v}}_{L_2}\})$, the hidden states of the mention's tokens $\{\mathbf{h}_i, ..., \mathbf{h}_j\}$ are obtained. Then we concatenate the hidden states of the first and the last tokens and feed them into a MLP to get the mention's embeddings: $\text{MLP}\left([\mathbf{h}_i||\mathbf{h}_j]\right)$

Contrastive Loss We introduce contrastive learning (Karpukhin et al., 2020; Gao et al., 2021) to learn a more robust representation of both mentions and entities. It is widely acknowledged that selecting negative examples could be decisive for learning a good model. To this end, we utilize both hard negatives and in-batch negatives to improve our model's ability to distinguish between gold entities and hard/general negatives. Let $e_{i,j}$ represent the j^{th} candidate entity of the i^{th} mention in a batch and let P_i denote the index of m_i's gold entity. The hard negatives are the other $K-1$ candidate entities retrieved in CR step except for the gold entity: $\{e^-_{i,k}\}^{k\in[1,K]}_{k\neq P_i}$. The in-batch negatives

Modality	Method	R@10	R@50	R@100
P	$P(e\|m)$	80.82	85.48	86.23
T	BM25	39.66	48.49	51.85
T	WikiVec	14.73	20.27	22.60
T	BLINK	63.63	73.15	76.03
V	CLIP	17.05	27.26	31.30
T+V*	BLINK+CLIP	66.96	77.18	80.53
P+V*	$P(e\|m)$+CLIP	85.26	90.27	91.30
P+T*	$P(e\|m)$+BLINK	86.36	91.78	93.21
P+T+V*	$P(e\|m)$+BLINK+CLIP	**86.37**	**91.91**	**93.35**

Table 4: Performance of candidate retrieval. R@K represents recall of the TopK retrieved entities. The modality of P, T, V represent the $P(e|m)$, textual context and visual context respectively. T+V and P+T+V represent the ensemble of different sub-methods. Results with * are generated using grid search over the Dev. dataset to find the best combination of different sub-methods.

are gold entities of other $B-1$ mentions in the mini-batch: $\{e^+_{b,P_b}\}^{b\in[1,B]}_{b\neq i}$, where B represents the batch size. The optimization objective is defined as the negative log likelihood of the ground-truth entity:

$$\mathcal{L}(m_i, E_{m_i}) = -\log \frac{e^{\Psi(m_i, e^+_{i,P_i})}}{e^{\Psi(m_i, e^+_{i,P_i})} + \sum^-}$$

$$\sum^- = \underbrace{\sum_{k=1,k\neq P_i}^{K} e^{\Psi(m_i, e^-_{i,k})}}_{\text{hard negatives}} + \underbrace{\sum_{b=1,b\neq i}^{B} e^{\Psi(m_i, e^+_{b,P_b})}}_{\text{in-batch negatives}} \quad (4)$$

Besides the above baselines, we also compare with the following classic baselines: 1) **Baselines of Textual Modality** include REL (Le and Titov, 2018), BERT (Devlin et al., 2019), BERT* enhanced with Yamada et al. (2020)'s pre-trained vectors and BLINK (Wu et al., 2020). 2) **Baselines of Visual Modality** include ResNet-50 and CLIP. 3) **Multimodal Baselines** include MMEL18 (Moon et al., 2018), MMEL20 (Adjali et al., 2020b). Details of the baselines can be found in Appendix **??**.

6 Experimental Results

6.1 Candidate Retrieval Results

As shown in Table 4: 1) Our model achieves 93.35% of R@100, which indicates most related entities can be recalled from the large 16M KB. For retrieval, each mention takes about 12ms of P(e|m), 40ms of BM25, 183ms of WikiVec and CLIP, 60ms of BLINK; 2) As for ensemble of different modalities, T + V achieves better results than V and T, which verifies that the information of different modalities are complementary;

In practice, we use grid search over the Dev. to find the best combination of different modalities. For example, when $K = 10$, the best E_m is generated with 80%P+ 10%T + 10%V.

Modality	Model	F1	P	R
T→ T	REL	60.52	67.93	54.56
	BLINK	66.74	70.93	63.03
	BERT	57.80	74.92	47.06
	BERT*	62.21	84.36	49.27
V → T	ResNet-50	34.01	44.06	27.69
	CLIP	35.26	36.68	33.41
T+V→ T	MMEL18	51.22	53.27	48.78
T+V→ T+V	MMEL20	37.44	38.48	36.46
	UNITER	69.37	73.72	65.51
	UNITER*	70.60	75.03	66.66
	LXMERT	72.19	76.72	68.17
	UNITER †	71.07	75.52	67.10
	UNITER* †	71.15	75.61	67.18
	LXMERT †	**73.22**	**77.81**	**69.14**

Table 5: Comparison with baselines with results averaged over 5 runs. Models with † are enhanced with contrastive learning. All the models use the same candidate entity set retrieved through $P(e|m)$+BLINK+CLIP with $K = 10$.

6.2 Entity Disambiguation Results

Following previous work, we report micro F_1, precision, recall in Table 5. According to the experimental results, we can see that: First, the proposed multimodal methods outperform all the methods with a single modality, which benefit from multimodal contexts. Besides, contrastive learning can even improve the performance. We reckon that contrastive learning improves the ability to distinguish entities. Second, the performance of textual baselines is better than the visual ones, which indicates the textual context still plays a dominant role in MEL. Third, the methods using transformers to model the interaction between modalities perform better than those with simple interaction (Moon et al., 2018; Adjali et al., 2020a), which verifies the importance of fusing different modalities.

Type	Image of m	Caption of m	Image of e*	e*
Object		The former Birka Princess in 2005, now called M/S Sea Diamond		MS_Sea_Diamond
Scene		Extra Match of the Women 50m freestyle finale.		Freestyle_swimming
Property		Vials of the COVID-19 vaccine with labels in English.		COVID-19
Scene & Property		The men's 400 m T53		Wheelchair_racing

Figure 6: Examples of the 'Visual Clues'.

6.3 Multimodal Analysis

The following experiments are performed on the ED task.

Are the multiple modalities complementary? We draw a Venn diagram of different modalities in Figure 8. The circle of Method i is calculated through $\frac{\#Hit_i}{|Dataset|}$ and the interaction of two circles are calculated through $\frac{\#(Hit_i \cap Hit_j)}{|Dataset|}$. One can see that the textual modality is dominant, while the visual modality provides complementary information. Specially, the multimodal method predicts more new entities of 11.56%, which verifies the importance of fusing two modalities.

Is it better to have multimodal contexts of both mentions and entities? We conduct an ablation study and report the experimental results in Table 6. We can see that the model with multimodal contexts of both mentions and entities achieves the best result. So linking multimodal mentions to multimodal entities is better than linking multimodal mentions to mono-modal entities as done in (Moon et al., 2018).

What visual clues are provided by the visual contexts? We randomly select 800 image-caption pairs from the test dataset, and then ask annotators to label each mention with the types of visual clues. The visual clues include 4 types: 1) **Object**: the image contains the entity object. 2) **Scene**: the image reveals the scene that the entity belongs to (e.g. a basketball player of the 'basketball game' scene). 3) **Property**: the image contains some properties of the entity (e.g. an American flag reveals the property of a person's nationality). 4) **Others**: other important contexts. Note that the four types of clues can be crossed and a sample could have no clues. We find that visual context is helpful for 60.54% mentions and 81.56% image-caption pairs. We report the contribution of different types of visual clues in Table 7. One can see that: 1) For scene clues and object clues, the T+V is 4.88% and 3.53% higher than T. So the multimodal model benefits a lot from the information of scene and objects in the images. 2) For property clues, the T+V achieves similar results with T, which shows fine-grained visual clues are not used well and indicates the direction of future research.

Model	F1	P	R
LXMERT	68.65	72.96	64.83
w/o V_m	64.14	68.16	60.56
w/o V_e	59.01	76.44	48.05
w/o V_m and V_e	62.21	84.36	49.27
w/o T_m	58.67	76.01	47.77
w/o T_e	45.66	48.53	43.12
w/o T_m and T_e	34.01	44.06	27.69

Table 6: Ablation study to analyze modality absence of mention and entity. W/o $T_{m/e}$ or $V_{m/e}$ stands for LXMERT trained without the corresponding inputs.

6.4 Case Study

We present several examples where multimodal contexts influence MEL in Figure 7. Example (a), (b) verify the helpfulness of the multimodal context.

Figure 7: Case study. Successful predictions and failed predictions for the underlined mention are shown.

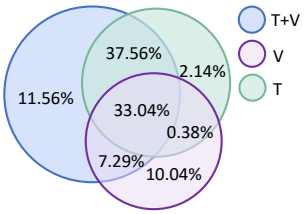


Figure 8: Venn diagram illustration of contributions of different modalities. We remove the input of the corresponding modality of LXMERT to get the results without re-training the model. To avoid the interference of $P(e|m)$, we also remove it from the model.

Visual Clues	Proportion	F_1	
		T	T+V
Object	45.40%	69.16	72.69
Scene	18.96%	62.80	67.68
Property	26.22%	70.37	70.83
Others	14.80%	71.43	72.73

Table 7: Model performance under different visual clues. T+V denotes the multimodal model LXMERT, and T represents the textual model BERT*.

From the error cases, we can see that the model still lacks such capabilities: 1) Eliminate the influence of unhelpful images (e.g., Example (c)); 2) Perform reasoning (e.g., inferring the "white house" from Example (d)'s image); 3) Alleviate over-reliance on P(e) (e.g., Example (e)).

7 Conclusion and Future Work

We propose WIKIDiverse, a manually-annotated Wikipedia-based MEL dataset collected from Wikinews. To overcome the weaknesses of existing datasets, WIKIDiverse covers a wide range of topics, entity types and ambiguity. We implement a series of baselines and carry out multiple experiments over the dataset. According to the experimental results, WIKIDiverse is a challenging dataset worth further exploration. Besides multimodal entity linking, WIKIDiverse can also be applied to evaluate the pre-trained language model, multimodal named entity typing/recognition, multimodal topic classification, etc. In the future, we plan to 1) utilize more than one images of each entity 2) adopt finer-grained multimodal interaction models for this task and 3) transfer the model to more general scenarios such as EL in articles.

Acknowledgement

This research was supported by the National Key Research and Development Project (No. 2020AAA0109302), National Natural Science Foundation of China (No. 62072323), Shanghai Science and Technology Innovation Action Plan (No. 19511120400), Shanghai Municipal Science and Technology Major Project (No. 2021SHZDZX0103) and Alibaba Research Intern Program.

Ethical Considerations

We collected publicly available Wikinews image-caption pairs without storing any personal data. During data cleaning, we remove the cases that contain pornographic, profane, and violent content.

We annotate the data using the crowdsourcing platform of Alibaba. To ensure that the crowd workers were fairly compensated, we paid them at an hourly rate of 15 USD per hour, which is a fair and reasonable rate of pay for crowdsourcing.

References

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020a. Building a multimodal entity linking dataset from tweets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4285–4292, Marseille, France. European Language Resources Association.

Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020b. Multimodal entity linking for tweets. In *Advances in Information Retrieval*, pages 463–478, Cham. Springer International Publishing.

Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 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*, pages 12466–12475.

Roi Blanco, Giuseppe Ottaviano, and Edgar Meij. 2015. Fast and space-efficient entity linking for queries. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 179–188. ACM.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Silviu Cucerzan. 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)*, pages 708–716, Prague, Czech Republic. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Hai-Tao Zheng, and Zhiyuan Liu. 2021. Few-nerd: A few-shot named entity recognition dataset. In *ACL-IJCNLP*.

Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. 2021. Multimodal entity linking: a new dataset and a baseline. *Multimedia*.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2619–2629, Copenhagen, Denmark. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

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

Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. 2011. Robust disambiguation of named entities in text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 782–792, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1595–1604, Melbourne, Australia. Association for Computational Linguistics.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3449–3460, Florence, Italy. Association for Computational Linguistics.

Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji. 2018. Visual attention model for name tagging in multimodal social media. In *Proceedings of ACL*, pages 1990–1999, Melbourne, Australia.

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

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 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)*, pages 2000–2008, Melbourne, Australia. Association for Computational Linguistics.

Wang Peng. 2021. Multimodal entity linking datasets benchmark.

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. *arXiv preprint arXiv:2103.00020*.

Lev Ratinov, Dan Roth, Doug Downey, and Mike Anderson. 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*, pages 1375–1384, Portland, Oregon, USA. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. *The probabilistic relevance framework: BM25 and beyond*. Now Publishers Inc.

Wei Shen, Jianyong Wang, and Jiawei Han. 2014. Entity linking with a knowledge base: Issues, techniques, and solutions. *IEEE Transactions on Knowledge and Data Engineering*, 27(2):443–460.

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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13035–13045.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Yadollah Yaghoobzadeh, Heike Adel, and Hinrich Schütze. 2016. Noise mitigation fordu neural entity typing and relation extraction. *arXiv preprint arXiv:1612.07495*.

Ikuya Yamada, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. Wikipedia2Vec: An efficient toolkit for learning and visualizing the embeddings of words and entities from Wikipedia. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 23–30. Association for Computational Linguistics.

Ikuya Yamada, Hiroyuki Shindo, Hideaki Takeda, and Yoshiyasu Takefuji. 2016. Joint learning of the embedding of words and entities for named entity disambiguation. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 250–259, Berlin, Germany. Association for Computational Linguistics.

Scott Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao. 2015. Semantic parsing via staged query graph generation: Question answering with knowledge base. *Proceedings of the Joint Conference of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing of the AFNLP*.

Li Zhang, Zhixu Li, and Qiang Yang. 2021. Attention-based multimodal entity linking with high-quality images. In *International Conference on Database Systems for Advanced Applications*, pages 533–548. Springer.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of AAAI*.