2021 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-3864-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICME51207.2021.9428274

MNRE: A CHALLENGE MULTIMODAL DATASET FOR NEURAL RELATION EXTRACTION WITH VISUAL EVIDENCE IN SOCIAL MEDIA POSTS

Changmeng Zheng†, *Zhiwei Wu*†, *Junhao Feng*†, *Ze Fu*†, *Yi Cai**

Key Laboratory of Big Data and Intelligent Robot (SCUT), MOE of China
School of Software Engineering, South China University of Technology, Guangzhou, China
sethecharm@mail.scut.edu.cn, ycai@scut.edu.cn

ABSTRACT

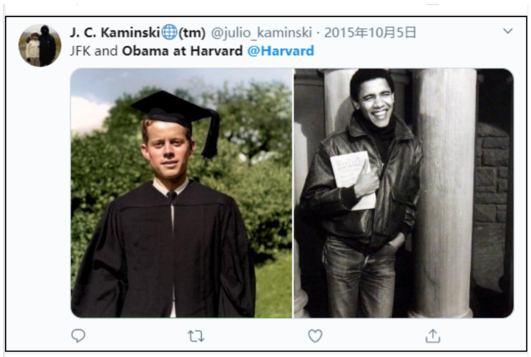
Extracting relations in social media posts is challenging when sentences lack of contexts. However, images related to these sentences can supplement such missing contexts and help to identify relations precisely. To this end, we present a multimodal neural relation extraction dataset (MNRE), consisting of 10000+ sentences on 31 relations derived from Twitter and annotated by crowdworkers. The subject and object entities are recognized by a pretrained NER tool and then filtered by crowdworkers. All the relations are identified manually. One sentence is tagged with one related image. We develop a multimodal relation extraction baseline model and the experimental results show that introducing multimodal information improves relation extraction performance in social media texts. Still, our detailed analysis points out the difficulties of aligning relations in texts and images, which can be addressed for future research. All details and resources about the dataset and baselines are released on https://github.com/thecharm/MNRE.

Index Terms— datasets, relation extraction, visual evidence, social media posts

1. INTRODUCTION

Relation extraction (RE) is the task of predicting attributes or relations between two named entities in a sentence. RE plays an important role in large-scale knowledge graph construction and benefits for many downstream tasks. Traditional kernel-based methods [1] or embedding methods [2] relied on heavy human-annotated data and is time-consuming and hard to generalize well.

Recently, neural network based methods achieve great success with different feature extractors [3]. Most of these works concern about newswire domain where sentences are formal and complete. However, the RE performance drops dramatically when texts are short and lack of contexts in social media posts. As a result, most of supervised neural models cannot precisely identify relations without enough contexts. Distant supervision is an alternative method to relieve this problem which leverages the alignment of knowledge bases and texts in sentences to automatically annotate relations [4]. However, distant supervision suffers from the wrong labeling problem, which is even worse when contexts are missing. Hence, it is necessary to supplement the missing semantic information with text-related contents like the related visual contents.



J. C. Kaminski🌐(tm) @julio_kaminski · 2015年10月5日
JFK and Obama at Harvard @Harvard

Named Entities:	*JFK* (PER), *Obama* (PER), *Harvard* (ORG)
Relations extracted by Test-based Methods:	(Residence, *JFK*, *Harvard*) (Residence, *Obama*, *Harvard*) (Spouse, *JFK*, *Obama*)
Gold Relations:	(Graduated at, *JFK*, *Harvard*) (Graduated at, *Obama*, *Harvard*) (Alumni, *JFK*, *Obama*)
Supporting Evidence:	Bachelor Cap, Gown, Book

Fig. 1. An example of multimodal relation extraction in Twitter. Previous text-based methods incorrectly identify the relations when lacking of contexts. However, we can extract relations precisely with the supporting evidence from visual contents.

Visual contents, such as the image posts in Twitter, can supplement the missing semantics and improve the performance of identifying relations. For example, in Figure 1, there are three entities in this sentence: "JFK", "Obama" and "Harvard". The main task of relation extraction is to identify

† indicates equal contribution. * Corresponding author

978-1-6654-3864-3/21/ $31.00 ©2021 IEEE

the relations of each entity pair. Previous works incorrectly classify the relations of "JFK" and "Harvard" as "Residence" and the "JFK" and "Obama" as "Spouse" due to the missing of contexts. However, we can know that "JFK" and "Harvard" are in the relation of "Graduated at" with the visual concepts "Bachelor cap" and "Gown". Still, the relations of "JFK" and "Obama" can be identified as "Alumni" with the guidance of all the visual objects about "Campus". Therefore, it is obvious to develop a multimodal methods to combine visual information into sentences to extract relations precisely. However, including widely-used sentence-level RE datasets [5], existing relation extraction datasets either concern about the problem of few-shot learning [6] or document-level RE [7]. There is no dataset for training and evaluating relation extraction models, which becomes a barrier in multimodal scenarios. To this end, we introduce MNRE - a challenge **M**ultimodal dataset for **N**eural **R**elation **E**xtraction task in social media posts.

Collecting such datasets is challenging because it requires the understanding of both visual and textual contents. Different from previous relation extraction datasets [6, 7] using distant supervision with human-annotated filtering, labeling relations in social media posts is harder with considerable annotation effort. We utilize a pretrained object detector to extract the visual objects from image posts for a complete and fine-grained understanding of visual contents. Then, some well-educated annotators are asked to tag relations based on entities and also, visual objects. This process improves the qualities of the MNRE dataset.

Another challenge for MNRE task is the alignment of vision and language. Many works have been proposed for multimodal representation and alignment with both datasets and methods, for example, the multimodal sentiment analysis task [8]. However, the alignments of vision and language are built on the simple concatenation of representation vectors, which cannot be utilized well to model the relationship of higher level semantics. For instance, the mapping of a bachelor cap to the concept of graduation is a gigantic gap beyond current works. We envisage a range of well-designed methods and resources for such a challenge that would boost the development of multimodal alignment towards a higher semantic level.

We further propose several multimodal neural relation extraction baselines and conduct experiments on different settings. Our experimental results show that text-based NRE models, even with the help of pretrained language models, still suffer a sharp decline in performance in social media posts (approximately 30% decreases in F1 value). At the same time, our multimodal NRE baseline achieves relatively higher results than previous text-based methods because of the guiding visual information.

Our main contributions can be summarized as follows:

• First, we introduce a new task called multimodal relation extraction and a human-annotated multimodal dataset which aims to systematically test the ability of NRE model in leveraging visual contents to supplement the missing contexts of social media posts, while other relation extraction datasets are only focused on textual contents.

• Second, we propose several multimodal baselines against previous state-of-the-art text-based NRE models and we show a significant improvement due to the merge of visual and textual information.

• Furthermore, we provide an in-depth and thorough analysis for different cases which will be an indicator for future research on the high level multimodal semantic fusion.

2. RELATED WORK

2.1. Relation Extraction in Social Media

Relation Extraction (RE) is the task of extracting semantic relationships from text, which usually occur between two or more entities. Although most neural models have achieved success in the relation extraction task, their models are mainly trained on newswire domain and lack of generalization to other domains. Recently, named entity recognition in social media has raised attention since named entities are a natural way to extract the key information of sentences and a fundamental process to construct a knowledge graph [9, 10].

Compared to NER in social media posts, there are fewer methods [13, 14] dealing with the problem of relation extraction in social media. Considering the short and ambiguous features of texts in social media, relation extraction is a harder task when the NER performance is not promising in such area. We propose a multimodal relation extraction dataset in social media posts to expect a better solution for extracting knowledge from such domain.

2.2. Multimodal Dataset

Constructing multimodal dataset for multimodal tasks is challenging when the understanding of both visual and textual contens is needed. There is much effort of previous works in attempting to construct a multimodal dataset for analyzing real world. Visual Question Answering 2.0 dataset [15] is a collection of 204K images and 1.1M questions. CMU-MOSI [16] is a multimodal sentiment analysis dataset which consists of 2199 opinion video clips. Zhang et al. [17] propose a multimodal named entity recognition dataset with 8724 labelled entities and 4819 images. Although many multimodal datasets are proposed with large labeling efforts like above, there are fewer datasets focusing on text-intensive tasks. We think our dataset can be a useful resource for measuring textual RE methods about generalization ability in social media area.

Dataset	# Img.	# Word	# Sent.	# Ent.	# Rel.	# Inst.	# Fact
SemEval-2010 Task 8 [11]	-	205k	10,717	21,434	9	8,853	8,383
ACE 2003-2004 [12]	-	297k	12,783	46,108	24	16,771	16,536
TACRED [5]	-	1,823k	53,791	152,527	41	21,773	5,976
FewRel [6]	-	1,397k	56,109	72,124	100	70,000	55,803
MNRE	**10,089**	**172k**	**14,796**	**20,178**	**31**	**10,089**	**9,933**

Table 1. Comparison of MNRE with existing sentence-level Relation Extraction datasets (Img.: image, Sent.: sentence, Ent.: entity, Rel.: relation, Inst.: instance, Fact: relational fact).

3. MNRE DATASET

3.1. Dataset Collection

We build the original corpus from three sources: two available multimodal named entity recognition datasets - Twitter15[9] and Twitter17[17], and crawling data from Twitter [1]. The Twitter data contain pairs of textual tweets and their associated images extracted from January to February 2019. Different from previous multimodal NER datasets, we do not pick some certain topics, such as sports or social events, for constructing a variety type of entities and relations. We preliminarily filter out non-English and redundant sentences manually, then we preserve the sentences with more than two entities. If a Twitter post has more than one image, we randomly select one of the images. Finally, we get a candidate set of 20000 instances of crawling Twitter data, 8357 of Twitter15 dataset and 4819 of Twitter17 dataset.

3.2. Twitter Name Tagging

We perform name tagging in collected Twitter data. According to our investigation, most relation extraction datasets identify named entities without considering the entity types. We utilize a pretrained NER tagging tool [2] for extracting both entities and their corresponding types. We think the entity type can help in multimodal relation extraction task since there are correlations between visual objects and textual entities.

3.3. Human Annotation

After tagging named entities, we employ four well-educated annotators to label the relations between entity pairs and filter out the wrong labelled sentences. We develop a tagging tool which simultaneously shows each entity pair in a sentence and the associated image. The annotators are asked to judge whether the relation could be inferred from both the sentence and the image post. Meanwhile, each annotator will give a confidence score that will be utilized to calculate a final relation type considering different view of all the annotators. The Twitter data are randomly assigned to each annotator. For guaranteeing the labeling quality, every instance will be decided by four annotators. We weight the confidence score of each annotator for a same label, and the label with a highest score will be preserved.

3.4. Dataset Statistics

The final dataset contains 31 relation categories with about 10k instances. The average number of words in each sentence is 11.62, which is far less than other relation extraction datasets in newswire domain. The shorter contents of social media texts reveal that extracting information in such domain is challenging due to the lack of contexts. For each sentence, we identify more than one entities in average. Moreover, our dataset provides 10k+ visual images with comparable data size to other text-only relation extraction datasets. Since the relation tags are fully human-annotated, we provide a relatively higher number of relation facts (entity-relation triplets). The statistics comparison of MNRE with existing sentence-level RE datasets is shown in Table 1.

3.5. Case Analysis

Visual information can help to extract entities and their relations. However, previous works [17, 9] only rely on the image regions with attentive maps, which are less-intensive in complicated visual scenarios. Zheng et al. [10] propose to leverage the mapping relations from visual objects to textual entities. We also find the mappings in our MNRE dataset. For example, in the first row of Figure 2, one can know the "Melo" is a name of a person and "nyk" is the name of a basketball team with the guidance of "person" and "jersey" objects. Moreover, we can learn that "Melo" and "nyk" is in the relation of "member of" with the person "wear" this jersey. A more complicated example is in the right of the first line. There are two people wearing the jerseys which is in the same style. The "colleague" relation can be extracted with the two identical jerseys.

We also indicate that some visual relations can be helpful in textual relation extraction. For instance, in the left part of the second row, the relation of the two people "Justin Bieber" and "Selena" is difficult to get only from the sentence. How-

[1] https://archive.org/details/twitterstream

[2] https://allennlp.org/elmo

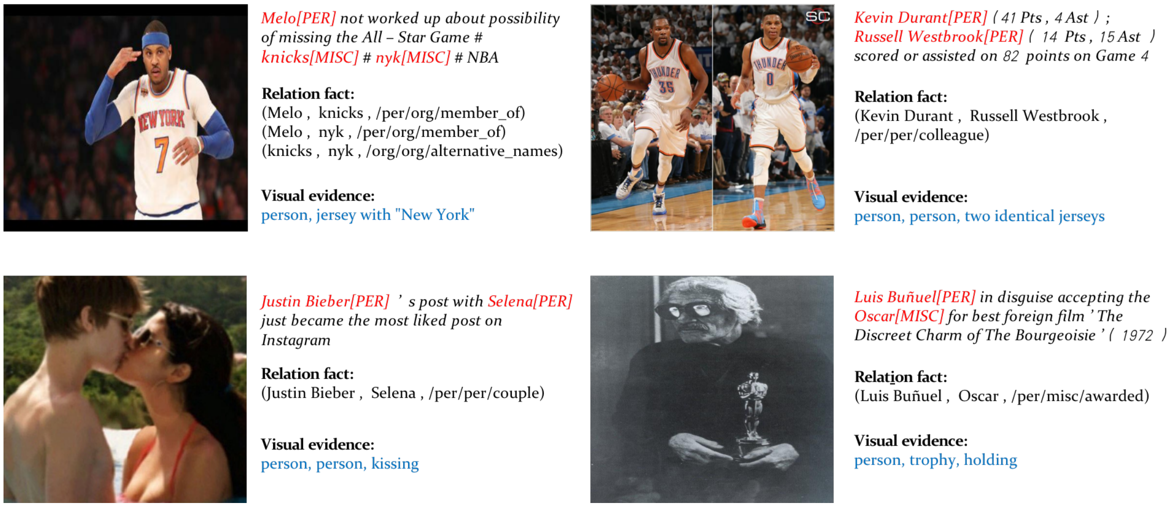

Melo[PER] not worked up about possibility of missing the All – Star Game # knicks[MISC] # nyk[MISC] # NBA
Relation fact:
(Melo , knicks , /per/org/member_of)
(Melo , nyk , /per/org/member_of)
(knicks , nyk , /org/org/alternative_names)
Visual evidence:
person, jersey with "New York"
Kevin Durant[PER] (41 Pts , 4 Ast) ;
Russell Westbrook[PER] (14 Pts , 15 Ast)
scored or assisted on 82 points on Game 4
Relation fact:
(Kevin Durant , Russell Westbrook , /per/per/colleague)
Visual evidence:
person, person, two identical jerseys
Justin Bieber[PER] ' s post with Selena[PER] just became the most liked post on Instagram
Relation fact:
(Justin Bieber , Selena , /per/per/couple)
Visual evidence:
person, person, kissing
Luis Buñuel[PER] in disguise accepting the Oscar[MISC] for best foreign film ' The Discreet Charm of The Bourgeoisie ' (1972)
Relation fact:
(Luis Buñuel , Oscar , /per/misc/awarded)
Visual evidence:
person, trophy, holding

Fig. 2. Four examples for illustrating the effectiveness of visual information in extracting relations. The first line shows that the visual objects and their attributes can help in identifying relations. Further more, we show that the interactions of person-to-person or person-to-object can also provide clues for classifying relations.

ever, we can know that they are couple from the image when they are kissing each other. The same situation can be found in the right part when "Luis Bunue" gains "Oscar" award with one person "is holding" a trophy.

4. EXPERIMENTAL RESULT AND ANALYSIS

4.1. Problem Definition

In this paper, our goal is to predict the relations in a function $F : (e_1, e_2, S, V) \rightarrow Y$. Here, e_1 and e_2 are the pre-extracted named entities. Given a sentence $S = (w_1, w_2, ..., w_n)$, the marked entities e_1 and e_2 and the visual contents $V = (v_1, v_2, ..., v_n)$, we need to classify the corresponding relation tag Y. The textual contents can be words or characters or both. Meanwhile, the visual contents can be image regions or detected objects or their attributes.

4.2. Baselines of Relation Extraction

We compare several relation extraction methods in our MNRE dataset. To explore the influence of incorporating visual information into text-based RE methods, we choose models from three aspects: CNN-based method, pretrained language model based method and distantly supervised method. **Glove+CNN** is a classic CNN-based model for relation extraction. Here we adopt an improved version of this model [18] which concatenates word embeddings with position embeddings. **BertNRE** The pretrained language model Bert has shown its strong generalization in multiple tasks. We adopt the fine-tuning version of Bert which is provided by Soares et al [3]. **Bert+CNN** We also develop a ablation model which leverages Bert representation for each word and a CNN to extract the local features. The model is designed to demonstrate that the image features are more adaptive to CNN-based methods. **PCNN** is a distantly supervised relation extraction model [4]. It is also a CNN-based model which improves the RE performance with a piecewise maxpooling.

4.3. Visual Feature Extraction

We extend several previous relation extraction baselines with visual contents. Here we develop three methods to incorporate visual information:

Image Labels The image label builds a way to bridge the vision and language since the label can be transformed into text embeddings and combined with word representations. The image labels are extracted by a pretrained object detection model. We represent each object label with an embedding layer. The label embeddings are concatenated and transformed into the same dimension as word embeddings for further classification.

Visual Objects Different from using label embeddings, we also explore to incorporate the visual features directly extracted by a pretrained object detector. Compared to only leveraging image labels, visual features may contain more implicit information such as attributes, textures et al. The way we deal with modality fusion is the same as using image labels.

Visual Attention To utilize the correlation of visual objects and texts, we adopt the same bi-linear attention as Zheng et al. [10]. The bi-linear attention layer can exploit the interactions between two modalities of each input channel.

Model	Acc.	Prec.	Rec.	F1
GloVe+CNN	68.56	53.55	45.21	49.03
GloVe+CNN (Lab.)	69.81	55.94	45.03	49.90
Glove+CNN (Obj.)	71.13	58.66	47.65	52.59
GloVe+CNN (Att.)	**72.69**	**62.25**	**46.72**	**53.38**
BertNRE	75.39	61.17	57.03	59.02
BertNRE (Lab.)	75.27	63.86	54.03	58.54
BertNRE (Obj.)	73.40	58.07	62.10	60.01
BertNRE (Att.)	**77.69**	**68.94**	**62.47**	**65.56**
Bert+CNN	74.02	68.49	49.34	57.36
Bert+CNN (Lab.)	72.93	57.05	55.72	56.36
Bert+CNN (Obj.)	74.72	62.50	60.03	61.24
Bert+CNN (Att.)	**76.36**	**65.28**	**61.72**	**63.45**
PCNN	73.49	67.22	45.03	53.93
PCNN (Lab.)	69.89	54.16	50.09	52.04
PCNN (Obj.)	72.15	63.85	45.40	53.07
PCNN (Att.)	72.00	58.85	48.03	52.89

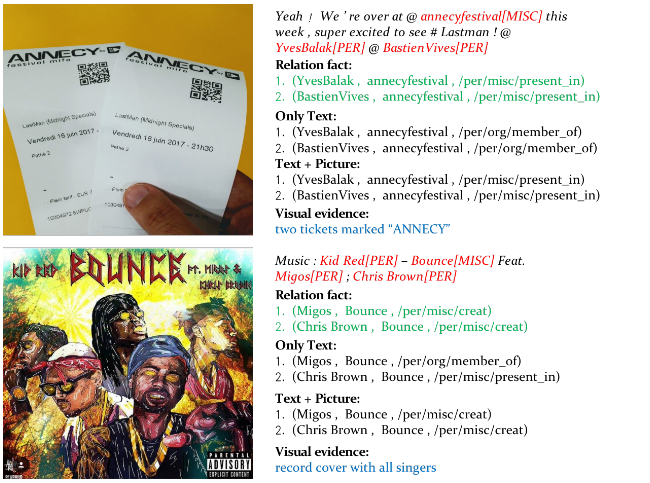
Table 2. The evaluation results (Accuracy, Precision, Recall and F1 value) of all the previous relation extraction baselines on four different settings (Lab.: Image Labels, Obj.: Visual Objects, Att.: Visual Attention).

4.4. General Results

We extend four relation extraction baselines with visual contents and compare their results in our MNRE dataset, which are shown in Table 2. The overall results show that simplely concatenating visual and textual representations can introduce noise for relation classification, especially for the situation that images and sentences are irrelevant. However, when we add the bi-linear attention to gain the correlation of texts and visual objects, the overall performance (including accuracy, precision, recall and F1 value) improves.

The visual contents show a sustainable improvement against the GloVe+CNN model. We achieve about 2 points improvement in accuracy and F1 value, which shows the visual contents can help to extract relations more precisely. BertNRE model gets a significantly higher result than CNN-based methods with fine-grained word level information. When we introduce the attentive module, the model achieves 5 points improvement against the base model with only concatenation of visual features or image labels. It demonstrates that the visual attention can capture the mapping relations of objects and entities. We also develop an ablation experiment that inputting Bert-representations to a CNN feature extractor, and the results show the introduction of visual features or visual attention can boost the NRE performance.

Interestingly, the results from incorporating visual information into distantly supervised models, for example, PCNN, demonstrate that visual contents are counterproductive to the NRE performance. Both accuracy and F1 values slightly decrease around 1 point. We think the piece-wise maxpooling method separates sentences with entities but each part of sentences lacks of correlations to visual objects. As a result, the model cannot infer where the right relations are from.



Yeah ! We ' re over at @ annecyfestival[MISC] this week , super excited to see # Lastman ! @ YvesBalak[PER] @ BastienVives[PER]

Relation fact:
1. (YvesBalak , annecyfestival , /per/misc/present_in)
2. (BastienVives , annecyfestival , /per/misc/present_in)

Only Text:
1. (YvesBalak , annecyfestival , /per/org/member_of)
2. (BastienVives , annecyfestival , /per/org/member_of)

Text + Picture:
1. (YvesBalak , annecyfestival , /per/misc/present_in)
2. (BastienVives , annecyfestival , /per/misc/present_in)

Visual evidence:
two tickets marked “ANNECY”

Music : Kid Red[PER] – Bounce[MISC] Feat. Migos[PER] ; Chris Brown[PER]

Relation fact:
1. (Migos , Bounce , /per/misc/creat)
2. (Chris Brown , Bounce , /per/misc/creat)

Only Text:
1. (Migos , Bounce , /per/org/member_of)
2. (Chris Brown , Bounce , /per/misc/present_in)

Text + Picture:
1. (Migos , Bounce , /per/misc/creat)
2. (Chris Brown , Bounce , /per/misc/creat)

Visual evidence:
record cover with all singers

Fig. 3. Error analysis of multimodal NRE model with compared text-based NRE model.

4.5. Error Analysis

In this section, we provide two cases in Figure 3. which indicates the multimodal RE model can perform better than text-only methods. The first case tells a story that two people are taking part in a festival. However, without the guidance of visual information (two tickets), traditional text-based methods incorrectly identify the relation of “YvesBalak” and “annecyfestival” as “member-of” other than “present-in”. The same situation happens in the second case when “Bounce” is the music record of “Migos, Chris Brown and Kid Red”. Previous text-based methods extract the relation “member-of” and “present-in” without the image information of a music record cover. Instead, the fusion of visual information helps to classify it into “creat” relations with most musicians appearing in the music cover image.

5. CONCLUSION

In this paper, we propose a new and high quality multimodal dataset for relation extraction. The dataset provides a new insight for tackling the problem of contexts lacking in social media posts by leveraging the relative image information. We also compare and analyze several relation extraction baselines with our multimodal extensions. The results demonstrate that previous state-of-the-art relation extraction models suffer performance decline in social media texts and a proper way of incorporating visual information can help to improve the RE performance.

6. ACKNOWLEDGEMENT

This work was supported by the Fundamental Research Funds for the Central Universities, SCUT (No.D2200150, D2201300), the Science and Technology Programs of Guangzhou (No.201704030076, 201902010046), National Natural Science Foundation of China (No. 62076100) and National Key Research and Development Program of China (Standard knowledge graph for epidemic prevention and production recovering intelligent service platform and its applications).

7. REFERENCES

[1] Min Zhang, Jie Zhang, Jian Su, and Guodong Zhou, “A composite kernel to extract relations between entities with both flat and structured features,” in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, 2006, pp. 825–832.

[2] Matthew R Gormley, Mo Yu, and Mark Dredze, “Improved relation extraction with feature-rich compositional embedding models,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1774–1784.

[3] Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski, “Matching the blanks: Distributional similarity for relation learning,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 2895–2905.

[4] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao, “Distant supervision for relation extraction via piecewise convolutional neural networks,” in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 1753–1762.

[5] Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D Manning, “Position-aware attention and supervised data improve slot filling,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 35–45.

[6] Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Yuan Yao, Zhiyuan Liu, and Maosong Sun, “Fewrel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation,” in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4803–4809.

[7] Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun, “Docred: A large-scale document-level relation extraction dataset,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 764–777.

[8] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, “Tensor fusion network for multimodal sentiment analysis,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1103–1114.

[9] Di Lu, Leonardo Neves, Vitor Carvalho, Ning Zhang, and Heng Ji, “Visual attention model for name tagging in multimodal social media,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1990–1999.

[10] Changmeng Zheng, Zhiwei Wu, Tao Wang, Cai Yi, and Qing Li, “Object-aware multimodal named entity recognition in social media posts with adversarial learning,” *IEEE Transactions on Multimedia*, 2020.

[11] Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz, “Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals,” *SEW-2009 Semantic Evaluations: Recent Achievements and Future Directions*, p. 94, 2009.

[12] George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel, “The automatic content extraction (ace) program-tasks, data, and evaluation.,” in *Lrec*. Lisbon, 2004, vol. 2, pp. 837–840.

[13] Zuoguo Liu and Xiaorong Chen, “Research on relation extraction of named entity on social media in smart cities,” *Soft Computing*, pp. 1–13, 2020.

[14] Gregory Brown, “An error analysis of relation extraction in social media documents,” in *Proceedings of the ACL 2011 Student Session*, 2011, pp. 64–68.

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, “Making the v in vqa matter: Elevating the role of image understanding in visual question answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6904–6913.

[16] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency, “Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.

[17] Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang, “Adaptive co-attention network for named entity recognition in tweets.,” in *AAAI*, 2018, pp. 5674–5681.

[18] Thien Huu Nguyen and Ralph Grishman, “Relation extraction: Perspective from convolutional neural networks,” in *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*, 2015, pp. 39–48.