MMEL: A Joint Learning Framework for Multi-Mention Entity Linking

Chengmei Yang[1, 2] Bowei He[2] Yimeng Wu[3] Chao Xing[3] Lianghua He[†1] Chen Ma[†2]

[1]Department of Computer Science and Technology , Tongji University , China
[2]Department of Computer Science, City University of Hong Kong, Hong Kong SAR
[3]Huawei Noah's Ark Lab

Abstract

Entity linking, bridging mentions in the contexts with their corresponding entities in the knowledge bases, has attracted wide attention due to many potential applications. Recently, plenty of multimodal entity linking approaches have been proposed to take full advantage of the visual information rather than solely the textual modality. Although feasible, these methods mainly focus on the single-mention scenarios and neglect the scenarios where multiple mentions exist simultaneously in the same context, which limits the performance. In fact, such multi-mention scenarios are pretty common in public datasets and real-world applications. To solve this challenge, we first propose a joint feature extraction module to learn the representations of context and entity candidates, from both the visual and textual perspectives. Then, we design a pairwise training scheme (for training) and a multi-mention collaborative ranking method (for testing) to model the potential connections between different mentions. We evaluate our method on a public dataset and a self-constructed dataset, NYTimes-MEL, under both text-only and multimodal scenarios. The experimental results demonstrate that our method can largely outperform the state-of-the-art methods, especially in multi-mention scenarios. Our dataset and source code are publicly available at https://github.com/ycm094/MMEL-main.

1 INTRODUCTION

Traditional entity linking aims at linking the mentions from the context to the corresponding entities in the knowledge graphs (KGs) [32, 19]. The main purposes of entity linking lie in bridging the web data with knowledge bases and then facilitating downstream information-retrieval applications, such as knowledge-based question answering [17, 41] and semantic search [8, 18].

Early approaches of entity linking mainly focus on addressing the entity ambiguity [19, 4]. For example, given the mention "Chaplin" in the context "A teenage Chaplin in the play Sherlock Holmes, in which he appeared between 1903 and 1906", we need to figure out that this mention should link to actor "Charlie Chaplin" rather than the composer "Christopher Chaplin". Here, "Charlie Chaplin" and "Christopher Chaplin" belong to a set of entity candidates. Although feasible, these early methods only leverage the textual information to guide the entity link task. Recently, a few works have explored the effectiveness of fusing the multimodal information, containing both visual and textual modalities, to improve the matching performance [15, 37, 39]. Also taking "Charlie Chaplin" as an example, with the corresponding images of the context and entity candidates, the model can make use of the visual information and then facilitate the final entity linking performance.

Although there have been studies in this field, we argue that there are still three potential avenues to improve. Firstly, these approaches **only take one mention into account each time, ignoring the potential connection between different mentions in the same context**. As the left part shown in Fig. 1, given a sentence "Francis X. Bushman, Chaplin and Anderson, photo taken at the Essanay Studio, Chicago in 1915", previous methods would like to link the mentions "Francis X. Bushman", "Chaplin", and "Anderson" to the corresponding KG entities one by one [37, 39]. Even though doable, these approaches fail to capture the potential connection among the mentions, such as the same era or a similar occupation. It is natural that there must be certain relationships between people when they appear in the same text. Therefore, considering the three mentions simultaneously in the above example can help find a few common characteristics and then facilitate the entity linking for all the mentions. Secondly, **considering contexts**

[†]Lianghua He and Chen Ma are both corresponding authors.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

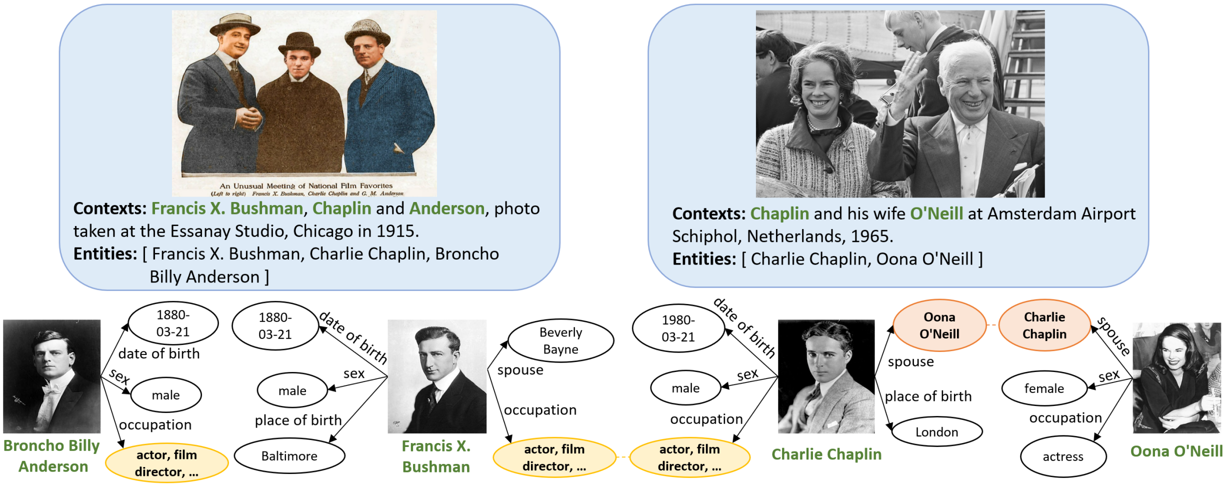

An Unusual Meeting of National Film Favorites
(Left to right) Francis X. Bushman, Charlie Chaplin and G. M. Anderson
Contexts: Francis X. Bushman, Chaplin and Anderson, photo taken at the Essanay Studio, Chicago in 1915.
Entities: [Francis X. Bushman, Charlie Chaplin, Broncho Billy Anderson]
Contexts: Chaplin and his wife O'Neill at Amsterdam Airport Schiphol, Netherlands, 1965.
Entities: [Charlie Chaplin, Oona O'Neill]
1880-03-21
date of birth
sex
male
occupation
actor, film director, ...
Broncho Billy Anderson
1880-03-21
date of birth
sex
male
place of birth
Baltimore
Francis X. Bushman
Beverly Bayne
spouse
occupation
actor, film director, ...
1980-03-21
date of birth
sex
male
occupation
actor, film director, ...
Charlie Chaplin
Oona O'Neill
spouse
place of birth
London
Charlie Chaplin
spouse
sex
female
occupation
actress
Oona O'Neill

Figure 1: The illustration of the multi-mention entity linking task. The top two figures show the contexts and entity candidates for two samples, while the bottom four graphs present the background knowledge for some entity candidates. The green font color stands for the mentions required to be linked in the contexts.

and entity candidates independently fails to capture the fine-grained features. Previous methods prefer to obtain the representation of contexts and entity candidates separately in advance and then design a customized module to learn the connection between them [37, 39]. By doing so, the obtained features of entity candidates are the same for different contexts and the context features are also fixed when linking to different entity candidates. In Fig. 1, two samples containing the mention "Chaplin" are given to show the significance of learning mention and entity candidate representations jointly. When learning the entity feature of "Charlie Chaplin", its representation in the left case should pay more attention to the "occupation" relation in the knowledge graph, while that in the right case is supposed to focus more on the "spouse" relation. Meanwhile, when linking to different entity candidates, fixing the context features fails to capture the uniqueness of each entity, leading to the worse performance. Take the left sample in Fig. 1 as an example, the context features are supposed to be obtained based on the corresponding entities, that is, the context features are different when given the two entity candidates "Charlie Chaplin" (positive one) and "Christopher Chaplin" (negative one). Therefore, it is crucial to learn the context-relevant entity features and entity-relevant context features. Thirdly, **not only the textual information, but also the visual modality should be counted in the multi-mention scenario.** Actually, named entities with multimodal contexts such as texts and images are ubiquitous in our daily life. Some works have explored the advantages of integrating both visual and textual information for entity-linking tasks, but they only focus on the single-mention scenario. Besides, there are some works proposed for collective entity linking [5, 45], considering the documents with multiple mentions, but these methods only consider the textual information. Different from them, we argue that, to address the multi-mention entity linking more practically, it is essential to take multimodal information into account and explore an effective joint learning framework to model the representations of contexts and entity candidates in both textual and visual modalities.

To tackle the aforementioned issues, we propose a novel **Mu**li-**M**ention **E**ntity **L**ink method, **MMEL**, consisting of a context-entity joint feature extraction module, a multimodal learning framework, and a multi-mention collaborative ranking with a pairwise training scheme. First, the joint feature extraction module is designed for taking full advantage of the contexts and entity candidates together to capture the corresponding representations more precisely. Second, the multimodal learning framework is proposed for considering both the visual and textual modalities to facilitate our multi-mention entity linking task. Then, the multi-mention collaborative ranking aims at linking multiple mentions in a context to the corresponding KG entities by taking into account the potential connection between different mentions. During training, based on contrastive learning, we design a novel pairwise training strategy to learn more distinguishable representations for mentions and entities. Even more importantly, to further contribute to the multi-mention entity linking, we construct a new open dataset: NYTimes-MEL with more than 10K multimodal samples extracted from the New York Times [34, 48] and Wikidata [36]. The experimental results on two datasets, Wiki-MEL [37] and NYTimes-MEL, show that our framework achieves consistently better performance compared with other state-of-the-art baselines in both text-only and multimodal multi-mention entity linking settings. In summary, our contributions are:

- To the best of our knowledge, we are the first to study the multi-mention entity linking task in the multimodal scenario and propose a novel framework with the joint feature extraction module to learn the representations of contexts and entity candidates together.

- To consider the potential connection between different mentions in the same context, we design a multi-mention collaborative ranking method accompanied by an effective pairwise training scheme.
- To evaluate our method on multi-mention entity linking task, we additionally construct a new dataset NYTimes-MEL. The experimental results under both text-only and multimodal settings demonstrate the superiority. The code and newly-constructed data are released publicly.

2 RELATED WORKS

Entity Linking. Entity linking, aiming at linking named entity mentions in the web text with their corresponding entities in a knowledge base, is critical for bridging web data and knowledge base [32]. Early works [4, 19, 35, 21, 31] mainly focus on textual entity link, in which both mentions and entities only possess textual modality information for processing. Apart from these methods, there are also works exploring the collective entity linking problem [5, 45, 49, 28]. For example, RRWEL [45] involves the recurrent random walk network to bridge the connection between different mentions, while NCEL [5] leverages graph neural network to capture the relationship among candidate entities of different mentions. However, these approaches are only based on the textual modality and fail to take full advantage of the multimodal information.

Due to the wide application of cameras and other photographic equipment, visual modality information such as images becomes easily accessible for these named entities. More and more recent works are proposed to utilize the visual modality information [20, 33] or multimodal information [25, 1, 2, 47, 15, 39, 37, 12] to improve the entity linking performance. Though previous works have considered different modality settings, few of them paid attention to the multi-mention entity link scenario which actually widely exists in the entity link task. In this work, we focus on this challenging scenario and take both the text-only and multimodal settings into account in our experiments.

Multimodal Learning. The entities in the real world usually have multimodal information, especially in textual and visual modalities. Many methods have been proposed to learn multimodal representations of such entities to better understand them. Mainstream approaches can be classified into two categories: joint representation [27, 24, 10] and collaborative representation [14, 40]. The difference between them lies that joint learning methods embed multimodal data into a joint representation space where each latent feature contains multimodal information, while collaborative representation methods learn the single-modal representation separately. Thanks to the huge amount of training data and sophisticated model structures, large multimodal pre-trained models [30, 26, 22] have been more and more utilized as the multimodal feature extractors. Besides, after obtaining the extracted latent features of each modality, how to fuse such information effectively to benefit the downstream tasks has also attracted wide attention [3, 42, 7]. In this work, we not only utilize multimodal pre-trained models to extract representations from raw data of textual and visual modalities, but also adopt different fusion strategies to fuse the information from both modalities.

3 METHODOLOGY

In this section, we first introduce the problem definition of multi-mention entity linking. Then, we elaborate our framework, in terms of the context-entity joint feature extraction for each modality of text and vision, the multimodal learning framework, the pairwise training scheme, and the multi-mention collaborative ranking, respectively. We summarize our framework—MMEL in Fig. 2.

3.1 PROBLEM DEFINITION

Given N samples, consisting of the input context set $X = \{x_i\}_{i=1}^{N}$ and the corresponding entity set $Y = \{y_i\}_{i=1}^{M}$, entity linking is defined as mapping a mention with its context to the correct entity in the KG. Here, each input context x with L tokens contains a mention x_m with L_m tokens, that is, $x_m \subset x$ and $L_m \leq L$. Each entity y is constructed by a subgraph, containing the entity itself, relevant relations in the relation set $\mathcal{R}$ and corresponding tail entities. It is worth noting that a context may contain multiple mentions. Therefore, different from the previous task definition that regards different mentions with the same context as different samples, we introduce a multi-mention entity linking task taking different mentions with the same context as one sample, that is, $\{x_m^i\}_{i=1}^{n} \subset x$ and $n \geq 1$.

It is worth noting that we also take multimodal information into account rather than only textual modality, that is, the input text $x = \{x_v, x_t\}$ and entity $y = \{y_v, y_t\}$. Here, the subscripts v and t represent visual and textual information, respectively.

3.2 CONTEXT-ENTITY JOINT FEATURE EXTRACTION

3.2.1 Textual Feature Extraction

Textual features are extracted from the pretrained language model BERT [11] and fine-tuned during the training stage. Previous methods always input the sequence of contexts and entity candidates independently, and then measure the relationship between them from the high-level features. In this way, the representations of contexts are all the same when linking the mention to different entity candidates and vice versa. In our multi-mention entity linking task, one context may contain multiple mentions to be matched. Therefore, we argue that the model should learn the context and entity representations jointly.

In order to consider the context and entity candidates jointly, we combine the context and entity together to learn the low-level joint features. Since the KG entities are stored in the form of a graph, we transform the graph into a sequence. For example, given the entity candidate $Q946745$, the corresponding sequence is "Sex: male. Date of birth: 1880-03-21. Place of birth: Little Rock. Occupation: actor, film producer, ...". In this way, the KG entity sequence can be treated in the same way as context. Especially, as shown in Fig. 2, inspired by the success of prompt learning [43, 9], we design a template "Context: x. Entity: y." to integrate the sequence of context and entities and then put the combined sequences into the BERT model to obtain the corresponding representations $f_t = \{T_i | T_i \in \mathbb{R}^{d_t}, i = 1, 2, ..., t_c + t_e\}$, where d_t represents the textual dimension, t_c and t_e refer to the token length of context and entities, respectively.

Since the pre-trained model BERT involves multiple self-attention layers, the obtained joint features f_t consider the content of contexts and entities simultaneously. To get the corresponding representations of contexts and entity candidates, we leverage two masks to separate the joint features and all the masks contain only 0 or 1 value. The first one is the context mask $\mathcal{M}_t^c = [|1|_{i=1}^{t_c}, |0|_{i=t_c+1}^{t_c+t_e}] \in \mathbb{R}^{t_c+t_e}$, while the second one is the entity mask $\mathcal{M}_t^e = [|0|_{i=1}^{t_c}, |1|_{i=t_c+1}^{t_c+t_e}] \in \mathbb{R}^{t_c+t_e}$. Using the above two masks, we multiply them with the joint features to get the context features f_t^c and entity features f_t^e as follows:

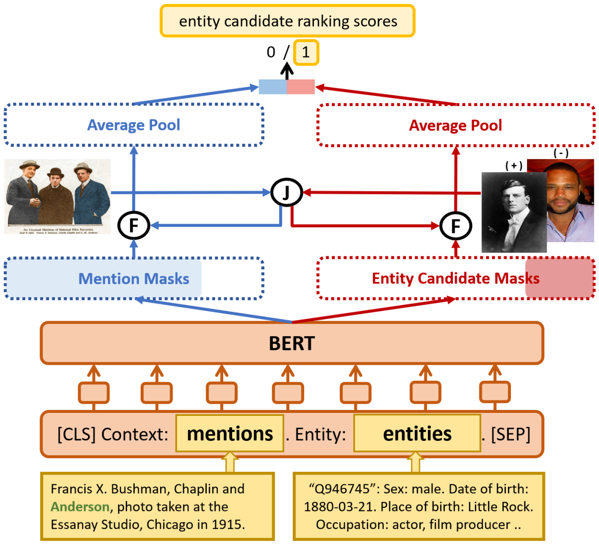
$$\begin{aligned} f_t^c &= \mathcal{M}_t^c \cdot f_t \in \mathbb{R}^{(t_c+t_e) \times d_t} \,, \\ f_t^e &= \mathcal{M}_t^e \cdot f_t \in \mathbb{R}^{(t_c+t_e) \times d_t} \,. \end{aligned} \tag{1}$$

3.2.2 Visual Feature Extraction

Visual features are extracted through the visual encoder of pretrained model CLIP [30]. Different from the joint textual feature extraction that learns the low-level representations, we first leverage the CLIP to extract the characteristics of the context and entity images, $f_v^c \in \mathbb{R}_v^d$ and $f_v^e \in \mathbb{R}_v^d$, respectively. Then, we adopt a single-layer perception to map the image features to the high-level space as follows:

$$\begin{aligned} \overline{f_v^c} &= \text{Reshape}(\text{ReLU}(f_v^c W_v + b_v)) \,, \\ \overline{f_v^e} &= \text{Reshape}(\text{ReLU}(f_v^e W_v + b_v)) \,, \end{aligned} \tag{2}$$

where $W_v \in \mathbb{R}^{d_v \times k d_v}$ and $b_v \in \mathbb{R}^{k d_v}$ are trainable parameters. After obtaining the visual features of contexts and entities, we reshape them to the new size, $\overline{f_v^c} \in \mathbb{R}^{k \times d_v}$ and $\overline{f_v^e} \in \mathbb{R}^{k \times d_v}$. Then we concatenate them together to get a new feature $f_v \in \mathbb{R}^{2k \times d_v}$ and adopt q modules, each of which consists of two multi-head self-attention and a feed-forward layer, to learn the context and entity image features jointly. Afterward, the joint image feature $\tilde{f_v}$ is obtained.



Francis X. Bushman, Chaplin and Anderson, photo taken at the Essanay Studio, Chicago in 1915.

"Q946745": Sex: male. Date of birth: 1880-03-21. Place of birth: Little Rock. Occupation: actor, film producer ..

Figure 2: The illustration of our MMEL framework. "F" and "J" denote multimodal fusion module and context-entity joint feature extraction for images, respectively. (+) and (-) denote the positive and negative entities.

Similar to the textual feature extraction, we then leverage two independent image masks $\mathcal{M}_v^c = [|1|_{i=1}^{k}, |0|_{i=k+1}^{2k}] \in \mathbb{R}^{2k}$ and $\mathcal{M}_v^e = [|0|_{i=1}^{k}, |1|_{i=k+1}^{2k}] \in \mathbb{R}^{2k}$ to get the corresponding context and entity image features as follows:

$$\begin{aligned} \tilde{f_v^c} &= (\mathcal{M}_v^c \cdot \tilde{f_v})[: k] \in \mathbb{R}^{k \times d_v} \,, \\ \tilde{f_v^e} &= (\mathcal{M}_v^e \cdot \tilde{f_v})[k :] \in \mathbb{R}^{k \times d_v} \,. \end{aligned} \tag{3}$$

3.3 MULTIMODAL LEARNING FRAMEWORK

Based on the above textual and visual features, in this section, we focus on merging the multimodal information for the final entity linking task. Specifically, for textual representations, we first leverage a single-layer perception to make the textual dimension and visual dimension the same:

$$\begin{aligned} \tilde{f_t^c} &= \text{ReLU}(f_t^c W_t + b_t)) \,, \\ \tilde{f_t^e} &= \text{ReLU}(f_t^e W_t + b_t)) \,, \end{aligned} \tag{4}$$

where $W_t \in \mathrm{R}^{d_t \times d_v}$ and $b_t \in \mathrm{R}^{d_v}$ are both trainable parameters. Then, following the previous work [37], we adopt a hierarchical multimodal co-attention module (MCM) to capture the correlations between the two modalities. This module first involves an attention layer, which makes visual features conduct self-attention and then put visual features as the query (Q), textual features as key (K) and value (V) to learn the cross-modal attention as follows:

$$A(Q,K,V) = \mathrm{softmax}(\frac{QW_q \cdot (KW_k)^\top}{\sqrt{d_k}})VW_v\,, \tag{5}$$

where $W_q \in \mathrm{R}^{d_v \times d_k}$, $W_k \in \mathrm{R}^{d_t \times d_k}$, and $W_v \in \mathrm{R}^{d_t \times d_k}$. Then, the attention layer is followed by a feed-forward layer FFN, residual connection, and layer normalization to better the multimodal learning performance as follows:

$$\begin{aligned} \tilde{f} &= \mathrm{LN}(\mathrm{A}(Q,K,V))\,, \\ \tilde{f} &= \mathrm{A}(Q,K,V) + \mathrm{FFN}(\tilde{f})\,, \end{aligned} \tag{6}$$

where LN denotes layer normalization, FFN contains two fully-connected (FC) layers and a ReLU activation function between them (i.e., FC-ReLU-FC). Based on the above MCM module, we can obtain the multimodal contexts features as follows:

$$\begin{aligned} f^c_{v,att} &= \max(\mathrm{MCM}(\tilde{f}^c_t, \tilde{f}^c_v, \tilde{f}^c_v)) \in \mathbb{R}^{d_v}\,, \\ f^c_{t,att} &= \max(\mathrm{MCM}(f^c_{v,att}, \tilde{f}^c_t, \tilde{f}^c_t) \in \mathbb{R}^{d_v}\,, \end{aligned} \tag{7}$$

where $\max$ denotes a max-pooling operation to capture the representative features. Similarly, the multimodal features of entities $f^e_{v,att}$ and $f^e_{t,att}$ can also be obtained through the above equation.

Later, we concatenate the joint visual and textual features of the context and also leverage a gated mechanism to fuse the multimodal representations:

$$g = \mathrm{softmax}(\mathrm{FFN}(\mathrm{concat}[f^c_{t,att}; f^c_{v,att}]))\,, \tag{8}$$

where $g \in \mathbb{R}^2$ refers to the importance of the textual features. Finally, the fused representation of contexts h^c is obtained through:

$$h^c = g \cdot [f^c_{t,att}; f^c_{v,att}]^\top \in \mathbb{R}^{d_v}\,, \tag{9}$$

and the fused representation of entity candidates $h^e \in \mathbb{R}^{d_v}$ can be obtained in the same way. Besides, we can also adopt other multimodal fusion modules, such as M-Encoder [7], to fuse the textual and visual information.

After obtaining the fused multimodal representations h^c and h^e through the hierarchical co-attention module and gated mechanism, we need to figure out whether the entity candidate matches the mention in the context. Note that there are multiple entity candidates when the model links the mention to the background KG entities. Besides, only one correct entity is regarded as the positive entity, and the other entity candidates are negative entities. Different from the above method that employs the contrastive learning to measure the distance between context and positive/negative entities [37], we leverage a single perception to discriminate the correlation between entity-relevant contexts and context-relevant entities since our framework learns the different context features for both positive and negative entities. Therefore, the matching score can be obtained via

$$m(x, y_i) = \mathrm{softmax}(\mathrm{concat}[h^c; h^e]W_m + b_m) \tag{10}$$

where $W_m \in \mathrm{R}^{2d_v \times 2}$, $b_m \in \mathrm{R}^2$ and $m(x, y_i) \in \mathrm{R}^2$. When h^e denotes the representations of positive entities, the ground truth of $m(x, y_i)$ is set to 1, while 0 for negative entities. In this way, the model can learn more fine-grained correlations between contexts and entity candidates to benefit our multi-mention entity linking task. *The multimodal learning framework is illustrated in Fig. 2.*

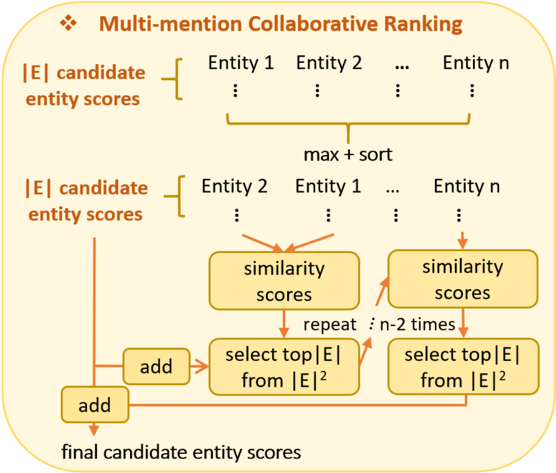


Figure 3: The detailed process of the multi-mention collaborative ranking.

3.4 PAIRWISE TRAINING SCHEME

In our multi-mention entity linking task, there is more than one mention in a given context. Different from the traditional single-mention entity linking, we pay more attention to the potential connection between different mentions rather than simply considering the correlation between mentions and entity candidates. As shown in the right sample in Fig. 1, given a context containing two mentions, we expect to learn the correlations (such as the spouse relationship and similar dates of birth) between them via the contrastive learning.

To be specific, for each mention in this context, there exist one positive entity and one negative entity extracted from the entity candidate set. Here, we leverage e^{pos}_1 and e^{neg}_1 to denote the positive and negative entities for the first mention, while e^{pos}_2 and e^{neg}_2 are the positive and negative entities for the second mention. It is worth noting that $e^{pos}_1 \neq e^{neg}_1 \neq e^{pos}_2 \neq e^{neg}_2$. Then we use the triplet loss in training to improve the fine-grained similarities between two positive entities and reduce the proximity between positive

Table 1: The statistics of datasets. The “single” and “multi” denote the number of samples with one or multiple mentions.

Datasets	Samples	Mentions	Text length	Mentions	train		dev		test	
					single	multi	single	multi	single	multi
Wiki-MEL	22,280	26,280	8.3	1.2	13,413	2,198	1,916	316	3,775	662
NYTimes-MEL	11,340	14,689	18.5	1.3	6,240	1,809	854	271	1,559	607

and negative entities as follows:

$$\mathcal{L}_{p1} = \max\{\mathrm{S}(h^{e_1^{pos}}, h^{e_2^{neg}}) - \mathrm{S}(h^{e_1^{pos}}, h^{e_2^{pos}}) + \gamma, 0\}\,,$$
$$\mathcal{L}_{p2} = \max\{\mathrm{S}(h^{e_2^{pos}}, h^{e_1^{neg}}) - \mathrm{S}(h^{e_2^{pos}}, h^{e_1^{pos}}) + \gamma, 0\}\,, \tag{11}$$

where $\mathrm{S}(a, b) = \sigma(\mathrm{FFN}(a, b)) \in \mathbb{R}$ denotes the similarity score calculated from the representations of entity a and b. γ refers to the margin value and σ is the sigmoid function. Therefore, the final pairwise loss is $\mathcal{L}_p = \mathcal{L}_{p1} + \mathcal{L}_{p2}$.

It is worth noting that there are some contexts containing more than two mentions and some have only one mention. For the contexts with multiple mentions, we only sample two mentions randomly each time to construct the training pair. For the contexts with only one mention, inspired by the latest works [16, 38], we make a copy of the context and the corresponding positive entity, and then sample another negative entity from the entity candidate set. Although the sequence of context and the positive entity is the same for two cases, through the language model BERT with plenty of dropout layers, the representations of context and entities are not the entirely same. Thus, the contrastive learning can be used as well.

3.5 MULTI-MENTION COLLABORATIVE RANKING

Although we adopt the pairwise training scheme to capture the correlation between different mentions, only two mentions are taken into account each time for both single and multiple-mention scenarios. When testing, we first figure out how many mentions are given in the context and then leverage the matching score $m(x, y_i)$ to rank the entity candidates for single-mention cases. For multi-mention cases, we design a novel ranking method based on the greedy algorithm to consider the correlation among different mentions.

Specifically, considering a context with three mentions, each mention has its own matching score. In this case, we leverage $m_{i,j} = m(x_i, y_{i,j})$ to denote the matching score between mention i and corresponding entity candidate j, where $i \in \{1, 2, 3\}$ and $j \in \{1, 2, ..., |E|\}$. Here, $|E|$ denotes the total number of entity candidates for each mention. Firstly, we select the max matching score of each mention and sort these max scores from the highest to the lowest according to different mentions. In this way, the first mention has the highest creditable entity matching score.

Different from traditional approaches that only consider the matching score for each mention separately, we also take into account the correlation between different mentions for our multi-mention entity linking task. Assuming that after sorting, the new mention order is x_2, x_1 and x_3, which means that the matching score of x_2 mention is the highest. Then, we leverage the $\mathrm{s}(x_2, x_1)$ to similarity scores between $|E|$ entity candidates of x_2 and $|E|$ entity candidates of x_1. It is worth noting that there are total $|E| \times |E|$ similarity scores $s(x_1, x_2) = \{S(x_1^a, x_2^b) = \sigma(\mathrm{FFN}(h^{e_{1,a}}, h^{e_{2,b}})); a = 1, 2, ..., |E|, b = 1, 2, ..., |E|\}$ for x_1 and x_2, where $h^{e_{1,a}}$ and $h^{e_{1,a}}$ are the representations of a-th and b-th entities for the mention x_1 and x_2, respectively. Unfortunately, these similarity scores show the exponential growth with the number of mentions. To reduce the time and space complexity, we only select the top-$|E|$ similarity scores and then average the similarity score $s(x_1^a, x_2^b)$ with the previous matching score $m_{1,a}$ and $m_{2,b}$. For features, we average the $h^{e_{1,a}}$ and $h^{e_{2,b}}$ to get the features for new $|E|$ combinations. After obtaining the top-$|E|$ new ranking scores, similarly, we calculate the correlations between the new combinations and x_3. In the end, there are total $|E|$ combinations of x_2, x_1 and x_3, representing the most possible correlations among the three mentions. We add the maximum value of similarity scores for entity candidate $x_{1,a}$ to the previous matching score $m_{1,a}$, to obtain the final ranking score. In this way, when conducting the multi-mention entity linking, we not only consider the matching score of entity candidates for each mention itself, but also the potential connection among different mentions. The process of our multi-mention collaborative ranking is illustrated in Fig. 3.

4 EXPERIMENTS

In this section, we conduct experiments on a public dataset and a self-collected dataset to demonstrate the effectiveness of our method. In the experiment, we mainly focus on the following questions:

- **RQ1:** Does our method achieve better performance than state-of-the-art methods on different datasets under both text-only and multi-modal settings?
- **RQ2:** How our method performs in both the single-mention and the multi-mention entity linking scenarios?
- **RQ3:** Whether our joint-learning framework and multi-mention collaborative ranking module with pairwise training can help improve performance? (Ablation study)

Table 2: Main results at Top-k accuracy (%) on Wiki-MEL and NYTimes-MEL dataset. The best results are shown in bold.

Modalities	Methods	Wiki-MEL				NYTimes-MEL			
		Top-1	Top-5	Top-10	Top-20	Top-1	Top-5	Top-10	Top-20
T	NCEL	2.1	10.6	21.1	41.3	8.2	11.3	18.4	31.5
T	ARNN	32.0	45.8	56.6	65.0	16.1	36.8	47.0	61.3
T	BERT	31.7	48.8	57.8	70.3	17.2	45.5	57.7	68.2
T	BLINK	30.8	44.6	56.7	66.4	14.1	35.5	47.2	58.2
T	GENRE	32.5	49.2	58.5	71.8	14.8	35.4	48.9	60.9
T	GHMFC-onlytext	34.1	51.3	60.4	72.5	16.6	38.6	50.5	62.1
T + V	JMEL	31.3	49.4	57.9	64.8	16.0	32.4	42.4	54.1
T + V	DZMNED-BERT	29.2	53.7	63.6	72.5	24.9	46.5	54.8	65.4
T + V	HieCoATT-Alter	40.5	57.6	69.6	78.6	16.7	35.2	44.6	62.0
T + V	GHMFC	43.6	64.0	74.4	85.8	17.1	40.7	51.7	64.1
T + V	LXMERT	20.6	46.9	67.3	87.6	16.4	49.8	62.8	74.7
T	MMEL-onlytext	40.2	71.2	84.2	93.6	36.8	66.9	78.6	89.9
T + V	MMEL-k1	65.0	89.1	94.4	97.2	43.4	72.8	83.1	**91.8**
T + V	MMEL-M-Encoder	67.7	90.5	95.9	**98.0**	**45.0**	**73.9**	**84.1**	91.7
T + V	MMEL	**71.5**	**91.7**	**96.3**	**98.0**	41.5	72.5	83.0	91.5

- **RQ4:** Whether our results in real cases are reasonable and persuasive? (Case Study)

4.1 EXPERIMENT SETUP

Datasets Construction. To the best of our knowledge, there is only one accessible dataset, Wiki-MEL [37], fitting for our proposed multi-mention entity linking task with multimodal information. Therefore, to further boost the research on this novel problem setting, we construct a new dataset, NYTimes-MEL, based on the images and captions collected from the New York Times [34, 48]. To find the corresponding entities, we employ the StanfordNLP tool [29] for each caption to conduct named entity recognition and regard the entities with "PERSON" type as the ground truth. For mention construction, we randomly select about 50% entities and replace them with the nick name. Then, following [37], we leverage wikidata [36] to obtain the images and 14 properties of each background KG entity. Finally, the samples containing invalid entities (cannot be extracted from the wikidata or without corresponding images) are removed and we divided the whole samples into training, validation and testing sets as 7:1:2. The statistics of two datasets are concluded in Table 1

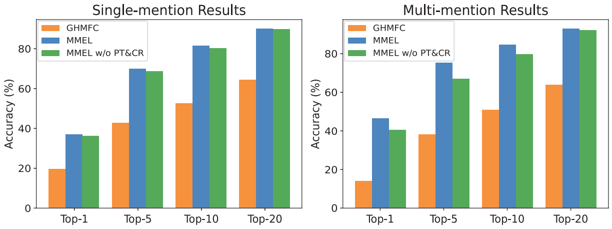


Figure 4: The results of NYTimes-MEL on single-mention and multi-mention samples. (PT: Pairwise Training scheme, CR: multi-mention Collaborative Ranking.)

Baselines. We divide the compared baseline methods into two categories, 1) text-only approaches with only textual modality, and 2) multimodal methods with both textual and visual information. In text-only setting, we compare our method with ARNN [13], BERT [11], BLINK [44], GENRE [4], and GHMFC-onlytext [37], while a collective entity linking method NCEL [6] is also involved to be compared. In the text-vision setting, we adopt JMEL [2], DZMNED-BERT [25], HieCoATT-Alter [24], GHMFC [37], and LXMERT [39] as baselines. In addition, we also design two variants of our method. The first one is **MMEL-k1**, which sets the hyper-parameter $k = 1$ to explore the impact on joint visual feature extraction. The second one is **MMEL-M-Encoder**, which replaces our MCM module with M-encoder [7] to explore the impact on different multimodal fusion strategies.

Implementation Details. In this section, we provide the implementation details of our method. Our MMEL framework is implemented with PyTorch on NVIDIA RTX A6000. We leverage the pre-trained base-uncased BERT model [11] as the textual encoder and CLIP model [30] as the visual encoder. We set the dimensions of textual and visual features, d_t and d_v, to 512 and 768. The number of stacked modules q is 2 and the new size of visual features k is 4. The learning rate is selected as 5e-5 and the dropout rate is set 0.2 to avoid overfitting. We leverage the AdamW [23] to optimize the whole parameters with the batch size 32. Following [37], we employ the longest common subsequence algorithm, common prefix and normalized edit distance between contexts and entities to obtain $|E| = 100$ candidate entities for each mention. The Top-k metrics are adopted to measure the performance of models and all the hyper-

Table 3: Results of MMEL ablation experiments on two datasets (JL: Joint Learning framework, v: visual information, t: textual information, PT: Pairwise Training scheme, CR: multi-mention Collaborative Ranking).

Datasets	**Methods**	**Metrics**			
		Top-1	Top-5	Top-10	Top-20
Wiki-MEL	MMEL	71.5	91.7	96.3	98.0
	- JL$_v$	41.2	71.7	84.8	92.5
	- JL$_t$	63.5	84.8	91.5	95.8
	- JL	27.8	47.2	56.2	65.8
	- PT&CR	63.7	88.8	94.5	96.3
NYTimes-MEL	MMEL	41.5	72.5	83.0	91.5
	- JL$_v$	40.7	70.9	82.0	90.0
	- JL$_t$	15.0	29.7	38.7	49.3
	- JL	14.3	28.5	37.5	48.3
	- PT&CR	38.3	67.8	80.0	91.0

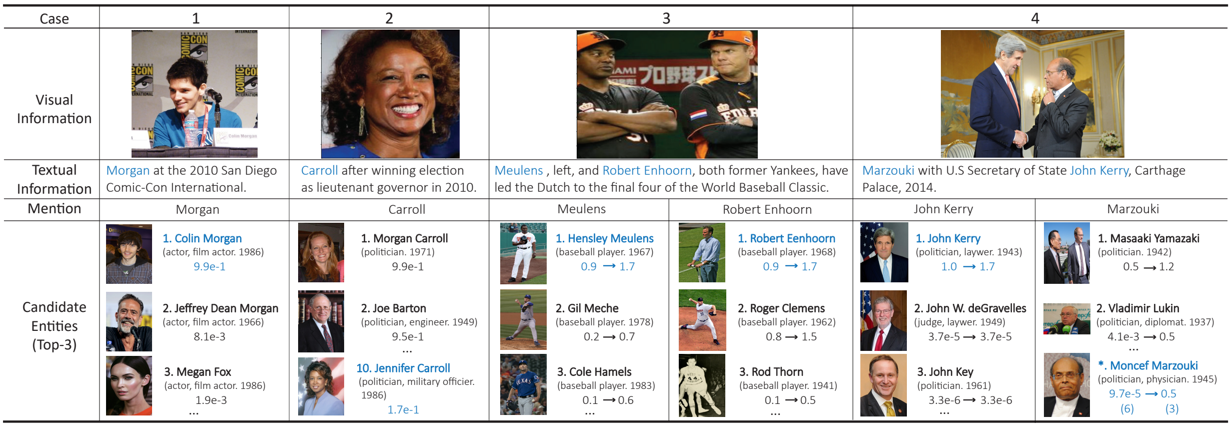
parameters are manually adjusted based on the top-5 result on the validation set.

4.2 RESULTS AND ANALYSIS

Results under Text-only and Multi-modal Settings (RQ1). As shown in Table 2, we have several observations. 1) In the text-only setting, our method achieves better performance compared with all the baselines with the improvements of 6.2%, 19.9%, 23.8%, and 21.1% for Top-1, 5, 10, and 20 accuracy results on the Wiki-MEL dataset, while about 19.6%, 28.3%, 20.9%, and 21.7% improvements on the NYTimes-MEL, respectively. The results illustrate the advantages of our MMEL framework to tackle the multi-mention entity linking task when only the textual information is available. 2) When only textual information is available, the results of collective entity linking method NCEL [6], which adopts a GCN to model the connection between the candidate entities of the current mention and the candidate entities of the neighbor mentions, are unsatisfying in our datasets and we attribute the poor performance to two reasons. The first one is that traditional collective entity linking methods [6, 46] target the document-level linking and fit for the cases with many mentions (more than 12 mentions for each context in [6, 46]). Whereas, the current multimodal entity linking task is only based on the sentence-level linking with no more than 5 mentions, and usually the sentence only has one mention. The second one is that these collective entity linking approaches ignore the negative impacts caused by negative candidate entities. Therefore, the contrastive learning in our framework, considering the correlation from both the positive and negative levels, can boost the linking performance to a large margin. 3) In the text-vision setting, our method also shows excellent performance under all the metrics with about 22.8% and 19.1% improvements on Wiki-MEL and NYTimes-MEL, respectively. Especially, the maximum improvement lies in the Top-1 accuracy of the MMEL, which is 27.9% higher than the GHMFC on Wiki-MEL, indicating the effectiveness of our method in the multimodal scenario. 4) It is worth noting that multimodal information can lead to better performance on the entity linking task. Compared with the MMEL-onlytext results, our MMEL has 31.3% and 4.4% improvements in the accuracy of Top-1 and Top-20 on Wiki-MEL, while 4.7% and 1.6% improvements on NYTimes-MEL. Considering the Top-1 as the difficult entity linking task and Top-20 as the simple task, these results illustrate that the usage of visual information can make the model capture the fine-grained features to facilitate the difficult entity linking task. 5) How to fuse the multimodal representation is also crucial in the entity linking task. Compared with the two variants of our method, the results show that our MMEL achieves the state-of-the-art performance on Wiki-MEL, but is not as good as MMEL-M-Encoder on NYTimes-MEL. Therefore, different multimodal fusion strategies can lead to different results and designing a universal approach is worth further exploring.

Results in Single-mention and Multi-mention Scenarios (RQ2). To illustrate the superiority of our MMEL framework for the multi-mention entity linking task, we divided the data in the test set into single-mention and multi-mention samples based on the number of mentions for each context. We compared our method with GHMFC and present the experimental results in Fig. 4, where we can observe that 1) the performance of our MMEL is 24.8% and 33.2% higher than GHMFC baseline for single-mention and multi-mention scenarios, respectively. Therefore, we can draw a conclusion that the improvement of our method mainly comes from the multi-mention entity linking performance. 2) Our proposed pairwise training scheme (PT) and multi-mention collaborative ranking (CR) have a positive impact on the entity linking task, especially for the multi-mention scenario. Without the PT and CR, the results show the 6.1%, 8.4%, and 0.8% decline on Top-1, Top-5, and Top-20 accuracies, respectively. These indicate that our PT and CR can help the model to link the entities more accurately and lead to excellent performance.

Ablation Study (RQ3). To validate the effectiveness of each module, we conduct the ablation study and the experimental results are shown in Table 3. From the table, we have the following observations. 1) For "- JL", our joint learning framework has made a qualitative leap in the entity linking task with about 40% improvements on both datasets. 2) For "- JL$_t$" and "- JL$_v$", the context-entity joint feature extraction is essential for both textual and visual modalities. In Wiki-MEL dataset, the results show that the visual joint learning plays an important role in the whole framework, while the textual feature extraction has a great impact on the linking accuracy in NYTimes-MEL dataset. 3) For "- PT&CR", our proposed pairwise training scheme and multi-mention collaborative ranking improve the final accuracy by 3.6% and



Case	1	2	3		4	
Visual Information						
Textual Information	Morgan at the 2010 San Diego Comic-Con International.	Carroll after winning election as lieutenant governor in 2010.	Meulens , left, and Robert Enhoorn, both former Yankees, have led the Dutch to the final four of the World Baseball Classic.		Marzouki with U.S Secretary of State John Kerry, Carthage Palace, 2014.	
Mention	Morgan	Carroll	Meulens	Robert Enhoorn	John Kerry	Marzouki
Candidate Entities (Top-3)	1. Colin Morgan (actor, film actor. 1986) 9.9e-1	1. Morgan Carroll (politician. 1971) 9.9e-1	1. Hensley Meulens (baseball player. 1967) 0.9 → 1.7	1. Robert Eenhoorn (baseball player. 1968) 0.9 → 1.7	1. John Kerry (politician, laywer. 1943) 1.0 → 1.7	1. Masaaki Yamazaki (politician. 1942) 0.5 → 1.2
	2. Jeffrey Dean Morgan (actor, film actor. 1966) 8.1e-3	2. Joe Barton (politician, engineer. 1949) 9.5e-1 ...	2. Gil Meche (baseball player. 1978) 0.2 → 0.7	2. Roger Clemens (baseball player. 1962) 0.8 → 1.5	2. John W. deGravelles (judge, laywer. 1949) 3.7e-5 → 3.7e-5	2. Vladimir Lukin (politician, diplomat. 1937) 4.1e-3 → 0.5 ...
	3. Megan Fox (actor, film actor. 1986) 1.9e-3 ...	10. Jennifer Carroll (politician, military officier. 1986) 1.7e-1	3. Cole Hamels (baseball player. 1983) 0.1 → 0.6 ...	3. Rod Thorn (baseball player. 1941) 0.1 → 0.5 ...	3. John Key (politician. 1961) 3.3e-6 → 3.3e-6 ...	*. Moncef Marzouki (politician, physician. 1945) 9.7e-5 → 0.5 (6) (3)

Figure 5: The cases for the entity linking task. The left two cases are single-mention and the right two cases are multi-mention. → refers to final entity candidate scores obtained through multi-mention collaborative ranking. $(\cdot)$ indicates ranking results.

2.9% on Wiki-MEL and NYTimes-MEL datasets, respectively. The results also illustrate that our pairwise training and multi-mention collaborative ranking can facilitate the model to tackle the difficult entity linking task since the improvement on Top-1 metric is more obvious.

Case Study (RQ4). We provide a few single-mention and multi-mention samples to illustrate the effectiveness of our MMEL in Fig. 5. The left two columns are single-mention cases, where we can observe that 1) our joint learning framework can capture the potential connection between contexts and entity candidates since the top-3 KG entities are all politicians when the context focuses on the topic of election. However, the model may still link the mention by mistake. We attribute the reasons to the large number of similar entities in the candidate set and the image variety issue [15]. The right two columns are multi-mention cases and we list the different ranking scores obtained before and after our multi-mention collaborative ranking. Here, we can observe that 2) our multi-mention collaborative ranking can help the model to link the entities more accurately (from 6th to 3rd), since it measures the potential connection between different mentions. That is, the entity candidates with less matching scores of mention "Marzouki" can be linked correctly when considering their relationships with entity candidates of the mention "John Kerry".

Runtime Discussion. Due to the involvement of joint learning framework, our method brings a trade-off between the inference time and ranking results. During testing, the running time is positively correlated with the number of candidate entities since we need to concatenate the mention and each candidate entity to obtain the corresponding representations. However, it is worth noting that our task focuses more on the fine ranking with the top-1 metric rather than the coarse ranking like the top-20 metric. Therefore, following [39], when given 10 candidate entities, the baseline GHMFC [37] will use 63.2s to process the whole test set with 5,256 cases in the Wiki-MEL dataset, while our framework will take 174.5s. Moreover, during the training stage, our proposed joint learning framework can train the model in parallel without training time overload. We also observe that the baseline GHMFC may reach its optimal state after about 70 epochs with about 110s per epoch on the Wiki-MEL dataset, while our method only takes 11 epochs with about 204s per epoch. These results illustrate that our method can converge faster than baselines.

5 CONCLUSION

Previous entity linking methods are mainly limited to the single-mention scenario and can hardly be generalized to the multi-mention scenario, restricting their performance correspondingly. In this paper, we first propose a joint learning framework to learn the features of contexts and entity candidates together, which can be employed in both text-only and multimodal settings. Then, we design a pairwise training scheme and a multi-mention collaborative ranking method to consider the potential connections between different mentions. The results on a public dataset and a self-constructed dataset also validate the effectiveness of our method. In future, we will explore the more efficient framework to tackle the multi-mention entity linking task with more useful multimodal information.

6 ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No.62171323), the National Key R&D Program of China (No.2020YFA0711400), the Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0100), the Fundamental Research Funds for the Central Universities, and the Start-up Grant (No.9610564) and the Strategic Research Grant (No.7005847) of City University of Hong Kong.

REFERENCES

[1] Omar Adjali, Romaric Besançon, Olivier Ferret, Hervé Le Borgne, and Brigitte Grau. Building a multimodal entity linking dataset from tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4285–4292, 2020.

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

[3] Pradeep K Atrey, M Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems*, 16:345–379, 2010.

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

[5] Yixin Cao, Lei Hou, Juanzi Li, and Zhiyuan Liu. Neural collective entity linking. In *COLING*, pages 675–686, 2018.

[6] Yixin Cao, Lei Hou, Juanzi Li, and Zhiyuan Liu. Neural collective entity linking. In *COLING*, pages 675–686, 2018.

[7] Xiang Chen, Ningyu Zhang, Lei Li, Shumin Deng, Chuanqi Tan, Changliang Xu, Fei Huang, Luo Si, and Huajun Chen. Hybrid transformer with multi-level fusion for multimodal knowledge graph completion. In *SIGIR*, pages 904–915, 2022.

[8] Tao Cheng, Xifeng Yan, and Kevin Chen-Chuan Chang. Entityrank: searching entities directly and holistically. In *VLDB*, pages 387–398, 2007.

[9] Yew Ken Chia, Lidong Bing, Soujanya Poria, and Luo Si. Relationprompt: Leveraging prompts to generate synthetic data for zero-shot relation triplet extraction. In *ACL*, pages 45–57, 2022.

[10] Guillem Collell, Ted Zhang, and Marie-Francine Moens. Imagined visual representations as multimodal embeddings. In *AAAI*, volume 31, 2017.

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

[12] Zhang Dongjie and Longtao Huang. Multimodal knowledge learning for named entity disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3160–3169, 2022.

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

[14] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013.

[15] Jingru Gan, Jinchang Luo, Haiwei Wang, Shuhui Wang, Wei He, and Qingming Huang. Multimodal entity linking: a new dataset and a baseline. In *ACM Multimedia*, pages 993–1001, 2021.

[16] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *EMNLP*, pages 6894–6910, 2021.

[17] Abhishek Gattani, Digvijay S Lamba, Nikesh Garera, Mitul Tiwari, Xiaoyong Chai, Sanjib Das, Sri Subramaniam, Anand Rajaraman, Venky Harinarayan, and AnHai Doan. Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach. *Proceedings of the VLDB Endowment*, 6(11):1126–1137, 2013.

[18] Bowei He, Xu He, Yingxue Zhang, Ruiming Tang, and Chen Ma. Dynamically expandable graph convolution for streaming recommendation. In *Proceedings of the ACM Web Conference 2023*, pages 1457–1467, 2023.

[19] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *EMNLP*, pages 782–792, 2011.

[20] Hexiang Hu, Yi Luan, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity linking.

[21] Tuan Lai, Heng Ji, and ChengXiang Zhai. Improving candidate retrieval with entity profile generation for wikidata entity linking. In *ACL (Findings)*, pages 3696–3711, 2022.

[22] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, pages 10965–10975, 2022.

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

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

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

[26] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *ECCV*, pages 529–544, 2022.

[27] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, pages 689–696, 2011.

[28] Minh C. Phan, Aixin Sun, Yi Tay, Jialong Han, and Chenliang Li. Pair-linking for collective entity disambiguation: Two could be better than all. *IEEE Trans. Knowl. Data Eng.*, 31(7):1383–1396, 2019.

[29] Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. Universal dependency parsing from scratch. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 160–170, Brussels, Belgium, October 2018.

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

[31] Ahmad Sakor, Isaiah Onando Mulang, Kuldeep Singh, Saeedeh Shekarpour, Maria Esther Vidal, Jens Lehmann, and Sören Auer. Old is gold: linguistic driven approach for entity and relation linking of short text. In *NAACL-HLT*, pages 2336–2346, 2019.

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

[33] Wenxiang Sun, Yixing Fan, Jiafeng Guo, Ruqing Zhang, and Xueqi Cheng. Visual named entity linking: A new dataset and a baseline. *arXiv preprint arXiv:2211.04872*, 2022.

[34] Alasdair Tran, Alexander Patrick Mathews, and Lexing Xie. Transform and tell: Entity-aware news image captioning. In *CVPR*, pages 13032–13042, 2020.

[35] Johannes M Van Hulst, Faegheh Hasibi, Koen Dercksen, Krisztian Balog, and Arjen P de Vries. Rel: An entity linker standing on the shoulders of giants. In *SIGIR*, pages 2197–2200, 2020.

[36] Denny Vrandecic and Markus Krötzsch. Wikidata: a free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, 2014.

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

[38] Shusen Wang, Bosen Zhang, Yajing Xu, Yanan Wu, and Bo Xiao. RCL: relation contrastive learning for zero-shot relation extraction. In *NAACL-HLT (Findings)*, pages 2456–2468, 2022.

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

[40] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *AAAI*, volume 33, pages 7216–7223, 2019.

[41] Chris Welty, J William Murdock, Aditya Kalyanpur, and James Fan. A comparison of hard filters and soft evidence for answer typing in watson. In *ISWC*, pages 243–256, 2012.

[42] Jennifer Williams, Ramona Comanescu, Oana Radu, and Leimin Tian. Dnn multimodal fusion techniques for predicting video sentiment. In *Proceedings of grand challenge and workshop on human multimodal language (Challenge-HML)*, pages 64–72, 2018.

[43] Hui Wu and Xiaodong Shi. Adversarial soft prompt tuning for cross-domain sentiment analysis. In *ACL*, pages 2438–2447, 2022.

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

[45] Mengge Xue, Weiming Cai, Jinsong Su, Linfeng Song, Yubin Ge, Yubao Liu, and Bin Wang. Neural collective entity linking based on recurrent random walk network learning. In *IJCAI*, pages 5327–5333, 2019.

[46] Mengge Xue, Weiming Cai, Jinsong Su, Linfeng Song, Yubin Ge, Yubao Liu, and Bin Wang. Neural collective entity linking based on recurrent random walk network learning. In *IJCAI*, pages 5327–5333, 2019.

[47] Li Zhang, Zhixu Li, and Qiang Yang. Attention-based multimodal entity linking with high-quality images. In *DASFAA*, pages 533–548, 2021.

[48] Wentian Zhao, Yao Hu, Heda Wang, Xinxiao Wu, and Jiebo Luo. Boosting entity-aware image captioning with multi-modal knowledge graph. *CoRR*, abs/2107.11970, 2021.

[49] Xiaoling Zhou, Yukai Miao, Wei Wang, and Jianbin Qin. A recurrent model for collective entity linking with adaptive features. In *AAAI*, pages 329–336, 2020.