electronics

MDPI

Article

Pre-Training and Fine-Tuning with Next Sentence Prediction for Multimodal Entity Linking

Lu Li [1,2], Qipeng Wang [3,4], Baohua Zhao [5], Xinwei Li [1], Aihua Zhou [5] and Hanqian Wu [3,4,*]

[1] School of Cyber Science and Engineering, Southeast University, Nanjing 210000, China; lilu-seu@seu.edu.cn (L.L.); lixinwei@seu.edu.cn (X.L.)
[2] Suzhou Centennial College, Suzhou 215000, China
[3] School of Computer Science and Engineering, Southeast University, Nanjing 210000, China; qipengwang@seu.edu.cn
[4] Key Laboratory of Computer Network and Information Integration of Ministry of Education, Southeast University, Nanjing 210000, China
[5] State Grid Smart Grid Research Institute Co., Ltd., Beijing 100000, China; zhaobaohua@geiri.sgcc.com.cn (B.Z.); zhouaihua@geiri.sgcc.com.cn (A.Z.)
* Correspondence: hanqian@seu.edu.cn

Abstract: As an emerging research field, more and more researchers are turning their attention to multimodal entity linking (MEL). However, previous works always focus on obtaining joint representations of mentions and entities and then determining the relationship between mentions and entities by these representations. This means that their models are often very complex and will result in ignoring the relationship between different modal information from different corpus. To solve the above problems, we proposed a paradigm of pre-training and fine-tuning for MEL. We designed three different categories of NSP tasks for pre-training, i.e., mixed-modal, text-only and multimodal and doubled the amount of data for pre-training by swapping the roles of sentences in NSP. Our experimental results show that our model outperforms other baseline models and our pre-training strategies all contribute to the improvement of the results. In addition, our pre-training gives the final model a strong generalization capability that performs well even on smaller amounts of data.

Keywords: entity linking; multimodal; pre-training

Citation: Li, L.; Wang, Q.; Zhao, B.; Li, X.; Zhou, A.; Wu, H. Pre-Training and Fine-Tuning with Next Sentence Prediction for Multimodal Entity Linking. *Electronics* **2022**, *11*, 2134. https://doi.org/10.3390/electronics11142134

Academic Editor: Kamil Dimililer

Received: 1 June 2022
Accepted: 5 July 2022
Published: 7 July 2022

1. Introduction

Entity linking (EL) aims to map an ambiguous mention to a unique entity of a target knowledge base (KB). It is an essential component of many information extraction and natural language understanding (NLU) pipelines [1]. There are already many results in this research field which can be roughly categorized into: (a) Local information that considering the mentions in the corpus individually [2]; (b) Global information that consider the impact of the preceding mentions that have already been processed [3].

Recently, with the rapid development of social media, the demand for MEL is rising because images and text often appear in pairs on these platforms (e.g., Twitter). MEL can play a significant role in uncovering the value behind these platform data. For instance, the results of MEL for users' tweets are useful for accurately profiling users (e.g., A basketball fan is likely to mention basketball players in his or her tweets). Therefore, increasing attention has been placed on the MEL. Further, it remains a challenge to process short spoken-text in the corpus from social media. However, by utilizing the image information, the performance can be dramatically improved [4]. As demonstrated in Figure 1, without relying on the additional input from the image, it is difficult to deduce the identity of Taylor Swift from the given text description (i.e., hand hearts taylor's version).

Figure 1. An example of tweets where taylor in the text is an ambiguous mention.

Nowadays, there have been many research results on MEL [5–7]. However, most current studies have a serious drawback that can not be ignored. These models usually require first obtaining visual and textual feature information separately and then require the design of a complex model for the integration of visual and textual features and the subsequent entity linking. In other words, these models usually divide MEL into modal fusion and entity linking. So, the whole process of MEL is very complex and difficult. Even if we directly use some existing multimodal pre-training models, such as LXMERT [8], ViLBERT [9], VL-BERT [10], VisualBERT [11], to obtain multimodal information from the corpus; because these models are not pre-trained specifically for MEL, the features we can get are not suitable for MEL and a complex model is still required.

To address this issue, we propose a paradigm of pre-training and fine-tuning for MEL via next sentence prediction (NSP) tasks. To the best of our knowledge, only our study focuses on improving MEL performance through pre-training. Our proposed approach not only has a very simple model structure, but also a very simple pre-training and fine-tuning process. Considering that the model needs to determine the relationship between mentions and entities by their features, we used three different types of Next Sentence Prediction(NSP) tasks to pre-train our model and swapped the roles of sentences to increase the amount of pre-trained data. These three types of tasks are, respectively, to determine the

inter-sentential relationship between text and images (i.e., mixed-model NSP), to determine the inter-sentential relationship between text and text (i.e., text-only NSP) and to determine the inter-sentential relationship between image text pairs (i.e., multimodal NSP). Here, mixed-model NSP and multimodal NSP means two different granularity multimodal tasks. Specifically, mixed-model will determine the relationship between different modalities from different sentence (e.g., a text from the first sentence and an image from the second sentence) while multimodal will treat this one text and one image as a whole and determine the relationship between these pairs. As for text-only, this task helps the original pre-trained model to adapt to the characteristics of social media texts (i.e., short and colloquial). After that, for fine-tuning stage, to better mine the knowledge in the pre-trained model, we firstly modify the text formatting as Sun did [12] to make the two sentences semantically smooth. Then, we fine-tuned the model with the same framework as the multi NSP pre-training. Our contributions are as follows:

- We introduce a paradigm of pre-training and fine-tuning for MEL. To the best of our knowledge, our work is first to improving MEL performance by pre-training.
- We introduce three different categories of NSP tasks to further pre-train and connect multimodal corpus and entity information in an appropriate way for fine-tuning.
- We conduct extensive experiments on a multimodal corpus based on Twitter, including both the general plain text and multimodal investigation.

2. Related Work

2.1. Multimodal Entity Linking

Compared to general Entity Linking that uses only textual information, MEL is an emerging research area and aims at improving the performance by also utilizing information from other modalities which always means visual information. Moon et al. [13] were the first to introduce and solve the task of Multimodal Named Entity Disambiguation (MNED) which can also be considered as MEL. They used a multimodal dataset constructed from a subset of Freebase [14] to evaluate their approach. Adjali et al. [4] constructed a multimodal entity linking dataset based on tweets and a multimodal KB in which every entity denotes a user of Twitter. They also proposed a multimodal entity linking model for whom a very important part is used to obtain joint multimodal representation [5]. Gan et al. [6] released their dataset named MultiModal Movie Entity Linking (M3EL). Their approach models the alignment of textual and visual mentions as a bipartite graph matching problem and solves it with an optimal-transportation-based linking method. In addition, to solve the MEL for Weibo which is also a social media platform, Zhang et al. [7] firstly removed the negative impact of noisy images by a two-stage image and text correlation mechanism and then captured the connection between mention representation and entity by a multiple attention mechanism.

2.2. Multimodal Pre-Training

Recently, Multimodal Pre-training has received a lot of attentions because of the success of previous work in single-modality areas (e.g., BERT [15]). Many researchers have proposed their multimodal pre-training models. On the one hand, many models try to obtain generic multimodal representations to facilitate downstream tasks. ViLBERT [9] is a multimodal two-stream model which was proposed based on the popular BERT architecture. It can learn task-agnostic joint representations of image content and text. Su et al. [10] proposed VL-BERT which can learn a new pre-trainable generic representation for visual-linguistic tasks and was designed to fit most of the visual-linguistic downstream tasks. Li et al. [16] introduced UNIMO which can effectively adapt to both single-modal and multimodal understanding and generation tasks and can learn more generalizable representations by allowing textual knowledge and visual knowledge to enhance each other in the unified semantic space. Zhang et al. [17] improved the performance across all vision language tasks by improving visual representations within VinVL. On the other hand, many models are pre-trained to perform well on specific tasks. For vision-and-

language navigation (VLN) tasks, Hao et al. [18] proposed PREVALENT which outperforms other methods significantly. Singh et al. [19] proposed STL-CQA whose performance was significantly better than other methods for Chart Question Answering (CQA). A new visual machine reading comprehension dataset (i.e., VisualMRC) is introduced by Tanaka et al. The model they proposed not only outperforms the base sequence-to-sequence models but also outperforms the SOTA VQA model [20].

3. Proposed Method

In this section, based on BERT, our pre-training and fine-tuning framework for MEL will be introduced. We will first define the MEL task, and then some terms we used are also explained. Next, we will introduce our three different categories of pre-training tasks, and Figures 2–4 show the overview of the tasks. Furthermore, the framework of fine-tuning for MEL is the same as the framework for multimodal pre-training in Figure 4.

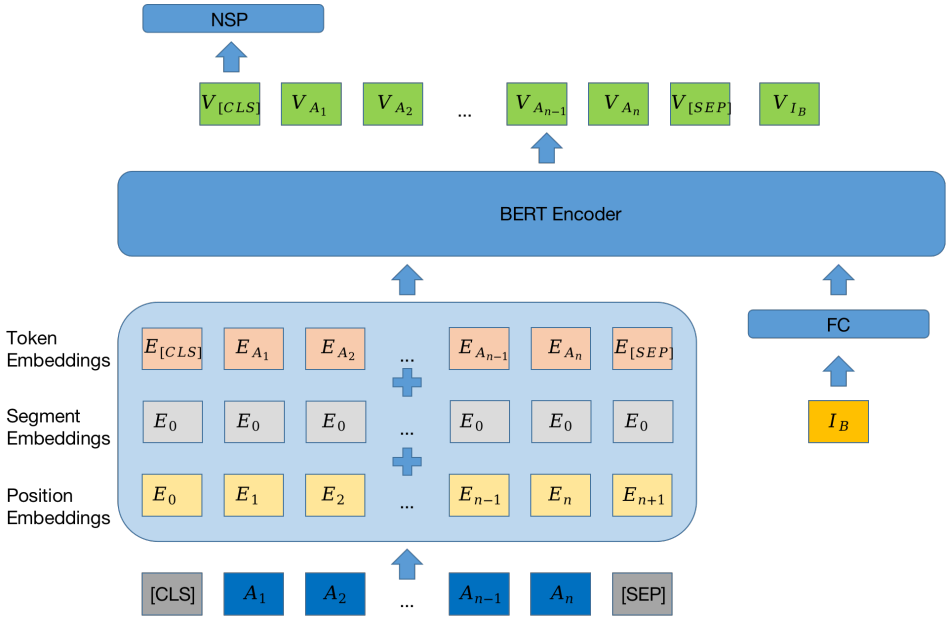

Figure 2. The framework of mixed-modal NSP pre-training. A_i means the text token and its embedding is obtained as general BERT. Token embedding is obtained by a bert embedding matrix. Segment embedding marks whether the token belongs to the first or second sentence. Position embedding marks the token's position.

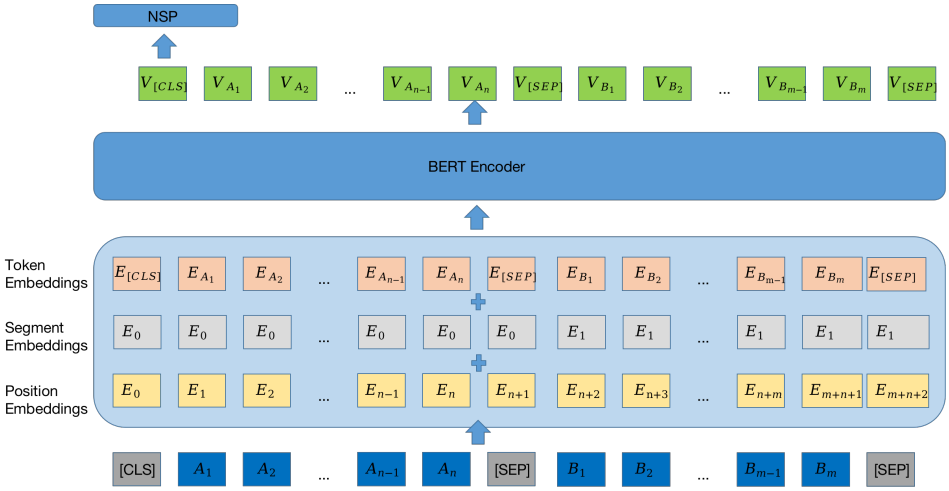

Figure 3. The framework of text-only NSP pre-training. It is exactly the same as the general BERT structure for NSP.

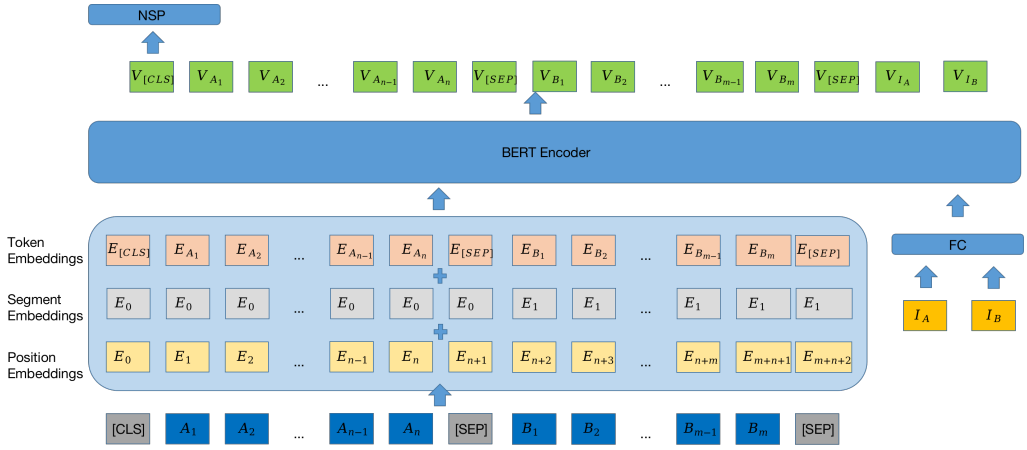

Figure 4. The framework of multimodal NSP pre-training and MEL fine-tuning. Text formats for input in the pre-training and fine-tuning phases are different.

3.1. Task Definition

Just like the general entity linking, the goal of MEL is also mapping the ambiguous mention in input to the entity in a KB. Specifically, in our task, we have the input set $Input = \{tweet_1, tweet_2, ..., tweet_n\}$ where $tweet_i$ denotes a tweet consisting of a text-image pair $(text_i, image_i)$, and the mention in the $text_i$ have been already marked as m_i. The KB in our task is a set of entities, i.e., $KB = \{e_1, e_2, ..., e_m\}$, where e_j denotes a twitter user and can be represented as a two-tuple of user information and timeline $e_j = (userInfo_j, timeLine_j)$. The $userInfo_j$ comes from user profile and includes *screenName* (e.g., @AndrewYNg), *userName* (e.g., Andrew Ng), *userLocation* (e.g., Palo Alto, CA, USA) and *userDescription* (e.g., Co-Founder of Coursera; Stanford CS adjunct faculty. Former head of Baidu AI Group/Google Brain. #ai #machinelearning, #deeplearning #MOOCs) and the $timeLine_j$ denotes all the tweets with images posted by this user.

Before we map the ambiguous mention to a certain entity, candidate generation is an essential step. A candidate set $Cand(m_i)$ according to the mention can be generated by surface form matching, dictionary lookup and prior probability. In this paper, entities whose *userNames* contain a mention are considered candidate entities for this mention.

Finally, a score between the mention m_i in the input tweet and the entity e_j in the candidate set $Cand(m_i)$ is obtained as an evaluation of correlation, and the entity with the highest score is considered to be the ground truth entity.

$$score_i^j = Score(m_i, e_j), e_j \in Cand(m_i)$$
$$e_i^t = argmax\, score_i^j$$

3.2. Representation

In this section, we will introduce the textual data format used in this paper and the way to the extract visual features. For the mention representation, the *mention context* denotes the context of the mention in the tweet, and *mention image* denotes the image of the tweet. We use the tokenizer to process the text and pre-trained ResNet50 model to obtain 2048-dimensional image features. The 2048-dimensional image features are finally converted to the same dimension as the text features by a fully-connected layer.

$$mention_{tokens} = Tokenizer(mention\ context)$$
$$Res_{mention} = ResNet50(mention\ image)$$
$$mention_{img} = W * Res_{mention} + b$$

W is a 2048×768 matrix and b is a 768-dimensional vector.

Considering that the entities in the KB we used are regular users of Twitter, we use their user profiles to construct the *entity context*. If the corresponding information is not

available, *NA* is selected as a default. *userDescription* uses up to the first 30 words. $tweet_1$, $tweet_2$ and $tweet_3$ mean the last 3 tweets in the *timeLine* of this user. In addition, considering the length of the *entity context*, each tweets selects at most the first 35 words.

The format of the *entity context* is: the username of *@userScreen* who lives in *userLocation* is *userName* and the description is *userDescription*. some recent tweets of the user include: $tweet_1$, $tweet_2$ and $tweet_3$.

The *entity context* is handled like *mention context*. For the visual features of entities, we process the all images from the user's *timeLine* like the mention images, and the final visual representation is obtained by averaging over all these image features.

$$entity_{tokens} = Tokenizer(entity\ context)$$

$$Res_{entity} = \frac{1}{|T(e)|} \sum_{img \in T(e)} ResNet50(img)$$

$$entity_{img} = W * Res_{entity} + b$$

$T(e)$ is the image set of entity's *timeLine* and $|T(e)|$ means the size of the set. W and b are the same as in the processing of $Res_{mention}$ above.

3.3. Pre-Training

In this section, we introduce three different types of NSP for pre-training, i.e., mixed-modal, text-only, and multimodal. For NSP, we have $sentence = \{text, image, entity_{context}, entity_{image}\}$ where *text* and *image* belong to the same input tweet and $entity_{context}$ and $entity_{image}$ belong to the same entity which is the ground truth entity of the mention in the previous input tweet. Therefore, unlike the general NSP task, the input information and the entity information parts of the same *sentences* we construct are related. With 50% probability, for each sentence in the set of sentences $sentences = \{sentence_1, sentence_2, ..., sentence_n\}$, we randomly select a sentence from the set as its next sentence $sentence^n = \{text^n, image^n, entity^n_{context}, entity^n_{image}\}$ and set the label *isNext* to 1. With another 50% probability, we select itself as the next sentence and set the label *isNext* to 0.

Next, for mixed-modal NSP, we use the *text* from *sentence* and $entity^n_{image}$ from $sentence^n$ to predict the label *isNext*. In addition to this, the $entity_{context}$ in *sentence* and the $image^n$ in $sentence^n$ are also used to predict the label *isNext*. This type of tasks can help our model to better interact text and image features to determine whether text are related to images.

As for text-only NSP, the *text* in *sentence* and the $entity^n_{context}$ in $sentence^n$ are used and it allows the model to adapt to the textual features of the tweets. The framework of this task is shown in Figure 3 and is the same as NSP in BERT [15].

Then, for multimodal NSP, the *text* and *image* in *sentence* and $entity^n_{context}$ and $entity^n_{image}$ in $sentence^n$ are used to predict the *isNext*. The model can improve the classification performance on multimodal corpus by this task.

Finally, the roles of *sentence* and *next sentence* in NSP are exchanged, i.e., the original *sentence* becomes new $sentence^n$ and original $sentence^n$ becomes new *sentence*. The amount of data used for pre-training is doubled in this way. For the above pre-training task, we use cross-entropy as the loss function. The goal of training is to minimize the loss value.

$$L_{i \in \{mix,text,multi\}} = \sum isNext \log predict + (1 - isNext) \log(1 - predict)$$

$$L_{pre} = L_{mix} + L_{text} + L_{multi}$$

isNext means the label we set for the NSP pre-training tasks and *predict* means the probability that our model predicts that the label is 1. L_i is the loss value of each type of pre-training task itself and L_{pre} is the final loss value for pre-training.

3.4. Linking

After pre-training the model, we further fine-tuned the model to make it perform better on MEL. Just as Sun's study [12] shows, the NSP task is better able to determine whether the two sentences are related if they are consistent with natural language logic.

So ***mention context*** is changed to ***mention context. mention is*** where *mention* denotes the mention in *mention context* and ***entity context*** is changed to ***userScreen. entity context*** where *userScreen* is from *entity context*. During the training phase, just as in the multimodal NSP task, the new mention context and the new entity context of $Cand(m)$ are fed into the model to determine whether the entity context is the next of the mention context. The loss function used here is also cross-entropy and the goal is to minimize the loss value.

$$L_{linking} = \sum isNext \log predict + (1 - isNext) \log(1 - predict)$$

4. Experiment

4.1. Dataset and Experiment Settings

The dataset we use is released by Adjali et al. [4]. Because of the constraint of Twitter, they only published the ids of the tweets they collected and we have to reconstructed the dataset and the KB. However, because of the temporal validity of the data, the collected data is very different from the one they use. We clean up the data that is not applicable due to the change in user's profiles, so that the ground truth entity of each mention in the dataset is in the KB, and each entity in the KB is a candidate entity of at least one mention. We will also release our code and data together. The statistics of the filtered data are shown in Table 1.

Table 1. Statistics of our reconstructed dataset and knowledge base.

	Number of Tweets	**Number of Mentions or Entities**
groundTruth	57,905	1553
KB	2,478,625	18,434

Table 2 shows the statistics of the datasets we divided. We first group the data according to the ground truth entity corresponding to the data and then divide these groups equally into random and unique parts. Next, the groups belonging to the unique part will be divided equally into *trainSet*, *devSet* and *testSet*. The groups belonging to the random part are regrouped into one group, and 10%, 80%, and 10% of this group of data are randomly added to *trainSet*, *devSet* and *testSet*. Finally, *trainSet07* and *trainSet05* are obtained by discarding 30% and 50% of the random part of the *trainSet*, respectively. The above way of data division requires more generalization ability of the model. Many entities in *testSet* will not be seen in pre-training or fine-tuning stage. Furthermore, *trainSet07* and *trainSet05* can also help demonstrate the generalization ability of our model.

Table 2. Statistics of the datasets divided from groundTruth.

	Unique	**Random**	**Totals**
trainSet	7000	3619	10,619
devSet	6923	3619	10,542
testSet	7795	28,949	36,744
trainSet07	7000	2533	9533
trainSet05	7000	1809	8809

As for the implementation details, the BERT-base is our base model and the accuracy [21] of linking is the evaluation metric because mentions have been marked in our dataset. Just as shown in Table 3, the max length of the text part we feed into our model is 300. For pre-training, the model is trained for 5 epochs with a batch size of 4. For fine-tuning, the model is trained for 4 epochs with a batch size of 16 and we go with negative samples as Zhang et al. [7] do. The learning rate are all set to 5×10^{-6} and decreases linearly. The optimizer for both is *AdamW*. We set the random seed to 0, 21, 42, 63 and 84, respectively, and average the results as the final result.

Table 3. Parameter settings for experiment.

Name	Value
max len	300
batch size for pre-training	4
batch size for fine-tuning	16
epochs for pre-training	5
epochs for fine-tuning	4
lr	5×10^{-6}
optimizer	AdamW

4.2. Baselines

There are not too many multimodal entity linking models available and many researchers proposed their models on their own datasets. We do not think any MEL model can be called the SOTA model because their results can not be fairly compared. We choose a recent model which also is based on a multimodal social media corpus as the baseline and we also demonstrate the effectiveness of pre-training by ablation studies. The models are as follows. **Att** is proposed by Zhang et al. [7] and it has excellent multimodal entity linking performance on the datasets they build which is also based on multimodal social media corpus. **FMEL** is the model we propose and it only use *trainSet* for fine-tuning. **PFMELm** is first pre-trained with *trainSet* with mixed-modal NSP and then uses *trainSet* for fine-tuning. **PFMELmt** is first pre-trained with *trainSet* with mixed-modal NSP and text-only NSP and then uses *trainSet* for fine-tuning. **PFMELmtm** is first pre-trained with *trainSet* with mixed-modal NSP, text-only NSP and multimodal NSP and then uses *trainSet* for fine-tuning. The pre-training tasks for **PFMELmtd** are the same as **PFMELmt**, but we double the amount of pre-trained data by swapping roles between sentences. **PFMEL** which is trained with all three types of pre-training tasks using post-doubling data is then trained with *trainSet* for fine-tuning. **PFMEL07** has the same pre-training process as the **PFMEL** and is fine-tuned with *trainSet*07. **PFMEL05** has the same pre-training process as the **PFMEL** and is fine-tuned with *trainSet*05.

In order to show that our model makes beneficial use of image information, we also compare our model with other models for general entity linking. Compared to the study of MEL, there have been many successful models, and models who have good performers in AIDA-CoNLL are selected as our baseline. The models are as follows. **CHOLAN** is proposed by Ravi et al. [22] and it is an end-to-end entity linking model with great performance. We just use the entity disambiguation part which is achieved by sentence classification. **Autoregressive** is proposed by De Cao et al. [23] and it is also an end-to-end entity linking model and outperforms state-of-the-art approaches on AIDA-CoNLL. We also just use the entity disambiguation part. **NSP-BERT** is proposed by Sun et al. [12] and it has the same structure as our proposed model when using plain text.

4.3. Results

We report in Table 4 the accuracy performance of models whose structures are very different. First, we can see that **NSP-BERT** achieves an accuracy of 73.946% and is the best of the models that only utilize textual features. This also can demonstrate that fine-tuning with NSP after formatting the two sentences is very effective for text-only EL. The reason for **NSP-BERT**'s high performance is that the knowledge contained in BERT was effectively mined by connecting the two sentences in a form that conforms to natural language. In the case of multimodal entity linking, the model we proposed has the best performance and achieves an accuracy of 75.192% and significantly outperforms another MEL model (i.e., **Att**). The accuracy of our model outperforms **NSP-BERT**'s 1.246% and it can effectively demonstrate that our model successfully utilizes the visual information. However, considering the slight drop in accuracy of **FMEL** compared to **NSP-BERT**, it is undesirable to input visual features and textual features together directly to the model and our pre-training strategy worked. One possible reason is that there is not much

association between images and text, and the image information may drag down the model performance as noise. Another possible reason is that the original BERT does not effectively use both image and text information, even though the images are closely related to the mention in the text.

Table 4. Results of different approaches. When our model is not pre-trained and uses only textual information, its structure is the same as NSP-BERT.

TEXT	**Dev**	**Test**
NSP-BERT	0.6045	0.73946
CHOLAN	0.52832	0.70818
Autoregressive	0.28302	0.59606
TEXT-IMAGE		
Att	0.32634	0.53424
FMEL	0.60406	0.73926
PFMEL	0.60268	0.75192

The result of the ablation study of our model is shown in Table 5. The accuracy of **PFMELm** outperforms **FMEL** 0.724% and it can demonstrate that mixed-modal NSP is useful. In addition, the effectiveness of text-only NSP is reflected by the accuracy improvement of **PFMELmt** compared to **PFMELm**. Whereas we can see that the accuracy of **PFMELmtm**, as well as **PFMELmtd** decreases, compared to **PFMELmt**, the final model has higher accuracy than **PFMELmt**, which can indicate that multimodal NSP and Sentence role exchange need to be implemented together to produce beneficial effects.

Table 5. Results of ablation study

	Dev	**Test**
FMEL	0.60406	0.73926
PFMELm	0.62462	0.7465
PMELmt	0.59428	0.74836
PFMELmtm	0.59008	0.74378
PFMELmtd	0.59906	0.7447
PFMEL	0.60268	0.75192

To illustrate the generalizability of our model, we reduced the size of the random part in *trainSet* during fine-tuning, and the results are shown in Table 6. Obviously, we can see that reducing the amount of data leads to a decrease in accuracy. However, after reducing the random part of the *trainSet* to 70%, the accuracy is still higher than that of the **FMEL**. Furthermore, even further be reduced to 50%, the accuracy is just only slightly lower.

Table 6. Results with different dataset sizes for fine-tuning.

	Dev	**Test**
FMEL	0.60406	0.73926
PFMEL05	0.59192	0.73916
PFMEL07	0.61292	0.74758
PFMEL	0.60268	0.75192

4.4. Error Analysis

To further understand our approach, we observed the specific results of linking. We found that our model sometimes does not distinguish well between entities of similar topics. For example, for the input shown in Figure 5, the mention *wiggins* should be linked to the Twitter user *andrew wiggins(@22wiggins)* who is an NBA basketball player. However, our model will actually link this mention to *jermaine wiggins(@jwiggs85)* who used to be an

NFL player. One important reason for this situation may be that our manually designed entity context is not perfect. Even if we can see nba in the text and a basketball player in the picture, we will still fail because of the lack of differentiation between entities. It is also possible that there are many tweets whose images do not relate to the text content. This allows the model to reduce the weight of image features when making judgments.

#badbeat from #nba sunday ! prop = o / u 6.5 rebs + asts for wiggins wiggins would pull down his 3rd reb to go along with his 4 asts on the night with 17 seconds left in the 4th quarter ! final = 7 rebs + asts = over. wiggins is

Figure 5. An example of a result error where wiggins should be linked to a basketball player rather than an NFL player.

5. Conclusions

As an essential component of many information extraction and NLU pipelines, the study of MEL can be of great value. In this paper, we proposed a paradigm of pre-training and fine-tuning for MEL. This paradigm is simple but effective. We designed three different categories of NSP tasks for pre-training based on the BERT-base. We do not need to modify the structure of BERT except for adding a fully-connected layer to transform the image feature dimension. Our experiments can clearly show that our approach which has a powerful generalization capability is superior to other baseline models and our pre-training strategies are contributing to our model. In future work, we will try to introduce other pre-training tasks to improve the performance. For example, in the case of multimodal inputs, we will treat the text part just like MLM in BERT. In addition, multimodal pre-training tasks can be introduced to further exploit image information. Considering that the entity context we designed in the paper may not be perfect, it would also be interesting to study how to represent a Twitter user more comprehensively by text mining.

Author Contributions: Conceptualization, L.L., Q.W. and H.W.; Data curation, Q.W.; Formal analysis, L.L. and B.Z.; Funding acquisition, H.W.; Investigation, L.L., Q.W. and X.L.; Visualization, B.Z. and X.L.; Validation, L.L. and H.W.; Resources, B.Z., A.Z. and H.W.; Supervision, B.Z., A.Z. and H.W.; Writing—original draft, Q.W.; Writing—review and editing, L.L. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Science and Technology Project of State Grid Corporation of China under grant 5700-202158177A-0-0-00.

Data Availability Statement: The code and data will be released here: https://github.com/numsi/mel_nsp (accessed on 1 May 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sevgili, O.; Shelmanov, A.; Arkhipov, M.; Panchenko, A.; Biemann, C. Neural Entity Linking: A Survey of Models Based on Deep Learning. *arXiv* **2021**, arXiv:2006.00575.
2. Csomai, A.; Mihalcea, R. Linking Documents to Encyclopedic Knowledge. *IEEE Intell. Syst.* **2008**, *23*, 34–41. [CrossRef]
3. Yang, X.; Gu, X.; Lin, S.; Tang, S.; Zhuang, Y.; Wu, F.; Chen, Z.; Hu, G.; Ren, X. Learning Dynamic Context Augmentation for Global Entity Linking. *arXiv* **2008**, arXiv:1909.02117.
4. Adjali, O.; Besançon, R.; Ferret, O.; Le Borgne, H.; Grau, B. Building a Multimodal Entity Linking Dataset From Tweets. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020; pp. 4285–4292.
5. Adjali, O.; Besançon, R.; Ferret, O.; Le Borgne, H.; Grau, B. Multimodal Entity Linking for Tweets. In *Advances in Information Retrieval*; Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2020; Volume 12035, pp. 463–478.
6. Gan, J.; Luo, J.; Wang, H.; Wang, S.; He, W.; Huang, Q. Multimodal Entity Linking: A New Dataset and A Baseline. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 993–1001.
7. Zhang, L.; Li, Z.; Yang, Q. Attention-Based Multimodal Entity Linking with High-Quality Images. In *Database Systems for Advanced Applications*; Jensen, C.S., Lim, E.P., Yang, D.N., Lee, W.C., Tseng, V.S., Kalogeraki, V., Huang, J.W., Shen, C.Y., Eds.; Springer: Berlin/Heidelberg, Germany, 2021; pp. 533–548.
8. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; pp. 5100–5111.
9. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2019; Volume 32.
10. Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; Dai, J. VL-BERT: Pre-Training of Generic Visual-Linguistic Representations. *arXiv* **2016**, arXiv:1908.08530.
11. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.J.; Chang, K.W. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv* **2019**, arXiv:1908.03557.
12. Sun, Y.; Zheng, Y.; Hao, C.; Qiu, H. NSP-BERT: A Prompt-Based Zero-Shot Learner Through an Original Pre-Training Task–Next Sentence Prediction. *arXiv* **2021**, arXiv:2109.03564.
13. Moon, S.; Neves, L.; Carvalho, V. 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), Melbourne, Australia, 15–20 July 2018; pp. 2000–2008.
14. Bast, H.; Bäurle, F.; Buchhold, B.; Haußmann, E. Easy Access to the Freebase Dataset. In Proceedings of the 23rd International Conference on World Wide Web—WWW '14 Companion, Seoul, Korea, 7–1 April 2014; pp. 95–98.
15. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
16. Li, W.; Gao, C.; Niu, G.; Xiao, X.; Liu, H.; Liu, J.; Wu, H.; Wang, H. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Virtual Event, 1–6 August 2021; pp. 2592–2607.
17. Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; Gao, J. VinVL: Revisiting Visual Representations in Vision-Language Models. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5575–5584.
18. Hao, W.; Li, C.; Li, X.; Carin, L.; Gao, J. Towards Learning a Generic Agent for Vision-and-Language Navigation via Pre-Training. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 13134–13143.
19. Singh, H.; Shekhar, S. STL-CQA: Structure-Based Transformers with Localization and Encoding for Chart Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 3275–3284.
20. Tanaka, R.; Nishida, K.; Yoshida, S. Visualmrc: Machine reading comprehension on document images. In Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; pp. 13878–13888.
21. Shen, W.; Wang, J.; Han, J. Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions. *IEEE Trans. Knowl. Data Eng.* **2014**, *27*, 443–460. [CrossRef]
22. Kannan Ravi, M.P.; Singh, K.; Mulang', I.O.; Shekarpour, S.; Hoffart, J.; Lehmann, J. CHOLAN: A Modular Approach for Neural Entity Linking on Wikipedia and Wikidata. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 504–514.
23. De Cao, N.; Aziz, W.; Titov, I. Highly Parallel Autoregressive Entity Linking with Discriminative Correction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021; pp. 7662–7669.