Multimodal Knowledge Learning for Named Entity Disambiguation

Dongjie Zhang, Longtao Huang, Ting Ma and **Hui Xue**
Alibaba Group
{yurui.zdj, kaiyang.hlt, mating.ma, hui.xueh}@alibaba-inc.com

Abstract

With the popularity of online social media, massive-scale multimodal information has brought new challenges to traditional Named Entity Disambiguation (NED) tasks. Recently, Multimodal Named Entity Disambiguation (MNED) has been proposed to link ambiguous mentions with the textual and visual contexts to a predefined knowledge graph. Existing attempts usually perform MNED by annotating multimodal mentions and adding multimodal features to traditional NED models. However, these studies may suffer from 1) failing to model multimodal information at the knowledge level, and 2) lacking multimodal annotation data against the large-scale unlabeled corpus. In this paper, we explore a pioneer study on leveraging multimodal knowledge learning to address the MNED task. Specifically, we first harvest multimodal knowledge in the Meta-Learning way, which is much easier than collecting ambiguous mention corpus. Then we design a knowledge-guided transfer learning strategy to extract unified representation from different modalities. Finally, we propose an Interactive Multimodal Learning Network (IMN) to fully utilize the multimodal information on both the mention and knowledge sides. Extensive experiments conducted on two public MNED datasets demonstrate that the proposed method achieves improvements over the state-of-the-art multimodal methods.

1 Introduction

Nowadays, online social media have become more and more important in our daily life. The massive-scale blogs posted on these social media hide valuable information that can be used to understand users and distill user preferences. However, how to extract valuable information is extremely challenging because the posts are always free-form, especially the text. Named Entity Disambiguation (NED) is such a critical task for extracting structured information, aiming to map ambiguous mentions from free-form texts to specific entities in a predefined knowledge graph. NED can benefit many downstream applications, such as information retrieval (Chen et al., 2021), question answering (Kandasamy and Cherukuri, 2020), relation extraction (Nguyen et al., 2017), etc.

Existing research on NED mainly focuses on texts and has been proven to be successful for well-formed texts. However, with the popularity of incorporating a mix of text and images on social media platforms (e.g. Twitter[1], Instargram[2], Snapchat[3], etc.), more ambiguous mentions appear in the short and noisy text. Due to the enormous number of mentions arising from incomplete and inconsistent expressions, the traditional text-only NED methods are limited in dealing with cross-modal ambiguity, making it difficult to link these mentions accurately. For example, on one hand, it is difficult to distinguish the mention *Swift* refers to **Taylor Swift** or **Ben Swift** from the textual context in Fig 1. On the other hand, due to the obstruction of eyes, hats, and other objects, the target person cannot be directly recognized from the image alone through face recognition techniques. When multimodal contexts in the post, as well as the historical knowledge, are combined, the correct entity **Ben Swift** can be predicted from the candidates. That is, the textual and visual features can complement each other.

Although some recent methods have achieved promising performance for the MNED task (Moon et al., 2018; Adjali et al., 2020a,b), challenges still exist. First, sufficient annotated corpus with texts and images is required to train a multimodal model, which is costly in practice (Abuczki and Ghazaleh, 2013). And lacking sufficient training data would limit the performance of neural models. Second, previous works mainly learn from the multimodal

[1]https://twitter.com/
[2]https://www.instagram.com/
[3]https://weibo.com/

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 3160 - 3169
December 7-11, 2022 ©2022 Association for Computational Linguistics

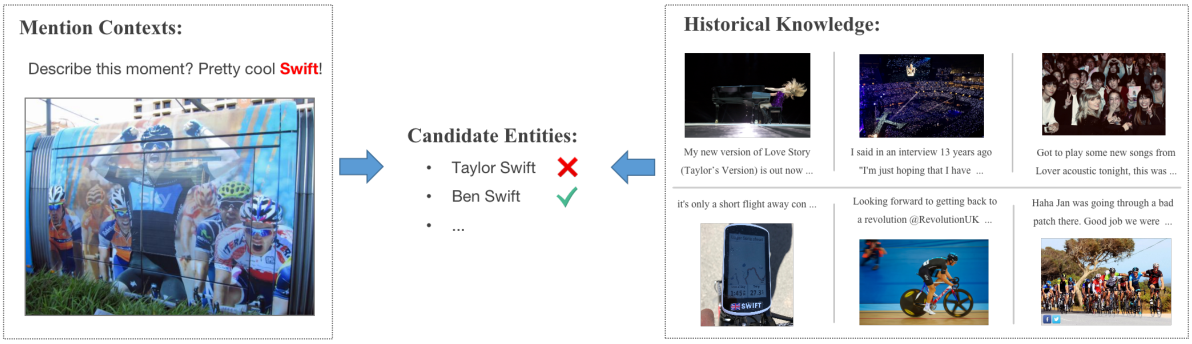

Figure 1: An example of named entity disambiguation. Because of the insufficiency of information, the mention *Swift* is ambiguous only from the textual context. And the correct entity *Ben Swift* can be disambiguated by considering multimodal contexts in the post and historical knowledge.

mention contexts and do not exploit available information at the knowledge level, without harvesting useful descriptions and historical data with visual features.

In this paper, we focus on solving MNED tasks at the knowledge level. To reduce the dependence on annotated data and fully use the unsupervised multimodal corpus, we firstly train a multimodal feature extractor by implementing a knowledge-guided transfer learning strategy. Then we enrich multimodal information at the knowledge level using a Meta-Learning aggregation strategy, aiming to obtain multimodal entities and mentions using a small number of knowledge annotations. Finally, we design an **I**nteractive **M**ultimodal learning **N**etwork (IMN) to flexibly utilize the multimodal information from both mention contexts and knowledge graph and integrate them.

The main contributions of this paper are summarized as follows:

- We propose a Meta-Learning method to utilize multimodal information at the knowledge level, and perform a knowledge-guided pre-training model to reduce the dependence on annotated data. To the best of our knowledge, this is the first work to introduce a multimodal pre-training model in the MNED task.

- We design an Interactive Multimodal Learning Network (IMN) to fully utilize the multimodal information on both the mention as well as knowledge sides.

- Comparative experiments conducted on two public MNED datasets show the proposed method outperforms state-of-the-art MNED methods.

The rest of the paper is organized as follows: In Section 2, we summarize the related work. In Section 3, we formulate the MNED task and introduce the proposed method in detail. In Section 4, we conduct extensive experiments and analyses. Finally, we conclude this work in Section 5.

2 Related Work

Multimodal Learning As an efficient mechanism of leveraging contextual information from multiple modalities in parallel, multimodal learning has been applied in a wide range of tasks in recent years (Elliott et al., 2015; Specia et al., 2016). In previous works, representation of different modalities was mostly obtained separately. For visual representation, CNN-based models such as VGG (Simonyan and Zisserman, 2014) , Google Inception (Szegedy et al., 2016), ResNet (He et al., 2016) are widely adopted in many multimodal tasks. Textual features are mostly represented by language models such as GloVe (Pennington et al., 2014), GPT (Radford et al., 2018), XLNet (Yang et al., 2019) etc. Recently, with the success of pre-training and self-supervised learning (Misra et al., 2016; Xie et al., 2017b), several multimodal transfer learning methods and architectures (Yu et al., 2021; Gao et al., 2020; Lu et al., 2019b; Qi et al., 2020) have been proposed, and have achieved state-of-the-art results on various vision language tasks, including Visual Question Answering, Visual Commonsense Reasoning, Region-to-Phrase Grounding, Image-text Retrieval, etc. VideoBERT (Sun et al., 2019) learns joint distributions over sequences of visual and linguistic tokens as multimodal features. Vision-and-Language BERTs (Lu et al., 2020, 2019a; Gao et al., 2020) extend BERT architecture to adapt multimodal input by extracting RoIs from images and regards as image tokens. Although these pre-training models can learn unsupervised features in unsupervised corpus, they still need further improvement in tasks that require additional knowledge. And we argue that the self-supervised models still requires guidance of knowledge.

Named Entity Disambiguation Traditional NED methods mainly focus on text-only corpus which can be divided into two categories, local methods and global methods (Barrena et al., 2018; Ganea and Hofmann, 2017). For local methods, each mention is disambiguated separately via hand-crafted features (Bunescu and Paşca, 2006; Mihalcea and Csomai, 2007) and contextual representations learned by neural networks (He et al., 2013; Eshel et al., 2017). Global methods(Nguyen et al., 2016; Le and Titov, 2018) jointly disambiguate mentions by taking into account the topical coherence among the referred entities in the same document(Fang et al., 2019). For the MNED task, the work from (Moon et al., 2018) is the first to utilize multimodal mention contexts via weighting the embeddings of images and words based on attention mechanism. The previous multimodal works primarily depend on sufficient training data with fully annotations on all mention modalities which is costly in practice(Abuczki and Ghazaleh, 2013). Although Moon et al. (2018) involve a zero-shot layer in their model to allow for disambiguation of unseen entities during training, the performance is limited if the multimodal information is incomplete in the training data. Inspired by recent success on multimodal knowledge graph (Xie et al., 2017a; Mousselly-Sergieh et al., 2018; Pezeshkpour et al., 2018),we aim at handle MNED tasks at the knowledge level, which is much easier than collecting and annotating multimodal corpus.

3 Proposed Method

3.1 Task Definition

Formally, the inputs are a set of multimodal posts $P = \{p^{(1)}, p^{(2)}, ..., p^{(n)}\}$ and a predefined knowledge graph $G = (E, R, H)$ that is composed of the entity set E, the relation set R and historical data of entities. Each input post $p \in P$ is denoted as $p = \{p_m, p_t, p_v\}$, where p_m is a mention that needs to be disambiguated, p_t is a sequence of words surrounding the mention in the post, and p_v is an image associated in the post. Note that the mention p_m can be obtained by other tasks such as Named Entity Recognition (Lample et al., 2016), which is beyond the scope of this paper. Then the target of MNED is to find the ground truth entity $e^+ \in E$ that p_m corresponds to.

3.2 Knowledge-Guided Pre-training Model

Before dealing with the input multimodal posts, we firstly build a pre-trained model to capture the inherent relationship between images and texts which is guided by the knowledge graph. In this transfer learning way, the model can better understand the content of different modalities and is helpful to overcome insufficiency of annotated multimodal corpus.

Knowledge Pre-training Architecture The pre-training model is composed of five parts, textual representation, visual representation, mention embedding, transformer encoder and training with adaptive loss. The multimodal inputs consist of textual and visual representation which is tokenized into a token and patch sequence according to Word-Pieces and Object Detection methods. We use the standard BERT(Devlin et al., 2018) pre-process method to get the textual sequence. Unlike traditional pipeline image representation techniques, we use an end-to-end method to obtain the visual representation. DEtection TRansformer(DETR)(Carion et al., 2020) approaches object detection as a direct set prediction problem which directly output the final set of objects in parallel. Given an input image, we take the fixed-length vector sequence of the output layer of DETR decoder as the visual representation. Each of the vectors corresponds to one image patch, we regard each patch as an "patch token". Mention embedding is initialized by Glove (Pennington et al., 2014).

The concatenation of the text token sequence, mention embedding and image patch sequence consists of the pre-training model inputs. Similar to (Gao et al., 2020),we adopt a pre-trained standard Transformer (Vaswani et al., 2017) as the matching backbone network of the pre-training model. The information of text tokens and image patches thus interact freely in multiple self attention layers. In order to ensure the multimodal comprehension ability as well as sensitiveness at the knowledge of the pre-training model, we mask mention tokens with a probability of 85% instead of random word masking.

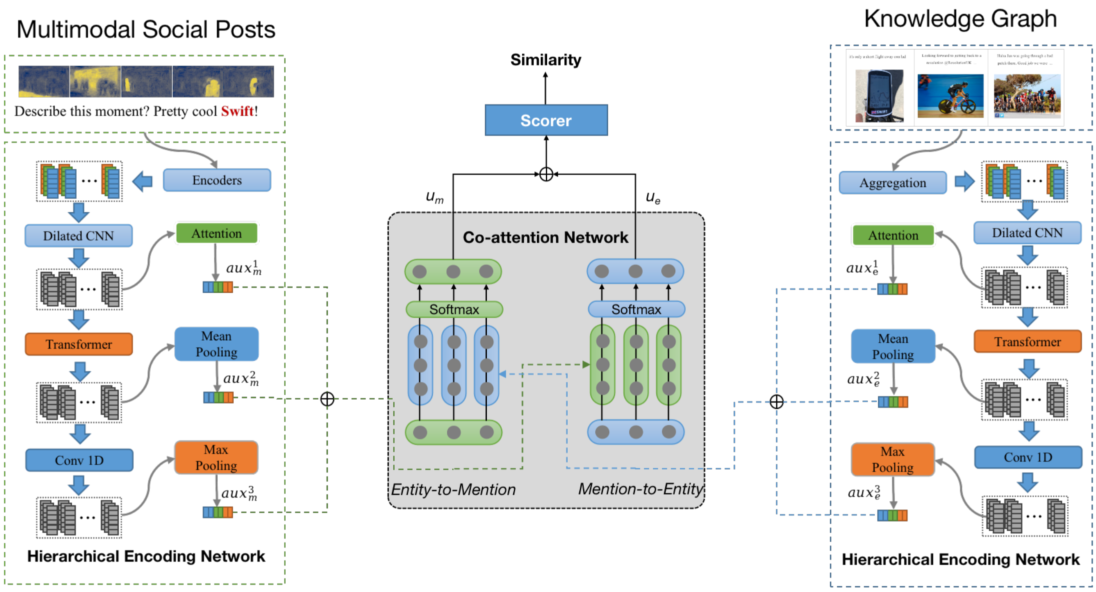

Figure 2: The overview of the IMN with hierarchical encoding network and co-attention network. The hierarchical encoding network contains a Dilated CNN layer, a Transformer layer and a Conv1D layer which map multimodal posts and entities to three levels of auxiliary spaces. Then a co-attention network is proposed to explicitly emphasize the cross-modal interactive features between mentions and entities using collaborative attention mechanism. Finally, a scorer function is applied to get similarity of mentions and entities.

3.3 Knowledge Prototype Construction

In spite of the multimodal mention contexts, we believe that multi-modal information at the knowledge level is potentially important for MNED tasks. Different from the previous textual representation methods, we prefer to establish multimodal representation at the knowledge level. Given an entity associated with many related historical posts containing images and texts, we simply select a part of the representative timeline tweets as the prototype. Specifically, we adopt three modalities representations to depict an entity based on timeline posts. The visual prototype of each entity e_v is acquired by aggregating the features of the k representative corresponding images. And features of an image can generated by many image identification such as DETR (Carion et al., 2020). Similarly, the textual prototype of each entity e_t is acquired by pre-trained language models such as BERT (Devlin et al., 2018). Meanwhile, the joint prototype of each entity e_o can be acquired by the hidden state of the knowledge-guided pre-training model described in previous subsections.

To select most representative support set from a large number of historical data, we build a similarity graph for each modality. The vertexes of the similarity graph are feature vectors obtained in previous steps. And the edges are the cosine similarity between the vertexes. Then top-k representative results are acquired by calculating the PageRank score (Page et al., 1999) of each vertex in the similarity graph. The multimodal prototypes of an entity can be acquired by the top-k PageRank vertexes, and we perform L2 regularization on each prototype. Finally, each entity is represented to three different modalities $e = \{e_v, e_t, e_o\}$ with the fixed length k.

For the multimodal posts, three different feature extractors is applied to obtain mention embeddings. For each post $p = \{p_m, p_t, p_v\}$, the visual embedding m_v and textual embedding m_t is generated by the same method used in entity representation process. The joint embedding m_j of the mention p_m is acquired by pre-trained model in section 3.2. Thus, each mention is embedded to three modalities $m = \{m_v, m_t, m_o\}$.

3.4 Interactive Meta Learning Network

The architecture of IMN is shown in Figure 2. Because of the huge semantic gap between different modalities, it is challenging to disambiguate entities in a high level embedding space (Liu et al., 2021). We first construct a hierarchical encoding network for mentions and entities, and design some unified auxiliary spaces, which are mapped by the outputs of encoders at different levels. Then a co-attention network is utilized to explicitly emphasize the cross-modal features between mentions and entities.

3.4.1 Hierarchical Encoding Network

The inputs of IMN include two parts: mention contexts and the candidate entity prototypes. Each part of inputs contains vector sequences of multiple modalities. The Hierarchical Encoding Network component is utilized to capture the inherent relationship between mention contexts and entity prototypes at different unified auxiliary space on multiple levels.

For mention contexts, each modality is encoded in a parallel way $m' = \{m'_1, m'_2, ..., m'_s\}$. We utilize a Dilated CNN (Yu and Koltun, 2015) layer to extract local features as for the first level embedding, we utilize a simple attention layer to aggregate the original embedding.

$$g = DilatedCNN(m') = \{g_1, g_2, ..., g_s\} \tag{1}$$

$$s_m^1 = \sum_i \alpha_i^1 g_i, \quad \alpha_i = \frac{exp(W_s^1 g_i)}{\sum_{j=1}^{s} exp(W_s^1 g_j)} \tag{2}$$

where W_s^1 is learnable weight matrix, s_m^1 denotes the encoding output at the first level. In order to represent mention and entity in a unified auxiliary space, s_m^1 is further mapped to a unified space embedding aux_m^1 by a full connected layer and a batch normalization layer after concatenation.

$$aux_m^1 = BN(W_a^1 s_m^1 + b_a^1), \tag{3}$$

where W_a^1 and b_a^1 are weight matrix and bias of the full connected layer.

In order to get higher level representation, we utilize a transformer layer (Vaswani et al., 2017) in which multimodal embeddings can be fully interacted through multiple attention architecture.

$$t = Transformer(g) = \{t_1, t_2, ..., t_s\} \tag{4}$$

The embedding of the second level space is aggregated by mean pooling, we utilize the similar full connected and batch normalization layers to generate auxiliary space embedding:

$$aux_m^2 = BN(W_a^2\, MeanPooling(t) + b_a^2), \tag{5}$$

The last layer is max pooling CNN(Conv1D). We utilize four Conv1D blocks with kernel size k = 2, 3, 4, 5. Finally, we concatenate the features generated from different CNN blocks as the output of last level encoder to obtain the embedding in the final unified auxiliary space.

$$c^k = MaxPooling(ReLU(Conv1D_k(t))), \tag{6}$$

$$c = [c^2; c^3, c^4, c^5], \tag{7}$$

$$aux_m^3 = BN(W_a^3\, c + b_a^3), \tag{8}$$

Similarly, we can get the entity representations $[aux_e^1, aux_e^2, aux_e^3]$ in the unified auxiliary space.

3.4.2 Co-attention Network

The Co-attention component implements a bidirectional interaction which can deal with the effect of different modalities from mention contexts to the knowledge graph and vice versa. We denote the two directions of effect as *entity-to-mention* and *mention-to-entity*, respectively.

The mention-to-entity attention is employed to compute the attention weights of entity embeddings in multiple unified auxiliary space with respect to mention embeddings. We employ the attention pooling mechanism as the aggregation strategy, the representation of the mention is an attentive combination of all entity representations.

$$b_{i,j} = aux_m^i ReLu(W_c^1\, aux_m^i + W_c^2\, aux_e^j + b_c), \tag{9}$$

$$\beta_{i,j} = Softmax(b_{i,j}) = \frac{exp(b_{i,j})}{\sum_j exp(b_{i,j})} \tag{10}$$

$$u_m^i = \sum_j \beta_{i,j}\, aux_e^j \tag{11}$$

$$u_m = MeanPooling(u_m^i) \tag{12}$$

3.4.3 Training

Given a set of multimodal posts which contain mentions and their corresponding entities, the training process is to minimize the ranking loss between the positive and negative pairs. Intuitively, the model is trained to produce a higher score between the representations of multimodal mention contexts and the ground-truth entity. Then the loss function is defined as:

$$L(m, e^+, e^-) = \sum_{e^- \in E} max(\gamma + f(m, e^+) - f(m, e^-), 0) \tag{13}$$

$$
\begin{aligned}
L_{fusion} = \tau_1 L(aux_m^1, aux_e^{1,+}, aux_e^{1,-}) \\
+\tau_2 L(aux_m^2, aux_e^{2,+}, aux_e^{2,-}) \\
+\tau_3 L(aux_m^3, aux_e^{3,+}, aux_e^{3,-}) \\
+\tau_4 L(u_m, u_e^+, u_e^-)
\end{aligned}
\quad (14)
$$

where f is cosine similarity, e^+ is the ground-truth corresponding entity of mention contexts m and e^- is the incorrect entity. γ is a margin parameter, τ_1, τ_2, τ_3 and τ_4 are the weights of the triplet losses in different levels.

We implement two learning tasks: knowledge learning and task learning. We match posts in support set, which usually have no mention, with candidate users and we can get the initialization parameters of the IMN on a specific task. The support set for knowledge learning is constructed manually with a few representative examples of an entity in practice. In this paper, we simply select a part of the representative timeline tweets as the support set using PageRank Network in Section 3.3. In task learning we fine tune model parameters through the MNED task.

In the inference stage, we only use the fusion space embedding $f(u_m, u_e)$ to calculate the similarity between mentions and entities without using any auxiliary spaces.

4 Experiments

4.1 Datasets

Measurement	Tweets-MEL	Weibo-MEL
# multimodal input posts	85K	25.6K
# distinct mentions in posts	1678	509
# entities in the knowledge graph	68K	501
# timeline tweets in the knowledge graph	2M	61.2k
avg.# length of posts	20.59	193.84
avg.# mentions in a post	1.15	1.23
avg.# candidate entities for each mention	17.24	500
avg.# timeline tweets of an entity	121	122.36

Table 1: Key statistics of the MNED dataset.

We conduct comparative experiments on two public multimodal entity disambiguation dataset (Adjali et al., 2020a; Zhou et al., 2021). Tweets-MEL collects text and images to jointly build a corpus of tweets with ambiguous mentions along with a Twitter KB defining the entities. The entities in the corpus are composed of popular twitter users including people, companies, and organizations. Weibo-MEL is a MNED corpus based on the social media Weibo, and including five construction stages: multimodal information extraction, mention extraction, entity extraction, triple construction and dataset construction. The overall statistics can be seen in Table 1.

4.2 Experimental Settings

Hyperparameters For the pre-training model, We use the default parameters of FationBert (Gao et al., 2020) and feature extractor parameters adopt the default configuration of original feature extraction model. For IMN, τ_1, τ_2, τ_3 and τ_4 is 0.5,0.5,0.5 and 1.0, the margin of the loss function is 0.2 and the epoch is 100 with a validation set for early stopping. We update the parameters using Adam optimizer (Kingma and Ba, 2014) with an initial learning rate of 0.001, the dropout rate is 0.2.

Evaluation Metrics For evaluation, we use standard micro P@1 accuracy(Adjali et al., 2020b; Moon et al., 2018) and R@3 (Moon et al., 2018) recall as metrics in our experiments. P@1 can intuitively reflect the precision of results. R@3 evaluates the matching quality by measuring whether the ground-truth entity is highly ranked.

4.3 Results and Analysis

4.3.1 Baselines

We compare our IMN model with both machine learning methods and multimodal deep learning methods. These benchmark methods are introduced as follows:

- **ET** (Adjali et al., 2020b): A feature-based machine learning model use the combination of multimodal features to build an Extra-Trees classifier for MNED task. **JMEL**: extracts the features of different modalities and learn a joint representation of tweets with a fully connected neural network.

- **ARNN** (Eshel et al., 2017): A text-only method for short noisy text, which uses an Attention RNN model to compute similarity between words and entity embeddings to disambiguate among candidates. **BERT** replaces the GRU layers with BERT

- **DZMNED** (Moon et al., 2018): The first proposed method for MNED by considering multimodal contexts, which adopts a CNN-LSTM hybrid network with modality attention. **DZMNED(BERT)**: replaces the Glove pre-training model with BERT.

4.3.2 Main Results

Table 2 shows the results of our model compared with baselines. In general, our IMN model achieves promising improvements over all the baselines on both P@1 and R@3 with the multimodal datasets. It can be observed that the pre-training methods are at an absolute advantage in both P@1 and R@3, which shows advantage of transfer learning and the necessity of jointly representing multimodal features for MNED task. Comparing to the multimodal method such as JMEL with traditional textual and visual representation methods, our model achieves 2.2% absolute improvement on P@1. The improvements indicate that the interaction between multiple modalities also adds performance gain by capturing the effect of different modalities from both the posts and the knowledge graph. In addition, adding more multimodal features can still supplement MNED tasks, even that the pre-trained representation already contain multimodal information. This affirms the advantage of IMN to capture information of different modal.

Model	Tweets-MEL		Weibo-MEL	
	P@1(%)	R@3(%)	P@1(%)	R@10(%)
ET	67.1	-	-	-
JMEL	80.3	-	-	-
ARNN	80.4	93.2	41.3	53.4
BERT	81.1	93.3	42.4	54.9
DZMNED	80.1	94.2	40.6	54.3
DZMNED(BERT)	82.0	94.4	46.3	55.5
IMN	**84.2**	**95.2**	**47.8**	**56.5**

Table 2: Comparison results with baselines on the multimodal dataset. The best performance is denoted with bold text. To be consistent with previous works, we use R@3 for Tweets-MET and R@10 for Weibo-Mel respectively.

4.3.3 Meta Learning Analysis

To investigate the effectiveness of our model on reducing training data by introducing multimodal knowledge, we randomly selected part of training data and compare our model with Zero-Shot model DZMNED. In Figure 3, our IMN method achieves the best overall performance, especially our method is significantly effective dealing with insufficient training data. This validates the advantage of involving multimodal information at the knowledge level.

4.3.4 Aggregating Statistics

In order to further study the dependency on annotated knowledge of IMN and the effect of different methods for entity support set construction, we conduct comparative experiments using different K values and two aggregation strategies and the results are shown in Figure 4. We can observe that the effect of PageRank performs better than random method especially for a small number of K values. It indicates that the features selected by the PageRank method are more representative and the influence of noise on the result is reduced to some extent. On the other hand, the best result is obtained when $k = 10$, the point can be inferred that great results can be achieved by maintaining only a small amount of high-quality data at the knowledge level. In this way we can reduce the dependency on annotated knowledge.

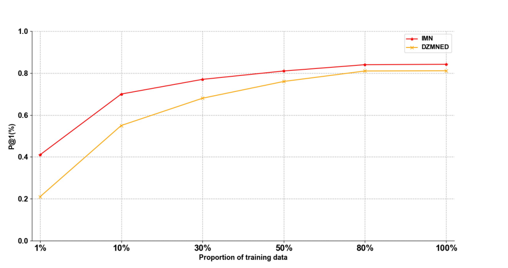

Figure 3: Performance comparison on part of training data.

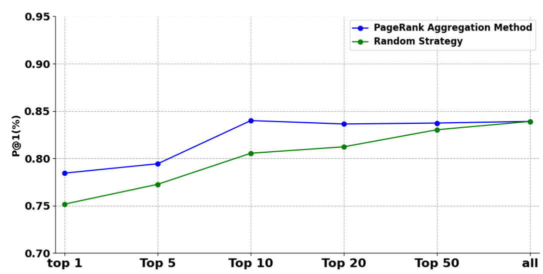

Figure 4: Results corresponding to different aggregation strategies. The abscissa represents the final aggregated number of entity historical data and the ordinate represents the corresponding precision.

4.3.5 Multimodalitiy Analysis

In this part, we perform a series of experiments to evaluate the performance of our model on dealing with the multimodal features on different input sides. As shown in Table 3, the pre-trained features outperform other single-modal features. Besides, we enrich multimodal features on the mention side and the entity side respectively. Results show that adding multimodal features from both sides can improve the model effect, and the multimodal features on the entity side has a more obvious contribution

Modal Side	Mention Modals			Entity Modals			Results	
	text	image	joint	text	image	joint	P@1(%)	R@3(%)
Single Modal	✓			✓			81.08	93.03
		✓			✓		77.56	90.93
			✓			✓	**82.19**	93.85
Mention Side	✓	✓		✓			82.38	94.16
	✓		✓	✓			82.38	94.14
	✓	✓	✓	✓			**82.70**	95.00
Entity Side	✓			✓	✓		83.10	95.06
	✓			✓		✓	83.19	95.11
	✓			✓	✓	✓	**83.21**	**95.11**

Table 3: Results of the Multimodalitiy Analysis. Single Modal indicates the effect of different modals when used alone. Mention Side and Entity Side refer to the enrichment means of multimodal information on the mention and the knowledge side respectively.

to the improvement of results. This points out a new direction for data annotating of MNED tasks: we can put the focus of data annotation on the production of multimodal knowledge, even if the input mention does not have multimodal contexts. In this way, the multimodal annotation dependence on the mention side can be greatly reduced.

4.3.6 Ablation Study

Model	Results	
	P@1(%)	R@3
IMN	**84.2**	**95.2**
- Knowledge Guided	83.5	95.1
- aux_1	83.9	95.1
- aux_1, aux_2	83.5	94.8
- aux_1, aux_2, aux_3	82.7	94.0

Table 4: Ablation tests for MNED. "-" means removing corresponding component of the model.

To investigate the effect of each component in our model, we conduct a set of ablation experiments as shown in Table 4. *IMN* is the complete proposed model. The notation '-' means removing some part of the model. From the experimental results we can observe that the performance drops obviously when auxiliary spaces are removed, which demonstrates the effectiveness of our interactive model. This proves the multimodal information from both the posts and the entities is helpful for the MNED task.

We also investigated the necessity of knowledge guidance in the pre-training process. Firstly, We implement the same mask strategy of Bert by treating mentions as normal words. Then, negative examples of each case are randomly selected from all tweets. We can observe that the overall accuracy will be reduced to a certain extent in Table 3. The result shows that the structure and historical information in the knowledge graph can be learned by a pre-training manner and is helpful to improve the effect of the MNED task.

5 Conclusion

We propose to solve MNED task at the knowledge level through multimodal Transfer Learning and Meta Learning. With large-scale unsupervised data and a small amount of annotated knowledge, our model significantly outperforms the state-of-the-art MNED methods. Experimental results show that enrich multimodal features at the knowledge level is more conducive to improving the effect of MNED models compared with mention contexts annotation.

There are still many points worth continuing to explore. In particular, the structural information in the knowledge graph which can be learned by knowledge representation models such as transE may also be useful. Besides, the prototype aggregation method still needs further exploration with graph learning models such as GCN etc.

6 Limitations

Our method requires additional multimodal knowledge and a large amount of unsupervised data for pre-training, which is additional burden to collect in practice. Besides, the performance of our model also depends on the feature extractors, how to combine more feature extractors and utilize more unified auxiliary space is still worth continuing exploration. Finally, our method does not consider the situation of multiple images in one post and entities lacking of multimodal knowledge .

References

Ágnes Abuczki and Esfandiari Baiat Ghazaleh. 2013. An overview of multimodal corpora, annotation tools and schemes. *Argumentum*, 9:86–98.

Omar Adjali, romaric Besancon, olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020a. Building a multimodal entity linking dataset from tweets. In *International Conference on Language Resources and Evaluation (LREC)*, Marseille, France.

Omar Adjali, romaric Besancon, olivier Ferret, Hervé Le Borgne, and Brigitte Grau. 2020b. Multimodal entity linking for tweets. In *European Conference on Information Retrieval (ECIR)*, Lisbon, Portugal.

Ander Barrena, Aitor Soroa, and Eneko Agirre. 2018. Learning text representations for 500k classification tasks on named entity disambiguation. In *CoNLL*, pages 171–180.

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *EACL*.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pages 213–229, Cham. Springer International Publishing.

Anthony Chen, Pallavi Gudipati, Shayne Longpre, Xiao Ling, and Sameer Singh. 2021. Evaluating entity disambiguation and the role of popularity in retrieval-based NLP. In *ACL/IJCNLP (1)*, pages 4472–4485. Association for Computational Linguistics.

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

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multilingual image description with neural sequence models. *arXiv preprint arXiv:1510.04709*.

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

Zheng Fang, Yanan Cao, Qian Li, Dongjie Zhang, Zhenyu Zhang, and Yanbing Liu. 2019. Joint entity linking with deep reinforcement learning. In *The World Wide Web Conference*, pages 438–447.

Octavian-Eugen Ganea and Thomas Hofmann. 2017. Deep joint entity disambiguation with local neural attention. In *EMNLP*, pages 2619–2629.

Dehong Gao, Linbo Jin, Ben Chen, Minghui Qiu, Peng Li, Yi Wei, Yi Hu, and Hao Wang. 2020. Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 2251–2260. Association for Computing Machinery.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhengyan He, Shujie Liu, Mu Li, Ming Zhou, Longkai Zhang, and Houfeng Wang. 2013. Learning entity representation for entity disambiguation. In *ACL*, pages 30–34.

Saravanakumar Kandasamy and Aswani Kumar Cherukuri. 2020. Query expansion using named entity disambiguation for a question-answering system. *Concurr. Comput. Pract. Exp.*, 32(4).

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *NAACL-HLT*, pages 260–270.

Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. In *ACL*, pages 1595–1604.

Hongying Liu, Ruyi Luo, Fanhua Shang, Mantang Niu, and Yuanyuan Liu. 2021. *Progressive Semantic Matching for Video-Text Retrieval*, page 5083–5091. Association for Computing Machinery, New York, NY, USA.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019a. Vilbert: Pretraining task-agnostic visio linguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019b. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rada Mihalcea and Andras Csomai. 2007. Wikify!: Linking documents to encyclopedic knowledge. In *CIKM*, pages 233–242.

Ishan Misra, C Lawrence Zitnick, and Martial Hebert. 2016. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer.

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.

Hatem Mousselly-Sergieh, Teresa Botschen, Iryna Gurevych, and Stefan Roth. 2018. A multimodal translation-based approach for knowledge graph representation learning. In *Joint Conference on Lexical and Computational Semantics,SEM@NAACL-HLT*, pages 225–234.

Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2017. J-REED: joint relation extraction and entity disambiguation. In *CIKM*, pages 2227–2230. ACM.

Thien Huu Nguyen, Nicolas Fauceglia, Mariano Rodriguez Muro, Oktie Hassanzadeh, Alfio Massimiliano Gliozzo, and Mohammad Sadoghi. 2016. Joint learning of local and global features for entity linking via neural networks. In *COLING*, pages 2310–2320.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Pouya Pezeshkpour, Liyan Chen, and Sameer Singh. 2018. Embedding multimodal relational data for knowledge base completion. In *EMNLP*, pages 3208–3218.

Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. 2020. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *WMT*, pages 543–553.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ruobing Xie, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2017a. Image-embodied knowledge representation learning. In *IJCAI*, pages 3140–3146.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017b. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 3208–3216.

Fisher Yu and Vladlen Koltun. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*.

Xingchen Zhou, Peng Wang, Guozheng Li, Jiafeng Xie, and Jiangheng Wu. 2021. Weibo-mel, wikidata-mel and richpedia-mel: Multimodal entity linking benchmark datasets. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers New Infrastructure Construction*, pages 315–320, Singapore. Springer Singapore.