News Image-Text Matching With News Knowledge Graph

ZHAO YUMENG, YUN JING, GAO SHUO, AND LIU LIMIN
College of Data Science and Application, Inner Mongolia University of Technology, Hohhot 010080, China
Inner Mongolia Autonomous Region Engineering and Technology Research Center of Big Data-Based Software Service, Hohhot 010080, China
Corresponding author: Yun Jing (yunjing_zoe163.com)
This work was supported in part by the National Natural Science Foundation of China under Grants 62062055, in part by the Natural Science Foundation of Inner Mongolia under Grant 2020MS06025, and in part by the Department of Science and Technology of Inner Mongolia under Grant 2019GG372.

ABSTRACT Image-text matching using the image caption method has made a great progress. However, there are many named entities in news text, and existing approaches are unable to directly generate named entities in the news image caption. It leads to a semantic gap between text and news image caption. Moreover, the existing methods lack the analysis of indirect relations between named entities. Therefore those approaches easily leads to relations error when generating news image caption. To generate the news image caption with named entities by analyzing the indirect relations between named entities. We propose a novel model. In details, we propose the TopNews dataset with related news, which aims to construct the relations between named entities as widely as possible. Then we develop the news knowledge graph by extracting named entities from TopNews dataset. Furthermore, we propose News Knowledge Driven Graph Neural Network (NKD-GNN). We utilize NKD-GNN to analyzing the whole relations of entities in news knowledge graph. In this way, we generate the news image caption with named entities. The results of extensive experiments based on TopNews dataset and common dataset demonstrate that our approach is effective in detecting the consistency of news images and text.

INDEX TERMS Image-text matching, named entity, indirect relations of named entities.

I. INTRODUCTION
Detecting the consistency of image and text by using the image caption method has attracted increasing attention in recent years [1]–[5]. Nevertheless, due to the semantic gap between news text and news image, calculating the consistency between news text and news image is still a challenging problem. Recently, some works have tried to overcome the semantic gap of news images and text. For example, in [6]–[8], these methods generate the news image caption with named entities. These methods first generate a template caption with placeholders for named entities. Then these methods connect the named entities of news text into a graph. Afterwards the best candidate for each placeholder is chosen via analyze the direct relation between named entities such as the co-occurrence rate of adjacent entities in the graph. Those approaches have made significant improvements in image-text matching. However, these methods are limited to ignoring the indirect relations between named entities, which easily leads to relations error when choosing entity candidates. As shown in Figure 1, the content of news is that Emmanuel Macron is speaking in the US Congress. Connect the named entities of news text into a graph according to the connection rules of the DBpedia. Adjacent entities have direct relations in the graph, otherwise entities have indirect relations. Then we get the caption ‘Trump delivers a speech from US Congress’ by selecting the entities with the highest co-occurrence rate of directly connected entities. We can find that the caption is wrong. This error has two reasons. On the one hand, these methods only consider the direct relations between adjacent named entities in the graph. Although trump and US Congress have the highest co-occurrence rate, they cannot accurately reflect news events. The fact of news is about Macron and US Congress, these have the indirectly relations. These methods perform poorly because these ignore the indirectly relations between named entities. We come to a conclusion, detect the consistency of news image-text needs to analyze the indirect
relations between named entities in news text. On the other hand, existing general knowledge graph such as DBpedia ConceptNet cannot accurately calculate the indirect relations between named entities in news stories. To analyze the whole relation of named entities, we need to calculate the relations of named entities accurately in news stories. For instance, Macron and US Congress do not have the directly relations in general knowledge graph. But in the news scene they often appear together. The general knowledge graph cannot reflect the relations of named entities in news story accurately. In the experiments section, we discovery that constructing the news knowledge graph is the basis for analyzing the whole relations between named entities.

We aim to generate news image caption with named entity by analyzing the whole relations of named entities. This introduces new challenges. On one hand, We need to accurately construct relations between named entities in news stories. On the other hand, We need to analyze the whole relations of entities in the news knowledge graph.

In this work, we propose a novel model to deal with the aforementioned challenges. Firstly, we propose the TopNews dataset which is about news image and text matching. To build relations between named entities in the broadest context possible, Topnews dataset includes not only news text, but also news related articles. Then, we extract the named entities from TopNews dataset and construct them to news knowledge graph. However, the news knowledge graph is non-Euclidean data. It is a challenging problem to fully analyze the relations of the entities in the news knowledge graph. Moreover, the news knowledge graph has some redundant entities that affect the analysis of named entity relations. We need avoid redundant entity feature vectors affecting. For this reason, we propose the NKD-GNN, which analyzes the whole relations between the named entities in news knowledge graph. We utilize NKD-GNN to select which entities relate to news images. In this way, we accurately constructed the relations between named entities and completely analyzing the relations between named entities.

The primary contributions of the proposed model are summarized as follows:

- We propose the TopNews dataset, which contains news text and images, additionally includes news related articles, and their publication dates.
- We develop the news knowledge graph by connecting named entities in the TopNews datasets. News knowledge graph precisely demonstrates the relations of names entities in a news story. Then, we propose the NKD-GNN that aims to analyze the whole relations between the named entities, and choose which named entities are related to news images.
- We conduct extensive experiments on the TopNews dataset. The result shows that our method can effectively detect the consistency between news image and text and our model outperforms the image-text matching model for news.

II. RELATED WORK

Our model is about image-text matching and news image caption. Hence, we will introduce related work in these two aspects.

A. IMAGE-TEXT MATCHING

Detecting the consistency of news images and text, is an application of image-text matching task. Image-text matching approaches capture visual-textual feature. Afterwards, these approaches, calculate the matching of visual and textual feature. In [9] the image feature and text feature are extracted by CNN and Skip-Gram Language Model, respectively. Ranking loss is then implemented for similarity learning. On the other hand, in [10], it encodes text by RNN and design a hinge-based triplet ranking loss to train the model. Recent successes of attention models for visual-textual learning tasks, such as for image captioning [11]–[14], motivate researchers to solve image-text consistency in the level of image regions and words. In [15], it incorporates a multimodal context-modulated attention scheme that can selec-
tively attend to a pair of instances of the image and sentence at each time step. In [16], it proposes Dual Attention Networks that attend to both specific words in the text and regions in images through multiple steps. Because of the restriction of CNN, each image is typically divided into a fixed number of regions (e.g., $7 \times 7$) of the same shape and size. This prevents models from consistency of words and small image objects more accurately. In [17], it detects the matching of news images and text by generating news image caption. In this progress, it uses the Conceptnet knowledge graph to explain the types of objects in news image, such as whether the chair in the image is a swivel chair or a toffee chair. This method improves the effect of news image-text matching detection. However, this method uses commonsense knowledge graphs to reason about the information of objects in the image, which leads to image caption errors. Therefore, we construct a news knowledge graph that reflects the knowledge of news scenes. Exploring the graph structure of vision-language tasks is a common solution. Despite the remarkable progress, those methods still suffer from cannot detect accurately the consistency of news images and text, when the semantic differences between the images and text are too huge. However, these methods have a shortcoming, when there are some specific named entities in the text. Existing approaches are unable to directly detect the consistency of news images and text. Therefore, our model aims to overcome the semantic gap by generating the news image caption with named entities.

### B. NEWS IMAGE CAPTION

News image captioning includes the article text as input and focuses on the types of images used in news articles. A key challenge here is to generate correct entity names. Existing approaches include extractive methods that use n-gram models to combine existing phrases [18] or simply retrieving the most representative sentences [19] in the article. Ramise et al. [8], [20] built an end-to-end LSTM decoder that takes both the article and image as inputs, but the model was still unable to produce names that were not seen during training. In [21], this method introduces the task where the machine generates captions without extra sentences about the novel object. In details, this method generates a sentence containing placeholders. The placeholder will then be filled with the correct word, resulting in a caption with novel object captions. However, we not only need to know the information of novel objects in the news image, but also need to know the fine-grained information in the news image, such as named entities. Due to the variety of named entity types in news images, the templates of news image captions are also more diverse. In [22]–[25], these methods connect the objects in the image as a graph. Afterwards the interpretative and reasoning image caption is generated by learning the semantic relations features between objects. However, the semantic differences between the images and text are too huge. These methods only depend on the semantic relationship between the objects in the image, the semantic difference between news image and text is difficult to overcome. Hence we need to utilize the external knowledge graph to explain the news image. To overcome the limitation of a fixed-size vocabulary, template-based methods have been proposed. An LSTM first generates a template sentence with placeholders for named entities, e.g., “PERSON speaks at BUILDING in DATE” [7], [20]. Afterwards the best candidate for each placeholder is chosen via a knowledge graph of entity combinations [6], or via sentence similarity [7], [20]. Those approaches have made significant improvement in the news image caption. However, these methods are limited to ignoring the indirect relations between named entities, which easily leads to relations error when filling entities. Our method proposes the NKD-GNN to analyze the whole relations between name entities in news knowledge graph. Thus, we are able to incorporate the best candidate into news image captions.

### III. MODEL

We aim to detect the consistency of news images and text by generating the news image caption with named entities. As illustrate in Figure 2 our model consists of four stages. Firstly, to accurately react entities relations in news stories, we develop a news knowledge graph. Secondly, we generate the news image template caption which consists of the objects of news images, and placeholders. Thirdly, we propose NKD-GNN that analyzes the whole relations and infer the indirectly relations between the named entities in the news knowledge graph. Then, NKD-GNN select the best candidate for each placeholder in the news image template caption. Finally, we use HCAN (Hybrid Co-Attention Network) [20] to detecting the consistency of the news image caption with named entities and news text.

### A. NEWS TEMPLATE CAPTION GENERATION

We generate news image template captions in which named entities placeholders are indicated along with their respective tags. In this process we follow the state-of-the-art image captioning model [27]. This paper follows the encoder-decoder model with a CNN encoder and LSTM decoder. At the encoding stage, we encode the image into a vector by using a pre-trained deep CNNs [28] which follows the state-of-the-art CNN with 19 convolutional layers based on the ImageNet dataset [29] and use the last fully-connected layer as the output of the encoding. At the decoding stage, we employ the language model based on Top-Down attention Long Short Term Memory (LSTM) [30] to decode image into news image captions. The news image caption generated in this way lacks named entities and has a semantic gap with the news text. In order to integrate the named entity into the news caption. We generate the news image template caption by using the WordNet [31] tool. In this work, we define slots as placeholders for words in the news image caption with the same types. We only generate placeholders for four types of words, such as ‘Person’ ‘Organization’ ‘Place’ ‘Building’. In detail, all words in the ‘Person’ semantic tree are replaced with <Person> placeholders, and words in the ‘Place’
semantic tree are replaced with <Place>. The <Organization> placeholder can be used to replace ‘a group of Sth’, then ‘building’ can be replaced by the <Building> placeholder. In later stages we will choose the best candidate for each placeholder in the news image template caption.

B. BUILD NEWS KNOWLEDGE GRAPH

We aim to analysis the whole relation of named entity in news story. Above all we need to construct the relations between named entities accurately. The use of knowledge graphs to build relations between entities is a common method. However, current knowledge graphs whose entities relations were calculated by the common sense database, such as, DBpedia, ConceptNet. Thus, these are unable to accurately reflect the relations between named entities found in the news story. We need to build a news knowledge graph. To construct the news knowledge graph, we need to extract named entities from the news story. Then, we need to define the weights of relations between these entities.

Firstly, we use SpaCy’s named entity recognizer [30] to extract the ‘PERSON’, ‘ORGANIZATION’, ‘PLACE’ and ‘BUILDING’ named entities in TopNews datasets. Then we construct these named entities to $V = \{v_1, v_2, \ldots, v_m\}$. $V$ is the set consisting of named entities involved in the TopNews datasets. The entities appearing together in a sentence are linked edges. Let $E = \{e_1, e_2, \ldots, e_m\}$ denotes the set consisting of all edges between named entities. We compute the edge weights between two named entities as is follows:

$$H_e = \frac{f_{v_h v_t}}{\max(f_{v_h}, f_{v_t})}$$

where $e \in E$, $H_e$ is the value of edge $e$. $e$ is an edge which connects two named entities in news knowledge graph. $v_h$ and $v_t$ are two named entities connected by the edge $e$. $f_{v_h v_t}$ is the co-occurrence frequency of two named entities in TopNews datasets, $f_{v_h}$ and $f_{v_t}$ are the individual frequencies of $v_h$ and $v_t$ respectively. For example, in Figure 4, Macron (Person) and Pairs (Place) co-occur frequently, therefore the edge between them has a larger weight.

Two named entities and the weight of the connection between them construct the three tuple $\langle v_h, e_i, v_t \rangle$. We construct news knowledge graph $G = (V, E)$ by composing set $V$ and set $E$. In news knowledge graph each node represents a named entity, each edge represents the weight of relation between two named entities. Moreover, each news has individual subgraph $g_s \in G$ that is build by named entities and those connections in the news related articles.

C. NEWS KNOWLEDGE DRIVEN GRAPH NEURAL NETWORK

To analyze the whole relations and infer the indirect relations between named entities in news knowledge graph, we propose NKD-GNN, and its structure is shown in Figure 3:

As illustrated in Figure 3, the NKD-GNN consists of four stages. In the first stage, in order to analyze the whole relations between the named entities, we utilize the graph neural network. In this process, each node need aggregates all edges and node information in the news knowledge graph. Then we obtain the vectors of entities. However, there exit many redundant entity feature vectors in news knowledge affecting the output of the model. Therefore, we use the attention mechanism to assign the weight to each entity vector, thereby
we calculate the news knowledge graph global representation vector \( N_g \). Furthermore, in news knowledge graph, the entity having most edges is the key entity, since the key entity reflects the focus of several related news articles. Thus, we compute the news knowledge graph representation vector \( N_g \) by taking linear transformation over the concatenation of the key entity vector \( N_b \) and global representation vector \( N_g \). After obtaining the news knowledge graph representation vector, we compute the probability for each entity by multiplying its vector \( v_i \) by news knowledge graph representation vector \( N_g \). Now, we describe our method in detail.

1) LEARNING ENTITY VECTOR ON NEWS KNOWLEDGE GRAPH

During the analysis of the whole relations between the named entities in news knowledge graph, each node needs to learn information from the other nodes. Because graph neural network can automatically extract features of news knowledge graphs with considerations of rich node connections. Thus, we use graph neural network to aggregate all edges and node information in the knowledge graph. In this process, each node constantly updates the information by aggregating information from other nodes. The goal of the graph operation is to let the nodes of news knowledge graph learn the relations between entities in the news scene. We first demonstrate the learning process of node vectors in news knowledge graph. Formally, for the node \( v_i \) of graph \( G_s \), the update functions are given as follows:

\[
\begin{align*}
  a_i^t &= A_{NKG,i} \left[ V_i^{t-1}, \ldots, V_n^{t-1} \right]^T H \\
  Z_i^t &= \sigma(W_c a_i^t + U_c V_i^{t-1}) \\
  r_i^t &= \sigma(W_r a_i^t + U_r V_i^{t-1}) \\
  \tilde{V}_i^t &= \tanh(W_o a_i^t + U_o (r_i^t \odot V_i^{t-1})) \\
  V_i^t &= (1 - Z_i^t \odot V_i^{t-1}) \odot \tilde{V}_i^t + Z_i^t \odot V_i^t 
\end{align*}
\]

where \( a_i^t \) is the updating gate, the weight matrix \( H \in \mathbb{R}^{d \times d} \), \( a_i^t \in \mathbb{R}^{d \times d} \) is the set of node vectors at time \( t - 1 \), \( A_{NKG} \in \mathbb{R}^{n \times n} \) is the adjacency matrix of the news knowledge graph, \( A_{NKG,i} \in \mathbb{R}^{1 \times n} \) is the block matrix corresponding to the th node, \( z_i^t \) is the reset gate, \( r_i^t \) is the update gate, \( \sigma(\cdot) \) is the sigmoid function and \( \odot \) is element-wise multiplication operator. The adjacency matrix of the news knowledge graph is shown in Figure 4, when \( i = 2 \), \( A_{NKG,2} = [0.67, 0, 0.33, 0.5, 0] \).

Formula (2) reflects the process of node \( v_i \) aggregating the information of its adjacent nodes in the news knowledge graph, and \( a_i^t \) aggregate the information from the neighbor nodes of \( q \in \mathbb{N}^d \). Formulas (3) and (4) respectively determine the information retained and discarded in the neighbor nodes. Formula (5) uses the vector \( V_i^{t-1} \) and the input \( a_i^t \) of node \( v_i \) at time \( t \) to obtain the candidate vector \( \tilde{v}_i^t \) of node. Formula (6) uses the vector \( V_i^{t-1} \) at \( t - 1 \) time and the node candidate vector \( \tilde{v}_i^t \) to calculate the node \( v_i \) representation vector \( v_i^t \).

2) NEWS KNOWLEDGE GRAPH EMBEDDINGS GENERATION

The nodes of the general knowledge graph are concepts in the common sense datasets. Unlike general knowledge graphs, in news knowledge graph the node is the entity of the TopNews, further the entities with the most edges is the key entity. Since, news knowledge graph was constructed by named entities of several news related articles, the key entity reflects the focus of several news related articles. Therefore, calculating the represent a vector of news knowledge graph,
consider not only the news knowledge graph global representation vector but also the key entity vector.

After feeding news knowledge graph into the graph neural networks, we obtain the vectors of all nodes. To represent the news knowledge graph as an embedding vector \( N_g \in \mathbb{R}^d \), we first consider the key entity vector \( N_b \). Then, we consider the news knowledge graph global representation vector \( N_g \) by aggregating all node vectors. However, entity vectors in the news knowledge graph global representation vector may have different levels of priority. We further adopt the soft-attention mechanism to better represent the news knowledge graph global representation:

\[
\alpha_i = q^T \sigma(W_1 v_i + W_2 v_j) \quad (7)
\]

\[
N_g = \sum_{i=1}^{n} \alpha_i V_i \quad (8)
\]

Then, we compute the news knowledge graph representation vector \( N_r \) by taking linear transformation over the concatenation of the key entity vector \( N_b \) and global representation vector \( N_g \). News knowledge graph representation \( N_r \) can be defined as:

\[
N_r = W_3 [N_g; N_b] \quad (9)
\]

where \( \alpha_i \) is the node coefficient, and the matrix parameter \( q \in \mathbb{R}^d \), \( W_1 \in \mathbb{R}^{d \times d} \) matrix and \( W_2 \in \mathbb{R}^{d \times d} \) matrix are the node weight matrix. \( N_g \) and \( N_b \) are compressed by \( W_3 \in \mathbb{R}^{d \times 2d} \) into the \( \mathbb{R}^d \) vector space.

**D. PREDICTION AND MODEL TRAINING**

After obtaining the news knowledge graph representation vector, we compute the score \( \hat{z} \) for each entity by multiplying its vector \( v_i \) and news knowledge graph representation vector \( N_r \). The functions are given as follows:

\[
\hat{Z}_i = N_r^T v_i \quad (10)
\]

Then we apply a softmax function to get the output vector of the model \( \hat{y} \).

\[
\hat{y} = \text{softmax}(\hat{Z}) \quad (11)
\]

where \( \hat{z} \) is the node’s score, \( \hat{y} \) denotes the probabilities of the entity filling in the placeholder of news image template caption. Subsequently, we choose the entity with the largest probability to fill the placeholder. As a result of this step, given the template caption “<PERSON> in a suite and tie standing in <PLACES>”, we obtain the ’Macron in a suite and tie standing in the US Congress’. For those placeholders that cannot be filled with named entities, we use general words to replace them, for example, using the word “Person” to replace the placeholder <PERSON>. For each news knowledge graph, the loss function is defined as the cross-entropy [32]. It can be written as follows:

\[
\mathcal{L}(\hat{y}) = - \sum_{i=1}^{m} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \quad (12)
\]

where \( y \) denotes the one-hot encoding vector of the most relevant entity in the news knowledge graph.

**E. COMPUTATIONAL CONSISTENCY METHOD**

In this work, we default the news image and text consistency when the news image caption with named entities matches a single sentence of the news text. Calculating whether an individual sentence match the news image caption with named entities is regarded as a text matching task. Thus, we follow the HCAN [26] model. It consists of three major components: firstly, a hybrid encoder module that explores three types of encoders: deep, wide, and contextual; secondly, a relevance consistency module with external weights for learning term consistency signals; finally, a semantic consistency module with co-attention mechanisms for context-aware representation learning. Since HCAN considers not only the semantic matching of two sentence, but also their relevance matching. Thus, it overcomes the sentence structure difference between the news image caption with named entities and news text.

**IV. EXPERIMENTS**

**A. DATASETS**

1) TopNews DATASETS

Previous news datasets do not include the consistency of news image and text. Thus, we propose the TopNews dataset. TopNews dataset contains 9000 news items from the Global Times for the period 01 December 2019 to 01 June 2020. It includes 1800 social, political, life, science and technology, and sports news respectively, randomly split into 5903 for training, 1686 for validation and 843 for testing. All of the news are image-text consistency, vote by expert group which is composed of 3 people. In order to build relations between named entities in the broadest context possible, Topnews dataset includes not only news text, but also news related articles. On average each news text has two related articles. In order to verify the performance of our model in image-text mismatching news. We randomly select 1000 news in the TopNews dataset, and randomly combined this text and images. Then we obtain 1000 image-text mismatching news and insert these into the test. Table 1 shows the data type of TopNews dataset.

| TABLE 1. Statistics of TopNews dataset. | Train | Validate | Test |
|-----------------------------------------|-------|----------|------|
| Number                                  | 5903  | 1686     | 1843 |
| Number of entity                        | 9189  | 2435     | 2729 |

Figure 5 shows the number of named entities in the different types of news. Generally the news text describes the details of the event. Thus, there are many named entities in news text such as famous persons, organization, building, place. Since political society and life news generally contain more events, these news has more named entities. However, science news reporting on novel technology and sports news is inclined to event introduction and reporting. Thus, there are few named entities in science news and sport news.
2) GoodNews DATASETS
“GoodNews [7]”, the news image captioning dataset in the literature. GoodNews dataset uses the New York Times API to retrieve the URLs of news articles ranging from 2010 to 2018. In total, GoodNews dataset has 466,000 images with captions, headlines and text articles, randomly split into 424,000 for training, 18,000 for validation and 23,000 for testing. In our work, we use the Goodnews dataset to verify the effectiveness of our method. We consider that the news of the dataset are image-text matching.

B. IMPLEMENTATION DETAILS AND METRICS
In this paper, we generate news knowledge graphs for every news. Then, we denote the one-hot encoding vector of most relevant entity in the news knowledge graph. There are 9200 named entities in the news related articles. Therefore, we set the dimensions of news knowledge graph representation vector to \( d = 9300 \). We use Gaussian distribution initialization parameters [33] that obey mean of 0 and a standard deviation of 0.1, and we set L2 penalty to \( 10^{-5} \). The mini-batch Adam optimizer is exerted to optimize these parameters. The initial learning rate was set to 0.01. The batch size was set to 100, and the learning rate was set to decay by 0.1 for every 3 epochs.

We use accuracy rate, precision rate, recall rate, F1-measure to evaluate the effectiveness of the proposed method.

C. EVALUATION OF NEWS KNOWLEDGE GRAPHS
In this section, to evaluate the effect of news knowledge graphs. We compare the news knowledge graph with DBpedia [34] and ConceptNet [35]. In detail, firstly, we generate the news image template caption by using the method of section 3.1. Secondly, we extract named entities from news related articles. Then we link these entities to graph according to the connection between the concepts in the ConceptNet and DBpedia respectively. Subsequently, in order to generate the news image caption with named entities. We utilize the pre-trained NKD-GNN choosing the best consistency named entities in graph to fill the placeholder of news image template caption. Finally, we use HCAN to detect the consistency of the news image caption with named entities and news text by using HCAN method.

| Method         | Accuracy | Precision | Recall | F1 Measure |
|----------------|----------|-----------|--------|------------|
| ConceptNet     | 68.5%    | 67.9%     | 68.3%  | 68.2%      |
| DBpedia        | 71.4%    | 70.8%     | 71.2%  | 71.0%      |
| News Knowledge graph | 81.5%    | 80.4%     | 80.7%  | 80.5%      |

Table 2 shows the results on two different knowledge graph in detecting the consistency of news images and text, then compare them with our method. According to the experiments, it is obvious that the proposed news knowledge graph achieves the best performance on TopNews dataset in terms of Accuracy, Precision, Recall, and F1 Measure. This evaluate the effectiveness of our proposed method. Regarding others knowledge graph, their performance is relatively poor. Since ConceptNet and DBpedia are large-scale knowledge graph that use large-scale commonsense dataset to construct the entities and these relations. Therefore, these knowledge graphs are not suitable for constructing relations between entities in news stories. On the contrary, in news knowledge graph we define the relation weight between named entities by counting these co-occurrence rates in a news story. Thus, our method improves the accuracy of news image and text consistency.

D. ABLATION EXPERIMENT OF NKD-GNN
In this section, to verify each step in the NKD-GNN method, we compare the news knowledge graph embedding strategy with the following three approaches: 1.GNN-Avg: This method only extracts the global representation vector of news knowledge graph with average pooling. In detail, it uses GNN [37] to learn each node vector in the news knowledge graph and the average sum of the node vectors to obtain the representation vector of the news knowledge graph. 2.GNN-AvgKey: This method fuses the key entity vector and the global representation vector of news knowledge graph. In details, it calculates the average sum of the node vectors to obtain the global representation vector of the news knowledge graph. Then, it takes linear transformation over these. 3.GNN-Att: This method uses GNN [36] to learn each entities vector in the news knowledge graph. It uses the attention mechanism to obtain the global representation vector of the news knowledge graph by aggregating node vectors.

The representation vectors of the news knowledge graphs are calculated by using the three methods and NKD-GNN. The results of methods with three different embedding strategies are given in Table 4.

The experimental results summarized in Table 4 show that the best results were obtained by using the NKD-GNN method. The results demonstrate that NKD-GNN can accu-
TABLE 3. The results of the comparative experiments.

| Visual Only | Methods          | TopNews       |          | GoodNews       |          |
|-------------|------------------|---------------|----------|----------------|----------|
|             | Accuracy | Precision | Recall  | F1-Measure | Accuracy | Precision | Recall  | F1-Measure |
| DAN [16]    | 61.7%     | 67.2%      | 61.4%   | 61.3%     | 65.1%     | 64.4%     | 64.6%   | 64.7%     |
| SCAN [10]   | 69.7%     | 69.2%      | 69.4%   | 69.5%     | 73.0%     | 72.5%     | 72.9%   | 72.8%     |
| VSRN [40]   | 71.4%     | 70.3%      | 70.4%   | 71.2%     | 74.3%     | 74.2%     | 73.8%   | 74.0%     |
| RRTC [41]   | 70.3%     | 69.7%      | 70.3%   | 70.1%     | 73.7%     | 73.4%     | 73.7%   | 73.5%     |
| Vns cree- VL [42] | 71.6%     | 71.1%      | 71.2%   | 71.3%     | 75.2%     | 74.7%     | 74.8%   | 75.1%     |
| Trip [43]   | 72.5%     | 71.9%      | 72.3%   | 72.4%     | 76.1%     | 75.5%     | 75.9%   | 76.3%     |

TABLE 4. The results of the ablation experiments.

| Method       | Accuracy | Precision | Recall | F1-Measure |
|--------------|----------|-----------|--------|------------|
| GNN-AVG      | 73.9%    | 73.4%     | 73.7%  | 73.5%      |
| GNN-ATT      | 74.2%    | 74.1%     | 74.0%  | 73.9%      |
| GNN-AVGKEY   | 74.1%    | 73.6%     | 73.9%  | 73.8%      |
| NKD-GNN      | 81.3%    | 80.4%     | 80.7%  | 80.5%      |

TABLE 5. The result of different types of news on our method.

| Method       | Accuracy | Precision | Recall | F1-Measure |
|--------------|----------|-----------|--------|------------|
| Society News | 85.7%    | 85.2%     | 85.3%  | 85.5%      |
| Politic News | 86.6%    | 86.1%     | 86.4%  | 86.3%      |
| Life News    | 84.4%    | 83.6%     | 83.9%  | 84.2%      |
| Sci News     | 74.3%    | 73.9%     | 74.1%  | 74.2%      |
| Sport News   | 78.4%    | 77.8%     | 78.2%  | 78.0%      |

2) COMPARISON WITH BASELINE METHODS

To demonstrate the overall performance of the proposed model, we compare our model with the baselines. The baseline methods are as follows:

1. DAN [16]: Firstly, this method splits image into some small areas and split text into words. Then, it uses attention net that can automatically match the word of text and image areas. 2. SCAN [39]: This method uses the attention mechanism to learn text and image representations. Then, it maps two features to the same vector space, and uses the cosine distance to measure the similarities between features in the text and images. 3. VSRN [40]: This method extracts images and text feature, respectively, then compare their features. In extracting text feature stage, it uses Gated Recurrent Unit to generate the feature of text. In extracting image feature stage, it uses the bottom-up attention method to extract the key objects in the image. Then it constructs relations between objects and uses GCN (Graph convolutional network) reasoning the relation between objects, in this way it generates image features. Finally, it compares image features with text features to detecting the consistency of image and text. 4. RRTC [41]: In this method, the region reinforcement network is built to infer fine-grained correspondence by considering the relationships of regions and re-assigning region-word similarities. Meanwhile, the topic constraint module is presented to summarize rately calculate news knowledge graph representation vectors. It is shown that calculating the represent a vector of news knowledge graph, consider not only the news knowledge graph global representation vector but also the key entity vector. Please note that GNN-AVGKEY, a downgraded version of NKD-GNN, still outperforms GNN-AVG and achieves almost the same performance as that of GNN-ATT. We discover that the key entity is crucial for news knowledge graph embedding. Thus, the key entity generally is the focus of several news. Furthermore, the tables show that GNN-ATT performs better than GNN-AVG with average pooling on TopNews dataset. It indicates that the news knowledge graph may contain some redundant entity vectors. Besides, it is shown that attention mechanisms are helpful in calculating the global representation vector of the news knowledge graph.

E. EVALUATION OF NEWS IMAGE–NEWS TEXT CONSISTENCY DETECTION

1) THE IMPACT OF DIFFERENT TYPES OF NEWS ON OUR METHOD

To verify the generalization ability of the new method, in this section we detect the consistency of image and text in different types of news. The results of the experiment are shown in Table 5.

As shown in Table 5, our method achieved better results when detecting the consistency of image and text in politic, social and life news. Since these news involves a large number of named entities people, organizations and places. More importantly, the focus of politic, social and life news are generally described around core entities. However, there are many advanced scientific equipment in the image of science news. Consequently, current object recognition methods are unable to accurately identify objects in the images of science news. Thus we cannot generate image caption for science news image accurate. Furthermore, the most sports news focuses on analyzing events and introducing games, this involves less named entities and less relations. So regarding sports and science news, the performance of our model is relatively poor.

2) COMPARISON WITH BASELINE METHODS

To demonstrate the overall performance of the proposed model, we compare our model with the baselines. The baseline methods are as follows:

1. DAN [16]: Firstly, this method splits image into some small areas and split text into words. Then, it uses attention net that can automatically match the word of text and image areas. 2. SCAN [39]: This method uses the attention mechanism to learn text and image representations. Then, it maps two features to the same vector space, and uses the cosine distance to measure the similarities between features in the text and images. 3. VSRN [40]: This method extracts images and text feature, respectively, then compare their features. In extracting text feature stage, it uses Gated Recurrent Unit to generate the feature of text. In extracting image feature stage, it uses the bottom-up attention method to extract the key objects in the image. Then it constructs relations between objects and uses GCN (Graph convolutional network) reasoning the relation between objects, in this way it generates image features. Finally, it compares image features with text features to detecting the consistency of image and text. 4. RRTC [41]: In this method, the region reinforcement network is built to infer fine-grained correspondence by considering the relationships of regions and re-assigning region-word similarities. Meanwhile, the topic constraint module is presented to summarize
the central theme of images, which constrains the original image deviation. 5:Unicoder-VL [42], a universal encoder that aims to learn joint representations of vision and language in a pre-training manner. This task tries to predict whether an image and a text describe each other. 6:Trip [43], a novel within-modality losses which encourage semantic coherency in both the text and image subspaces, which does not necessarily align with visual coherency. Our method ensures that not only are paired images and texts close, but the expected image-image and text-text relationships are also observed. Our approach improves the results of cross-modal retrieval on four datasets compared to five baselines. 7:Transformer + RoBERTa [44], an end-to-end model for news image captioning with a novel combination of sequence-to-sequence neural networks, language representation learning, and vision subsystems. In particular, it address the knowledge gap by computing multi-head attention on the words in the article, along with faces and objects that are extracted from the image. It address the linguistic gap with a flexible byte-pair-encoding that can generate unseen words. 8:image transformer [45], this method which consists of a modified encoding transformer and an implicit decoding transformer, motivated by the relative spatial relationship between image regions. This design widens the original transformer layer’s inner architecture to adapt to the structure of images. 9:ETA-Transformer [46], a novel model that enables the Transformer to exploit semantic and visual information simultaneously. Furthermore, it use Gated Bilateral Controller to guide the interactions between the multimodal information. 10:Entity Aware [6]: This method constructs named entities from related articles to graph. Subsequently, it chooses the proper named entities with the highest occurrence from the directly connected entities. Then it inserts these into the appropriate placeholder in news image template captions to generate the interpretable caption of news images. Finally, we use the method of section 3.5 to detect the consistency of news text and the interpretable caption of news images. 11:GN [7]: This method uses news images and related news text as input to generate the news image caption with named entities. Then, we use the method in section 3.5 to detect the consistency of the news image caption with named entities and news text. The results of verifying the methods are shown in Table 3.

3) RESULTS ANALYSIS

As shown in Table 3, we divide the results into two groups. We can see that our model outperforms the compared models across all metrics. As listed in TopNews dataset, our model improves the Accuracy from 75.6 to 81.3, the Precision from 75.0 to 80.4, the Recall from 75.3 to 80.7, the F1 Measure from 75.4 to 80.5. As listed in Table 5, our model also outperforms the comparable model across all the metrics. We can discover that our model has achieved better results on GoodNews datasets. Because there is no news related articles in the GoodNews dataset. We use the named entities of news text to construct a news knowledge graph. Then use the news knowledge graph to generate news image caption. The news image caption generated in this way has similar background knowledge of the news text. In the first group of the compared models, we discover that these methods extracts the visual features from the news image when detecting consistency of news image and text. It fails to bridge the semantic gap between the news image and text, which cause these methods being ineffective. In the second group of the compared models, we discover that Entity-Aware generally outperforms the GN. Compared to GN, Entity-Aware extracts named entities from news related articles and maps them into a graph, moreover Entity-Aware analyzes the relation of directly connected entities which obtain the direct relations of named entities in the graph. However, in addition to direct relations, we also need to analyze the indirect relations of entities in the graph. Furthermore, we discover that some news images contain metaphors, such as ‘stars and stripes’ meaning the USA, and the ‘five-star red flag’ meaning China. When these metaphors appear in news images, our method is unable to accurately generate the news image caption with named entities.
4) QUALITATIVE ANALYSIS

Figure 6 shows two examples that are detecting the consistency of news images and text of our model. Figure 6(a) is the image-text consistency news. Figure 6(a) news text is about the UEFA Champions League sports event. Figure 6(a) news image shows Timo Werner playing soccer. Generating the news image caption with named entities by using our model is <Timo Werner is playing soccer in Cologne> where the news image caption with named entities matches the news text. Thus, our model determines this news is image-text consistency. Figure 6(b) is the image-text mismatching news. Figure 6(b) news text is about changes in customer behavior during a time of economic downturn. Figure 6(a) news image shows a group of Federal Agents. Generating the news image caption with named entities by using our model is <Federal agents standing in Washington>. The news image caption with named entities is completely irrelevant to the corresponding news text. Thus, our model determines this news is image-text mismatching. Figure 6(c) is the image-text matching news. Figure 6(c) news text is about China-U.S. Trade issues. Figure 6(c) news image is the Chinese flag and the American flag. Generating the news image caption with named entities by using our model is <A red and white flag is in front of a red flag>. The news image caption with named entities is completely irrelevant to the corresponding news text. Thus, our model determines this news is image-text mismatching. Our method detects errors. In this news image, the five-star red flag heralds China, and the stars and stripes herald the United States. The existing computer vision technology is difficult to analyze the abstract semantics in images. When these metaphors appear in news images, our method is unable to accurately generate the news image caption with named entities. So our method is less effective for this kind of news.

V. CONCLUSION

In this paper, we propose a novel method to detecting the consistency of news images and text. In this process, firstly to construct the relations accurately between named entities in a news story. We build the news knowledge graphs by extracting named entities in Topnews. Subsequently, to analyzing the whole relations and infer the indirect relations between named entities in news knowledge graph, we propose the NKD-GNN to choose named entities to fill-in the placeholders in the template caption. Finally, we detect the consistency of the news image caption with named entities and text. The results of extensive experiments based on TopNews dataset demonstrate that our method is effective in detecting the consistency of news images and text. But, our method cannot extract the metaphors news images. In future work, we will aim to explore metaphors in news images and improve the accuracy in news image-text matching task.

REFERENCES

[1] L. Ma, Z. Lu, L. Shang, and H. Li, “Multimodal convolutional neural networks for image-text matching task,” in Proc. IEEE Int. Conf. Comput. Vis., Dec. 2015, pp. 2623–2631.
[2] L. Wang, Y. Li, J. Huang, and S. Lazebnik, “Learning two-branch neural networks for image-text matching tasks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 394–407, Feb. 2019.
[3] R. Liu, Y. Zhao, S. Wei, L. Zheng, and Y. Yang, “Modality-invariant image-text embedding for image-sentence matching,” ACM Trans. Multimedia Comput., Commun., Appl., vol. 15, no. 1, pp. 1–19, Feb. 2019.
[4] T. Wang, X. Xu, Y. Yang, A. Hanjalic, H. T. Shen, and J. Song, “Matching images and text with multi-modal tensor fusion and re-ranking,” in Proc. 27th ACM Int. Conf. Multimedia, Oct. 2019, pp. 12–20.
[5] C. Alberti, J. Ling, M. Collins, and D. Reitter, “Fusion of detected objects in text for visual question answering,” in Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process., 2019, pp. 2131–2140.
[6] D. Lu, S. Whitehead, L. Huang, H. Ji, and S.-F. Chang, “Entity-aware image caption generation,” in Proc. Empirical Methods Natural Lang. Process, 2018, pp. 1–11.
[7] A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas, “Good news, everyone! Context driven entity-aware captioning for news images,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 12466–12475.
[8] A. Ramisa, F. Yan, F. Moreno-Noguer, and K. Mikolajczyk, “BreakingNews: Article annotation by image and text processing,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 5, pp. 1072–1085, May 2018.
[9] R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” 2014, arXiv:1411.2539. [Online]. Available: http://arxiv.org/abs/1411.2539.
[10] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, “Show, attend and tell: Neural image caption generation with visual attention,” in Proc. Int. Conf. Mach. Learn., 2016, pp. 2048–2057.
[11] J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing when to look: Adaptive attention via a visual Sentinel for image captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., vol. 6, Jul. 2017, pp. 375–383.
[12] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, “Image captioning with semantic attention,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2016, pp. 4651–4659.
[13] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, “Areas of attention for image captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Oct. 2017, pp. 1242–1250.
[14] Y. Huang, W. Wang, and L. Wang, “Instance-aware image and sentence matching with selective multimodal LSTM,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 2310–2318.
[15] Y. Huang, W. Wang, and L. Wang, “Instance-aware image and sentence matching with selective multimodal LSTM,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 299–307.
[16] H. Nan, J.-W. Ha, and J. Kim, “Dual attention networks for multimodal reasoning and matching,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jul. 2017, pp. 232–241.
[17] Z. Chen, H. Xu, S. Gao, and Y. Wang, “Semantic generation mechanism of news images based on concept-net,” in Proc. IEEE Int. Conf. Power Electron., Comput. Appl. (ICPECA), Jan. 2021, pp. 245–248.
[18] P. Sharma, N. Ding, S. Goodman, and R. Soricut, “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,” in Proc. 56th Annu. Meeting Assoc. Comput. Linguistics. Melbourne, VIC, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2556–2565.
[19] S. Subramanian, R. Li, J. Pilault, and A. J. Pal, “On extractive and abstractive neural document summarization with transformer language model,” 2019, arXiv:1909.03186. [Online]. Available: https://arxiv.org/abs/1909.03186.
[20] Y. Jing, X. Zhiwei, and G. Guanglai, “Context-driven image caption with global semantic relations of the named entities,” IEEE Access, vol. 8, pp. 143584–143594, 2020.
[21] Y. Wu, L. Zhu, L. Jiang, and Y. Yang, “Decoupled novel object captioner,” in Proc. 26th ACM Int. Conf. Multimedia, Oct. 2018, pp. 1029–1037.
[22] T. Yao, Y. Pan, Y. Li, and T. Mei, “Exploring visual relationship for image captioning,” in Proc. Eur. Conf. Comput. Vis. (ECCV), Sep. 2018, pp. 684–699.
[23] R. Hu, A. Rohrbach, T. Darrell, and K. Saenko, “Language-conditioned graph networks for relational reasoning,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 10294–10303.
[24] T. Yao, Y. Pan, Y. Li, and T. Mei, “Hierarchy parsing for image captioning,” in Proc. IEEE/CVF Int. Conf. Comput. Vis., Oct. 2019, pp. 2621–2629.
ZAO YUMENG received the B.S. degree from the Inner Mongolia University of China, Hohhot, China, in 2018. He is currently pursuing the M.S. degree with the College of Data Science and Application, Inner Mongolia University of Technology, Hohhot. His research interests include pattern recognition and deep learning.

YUN JING received the B.S. degree from Inner Mongolia University, Hohhot, China, in 2003, and the M.S. degree from Beijing Information Science and Technology University, Beijing, China, in 2009. She is currently pursuing the Ph.D. degree with Inner Mongolia University. She has been with Inner Mongolia University, since 2013. Her research interests include image processing, image caption, and machine learning algorithms. She is currently a Lecturer with the College of Data Science and Technology, Inner Mongolia University of Technology.

GAO SHUO received the B.S. degree from Weifang University, Weifang, China, in 2018. He is currently pursuing the M.S. degree with the College of Data Science and Application, Inner Mongolia University of Technology, Hohhot, China. His research interests include pattern recognition and deep learning.

LIU LINLIN received the B.S. degree from Inner Mongolia University, Hohhot, China, in 1987, and the M.S. degree from Tsinghua University, Beijing, China, in 2002. He is a Professor and a M.S. Supervisor. His research interest includes internet public opinion analysis.

* * *