Entity-aware Image Caption Generation

Di Lu[1], **Spencer Whitehead**[1], **Lifu Huang**[1],**Heng Ji**[1], **Shih-Fu Chang**[2]
[1]Computer Science Department, Rensselaer Polytechnic Institute
{lud2,whites5,huangl7,jih}@rpi.edu
[2]Computer Science Department, Columbia University
sfchang@cs.columbia.edu

Abstract

Current image captioning approaches generate descriptions which lack specific information, such as named entities that are involved in the images. In this paper we propose a new task which aims to generate informative image captions, given images and hashtags as input. We propose a simple but effective approach to tackle this problem. We first train a convolutional neural networks - long short term memory networks (CNN-LSTM) model to generate a template caption based on the input image. Then we use a knowledge graph based collective inference algorithm to fill in the template with specific named entities retrieved via the hashtags. Experiments on a new benchmark dataset collected from Flickr show that our model generates news-style image descriptions with much richer information. Our model outperforms unimodal baselines significantly with various evaluation metrics. [1]

1 Introduction

As information regarding emergent situations disseminates through social media, the information is presented in a variety of data modalities (*e.g.* text, images, and videos), with each modality providing a slightly different perspective. Images have the capability to vividly represent events and entities, but without proper contextual information they become less meaningful and lose utility. While images may be accompanied by associated tags or other meta-data, which are inadequate to convey detailed events, many lack the descriptive text to provide such context. For example, there are 17,285 images on Flickr from the Women's March on January 21, 2017,[2] most of which contain only a few tags and lack any detailed text descriptions. The absence of context leaves individuals with no knowledge of details such as the purpose or location of the march.

Image captioning offers a viable method to directly provide images with the necessary contextual information through textual descriptions. Advances in image captioning (Xu et al., 2015; Fang et al., 2015; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015) are effective in generating sentence-level descriptions. However, sentences generated by these approaches are usually generic descriptions of the visual content and ignore background information. Such generic descriptions do not suffice in emergent situations as they, essentially, mirror the information present in the images and do not provide detailed descriptions regarding events and entities present in, or related to, the images, which is imperative to understanding emergent situations. For example, given the image in Figure 1, the state-of-the-art automatically generated caption is '*Several women hold signs in front of a building.*', which is lacking information regarding relevant entities (*e.g.* ' *Junior doctors*', '*Tories*').

In our work, we propose an ambitious task: **entity-aware image caption generation**: automatically generate an image description that incorporates specific information such as named entities, relevant to the image, given limited text information, such as associated tags and meta-data (*e.g.* time of photo and geo-location). Our approach to this task generally follows three steps. First, instead of directly generating a sentence for an image, we generate a ***template*** sentence with fillable slots by training an image captioning architecture on image-caption pairs, where we replace the entities from the captions with slot types indicating the type of entity that should be used

[1] Datasets and programs: https://github.com/dylandilu/Entity-aware-Image-Captioning

[2] https://en.wikipedia.org/wiki/2017_Women%27s_March

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4013–4023
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

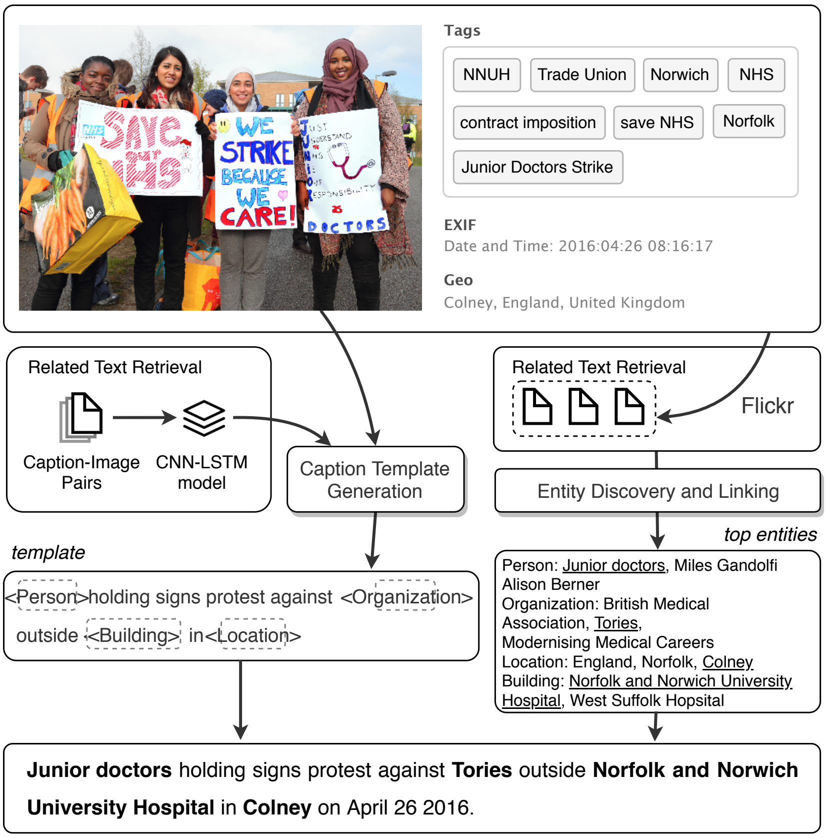
Tags
NNUH
Trade Union
Norwich
NHS
contract imposition
save NHS
Norfolk
Junior Doctors Strike
EXIF
Date and Time: 2016:04:26 08:16:17
Geo
Colney, England, United Kingdom
Related Text Retrieval
Caption-Image Pairs
CNN-LSTM model
Caption Template Generation
Related Text Retrieval
Flickr
Entity Discovery and Linking
top entities
template
<Person>holding signs protest against <Organization>
outside <Building> in<Location>
Person: Junior doctors, Miles Gandolfi
Alison Berner
Organization: British Medical
Association, Tories,
Modernising Medical Careers
Location: England, Norfolk, Colney
Building: Norfolk and Norwich University
Hospital, West Suffolk Hospital
Junior doctors holding signs protest against **Tories** outside **Norfolk and Norwich University Hospital** in **Colney** on April 26 2016.

Figure 1: The overall framework.

to fill the slot. A ***template*** for the example in Figure 1 is: <Person> holding signs protest against <Organization> outside <Building> in <Location>.

Second, given the associated tags of an image, we apply entity discovery and linking (EDL) methods to extract specific entities from previous posts that embed the same tags. Finally, we select appropriate candidates for each slot based upon the entity type and frequency. For example, we select the person name '*Junior doctors*' to fill in the slot <Person> because it co-occurs frequently with other entities such as '*Tories*' in the text related to the tags $\#NHS, \#JuniorDoctorsStrike$. This framework offers a distinct advantage in that it is very flexible, so more advanced captioning or EDL methods as well as other data can be used almost interchangeably within the framework.

To the best of our knowledge, we are the first to incorporate contextual information into image captioning without large-scale training data or topically related news articles and the generated image captions are closer to news captions.

2 Approach Overview

Figure 1 shows the overall framework of our proposed model. Given an image with associated tags and other meta-data, such as geographical tags and EXIF data,[3] we first feed the image into a template caption generator to generate a sentence composed of context words, such as "*stand*", and slots, such as <person>, to represent missing specific information like named entities (Section 3). The template caption generator, which follows the encoder-decoder model (Cho et al., 2014) with a CNN encoder and LSTM decoder (Vinyals et al., 2015), is trained using news image-template caption pairs.

We then retrieve topically-related images from the Flickr database, which have the same tags as the input image. Next, we apply EDL algorithms to the image titles to extract entities and link them to external knowledge bases to retrieve their fine-grained entity types. Finally, for each slot generated by the template generator, we choose to fill the slot with the appropriate candidate based on entity type and frequency (Section 4).

[3] EXIF data contains meta-data tags of photos such as date, time, and camera settings.

3 Template Caption Generation

Language models (LM) are widely used to generate text (Wen et al., 2015; Tran and Nguyen, 2017) and play a crucial role in most of the existing image captioning approaches (Vinyals et al., 2015; Xu et al., 2015). These models, learned from large-scale corpora, are able to predict a probability distribution over a vocabulary. However, LM struggle to generate specific entities, which occur sparsely, if at all, within training corpora. Moreover, the desired entity-aware captions may contain information not directly present in the image alone. Unless the LM is trained or conditioned on data specific to the emergent situation of interest, the LM alone cannot generate a caption that incorporates the specific background information. We address this issue by only relying on the LM to generate abstract slot tokens and connecting words or phrases, while slot filling is used to incorporate specific information. This approach allows the LM to focus on generating words or phrases with higher probability, since each slot token effectively has a probability equal to the sum of all the specific entities that it represents, thereby circumventing the issue of generating lower probability or out-of-vocabulary (OOV) words.

In this section, we describe a novel method to train a model to automatically generate template captions with slots as '*placeholders*' for specific background information. We first present the schemas which define the slot types (Section 3.1) and the procedure to acquire training data for template generation (Section 3.2). Finally, we introduce the model for template caption generation (Section 3.3).

3.1 Template Caption Definition

Named entities are the most specific information which cannot be easily learned by LM. Thus, in this work, we define slots as placeholders for entities with the same types. We use the fine grained entity types defined in DBpedia [4] (Auer et al., 2007) to name the slots because these types are specific enough to differentiate between a wide range of entities and still general enough so that the slots have higher probabilities in the language model. For example, `Person` is further divided into `Athlete`, `Singer`, `Politician`, and so on. Therefore, a template caption like '`Athlete` celebrates after scoring.' can be generated by the language model through leveraging image features, where the slot `Athlete` means a sports player (e.g., Cristiano Ronaldo, Lionel Messi).

3.2 Acquisition of Image-Template Caption Pairs

High quality training data is crucial to train a template caption generator. However, the image-caption datasets used in previous work, such as Microsoft Common Objects in Context (MS COCO) (Lin et al., 2014) and Flickr30K (Rashtchian et al., 2010), are not suitable for this task because they are designed for non-specific caption generation and do not contain detailed, specific information such as named entities. Further, manual creation of captions is expensive. In this work, we utilize news image-caption pairs, which are well written and can be easily collected. We use the example in Figure 2 to describe our procedure to convert image-caption to image-template caption pairs: **preprocessing**, **compression**, **generalization**.

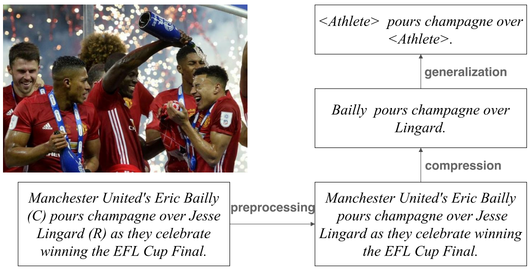
<Athlete> pours champagne over <Athlete>.
generalization
Bailly pours champagne over Lingard.
compression
Manchester United's Eric Bailly (C) pours champagne over Jesse Lingard (R) as they celebrate winning the EFL Cup Final.
preprocessing
Manchester United's Eric Bailly pours champagne over Jesse Lingard as they celebrate winning the EFL Cup Final.

Figure 2: Procedure of Converting News Captions into Templates.

Preprocessing: We first apply the following pre-processing steps: (1) remove words in parentheses, such as *'(C)'* and *'(R)'* in Figure 2, because they usually represent auxiliary information and are not aligned with visual concepts in images; (2) if a caption includes more than one sentence, we choose the longer one. Based on our observation, shorter sentences usually play the role of background introduction, which are not aligned with the key content in images; (3) remove captions with less than 10 tokens because they tend to be not informative enough. The average length of the news image captions is 37 tokens.

Compression: The goal of compression is to make news captions short and aligned with images as much as possible by keeping informa-

[4] We use the sixth level entity types in Yago ontology (Wordnet types only).

tion related to objects and concepts in the images, which are usually subjects, verbs and objects in sentences. In this paper, we propose a simple but efficient compression method based on dependency parsing. We do not use other complicated compressions (Kuznetsova et al., 2014) because our simple method achieves comparative results on image caption dataset. We first apply the Stanford dependency parser (De Marneffe and Manning, 2008) on preprocessed captions. Then, we traverse the parse tree from the root (*e.g.‘pours’*) via <governor, grammatical relations, dependent> triples using breadth-first search. We decide to keep a dependent or not based on its grammatical relation with the governor. Based on our observations, among the 50 grammatical relations in the Stanford dependency parser, we keep the dependents that have the following grammatical relations with their governors: *nsubj, obj, iobj, dobj, acomp, det, neg, nsubjpass, pobj, predet, prep, prt, vmod, nmod, cc*.

Generalization: The last step for preparing training data is to extract entities from captions and replace them with the slot types we defined in Section 3.1. We apply Stanford CoreNLP name tagger (Manning et al., 2014) to the captions to extract entity mentions of the following types: Person, Location, Organization, and Miscellaneous. Next, we use an English Entity Linking algorithm (Pan et al., 2015) to link the entity mentions to DBpedia and retrieve their fine-grained types.[5] We choose the higher level type if there are multiple fine-grained types for a name. For example, the entity types of *Manchester United*, *Eric Bailly*, and *Jesse Lingard* are *SoccerTeam*, *Athlete*, and *Athlete*, respectively. For entity mentions that cannot be linked to DBpedia, we use their coarse-grained entity types, which are the outputs of name tagger.

Finally, we replace the entities in the compressed captions with their corresponding slots:

Generalized Template: <Athlete> pours champagne over <Athlete>.

3.3 Generation Model

Using the template caption and image pairs *(S, I)* as training data, we regard the template caption generation as a regular image captioning task. Thus, we adapt the encoder-decoder architecture which is successful in the image captioning task (Vinyals et al., 2015; Xu et al., 2015). Our model (Figure 3) is most similar to the one proposed in (Vinyals et al., 2015). Note, other captioning methods may easily be used instead.

Encoder: Similar to previous work (Vinyals et al., 2015; Xu et al., 2015; Karpathy and Fei-Fei, 2015; Venugopalan et al., 2017), we encode images into representations using a ResNet (He et al., 2016) model pre-trained on the ImageNet dataset (Deng et al., 2009) and use the outputs before the last fully-connected layer.

Decoder: We employ a Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) based language model to decode image representations into template captions. We provide the LSTM the image representation, I, as the initial hidden state. At the t^{th} step, the model predicts the probabilities of words/slots, y_t, based on the word/slot generated at last time step, y_{t-1}, as well as the hidden state, s_t.

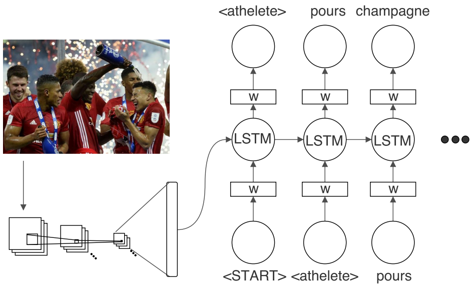

Figure 3: LSTM language generator.

4 Template Caption Entity Population

With the generated template captions, our next step is to fill in the slots with the appropriate specific entities to make the caption complete and entity-aware. In this section, we expand our method to extract candidate entities from contextual information (*i.e.*, images in Flickr with the same tags). Once we extract candidate entities, we apply the Quantified Collective Validation (QCV) algorithm (Wang et al., 2015), which constructs a number of candidate graphs and performs collective validation on those candidate graphs to choose the appropriate entities for the slots in the template caption.

[5] This yields a total of 95 types after manually cleaning.

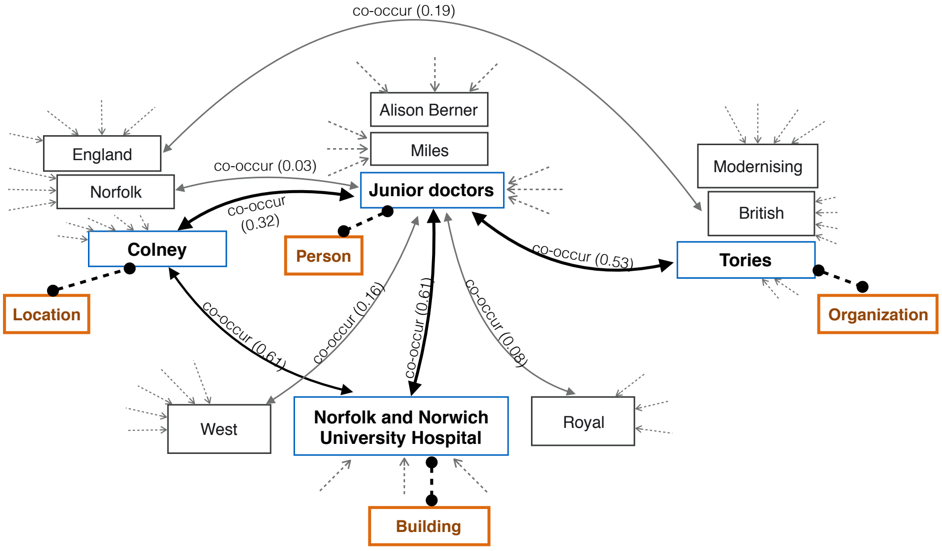

Figure 4: Knowledge graph construction.

4.1 Candidate Entity Retrieval

Users on social media typically post images with tags that are event-related (e.g. #occupywallstreet), entity-related (e.g. #lakers), or topic-related (e.g. #basketball). On our Flickr testing dataset, the average number of tags associated with an image is 11.5 and posts with the same tags likely share common entities. Therefore, given an image and its tags, we retrieve images from Flickr with the same tags by a window size of seven-day based on taken date of the photo, and then utilize the textual information accompanying the retrieved images as context. We filter out the high frequency hashtags($>$ 200 in testing dataset). Because some common tags, such as '#concert', appear in lots of posts related to different concerts.

Given the related text, we apply EDL algorithms (Pan et al., 2015) to extract named entities and link them to DBpedia to obtain their entity types. For each entity type, we rank the candidates based on their frequency in the context and only keep the top 5 candidate entities.

4.2 Quantified Collective Validation

Each slot in the template must be filled with an entity from its corresponding candidate entities. We can regard this step as an entity linking problem, whose goal is to choose an entity from several candidates given an entity mention. We utilize the QCV algorithm to construct a number of candidate graphs for a given set of slots (Wang et al., 2015), where each combination of candidate entities substituted into the slots yields a different graph (Figure 4). For each candidate combination graph, G_c^i, we compute the edge weights between each pair of candidates in the graph as

$$H_r = \frac{f_{c_h c_t}}{max(f_{c_h}, f_{c_t})} \qquad (1)$$

where $r \in E(G_c^i)$ is an edge in G_c^i, c_h and c_t are the head candidate and tail candidate of the edge, $f_{c_h c_t}$ is the co-occurrence frequency of the pair of candidates, and f_{c_h} and f_{c_t} are the individual frequencies of head candidate and tail candidate, respectively. For example, in Figure 4, *Colney* (Location) and *Junior doctors* (Person) co-occur frequently, therefore the edge between them has a larger weight.

We compute the summed edge weight, $\omega(G_c^i)$, for each G_c^i by

$$\omega(G_c^i) = \sum_{r \in E(G_c^i)} H_r \qquad (2)$$

and select the combination of candidates with the largest $\omega(G_c^i)$ to fill in the set of slots.

As a result of this process, given the template: '<Person> holding signs protest against <Organization> outside <Building> in <Location>.', we obtain an entity-aware caption: '***Junior doctors*** *holding signs protest against* ***Tories*** *outside* ***Norfolk and Norwich University Hospital*** *in* ***Colney***'.

4.3 Post-processing

Some images in Flickr have EXIF data which gives the date that an image is taken. We convert this information into the format such as '***April 26 2016***' and add it to the generated captions as post-processing, by which we obtain the complete caption: '*Junior doctors holding signs protest against Tories outside Norfolk and Norwich University Hospital in Colney on **April 26 2016**.*'. We leave the generated caption without adding date information if it is not available. For those slots that cannot be filled by names, we use general words to replace them, such as using the word *'Athlete'* to replace the slot `Athlete`.

5 Experiments

5.1 Data

We require images with well-aligned, news-style captions for training. However, we want to test our model on real social media data and it is difficult to collect these informative captions for social media data. Therefore, we acquire training data from news outlets and testing data from social media. We select two different topics, social events and sports events, as our case studies.

	Train	Dev	Test
Number of Images	29,390	4,306	3,688
Number of Tokens	1,086,350	161,281	136,784
Social Event	6,998	976	872
Sports Event	22,392	3,330	2,816
Person	32,878	4,539	4,156
Location	42,786	5,657	5,432
Organization	17,290	2,370	2,124
Miscellaneous	8,398	1,103	1,020

Table 1: Statistics of datasets for template generation.

Template Generator Training and Testing. To train the template caption generator, we collect 43,586 image-caption pairs from Reuters[6], using topically-related keywords[7] as queries. We do not use existing image caption datasets, such as MSCOCO (Lin et al., 2014), because they do not contain many named entities. After the compression and generalization procedures (Section 3.2) we keep 37,384 images and split them into train, development, and test sets. Table 1 shows the statistics of the datasets.

Entity-aware Caption Testing. Since news images do not have associated tags, for the purpose of testing our model in a real social media setting, we use images from Flickr for our caption evaluation[8], which is an image-centric, representative social media platform. We use the same keywords as for template generator training to retrieve multi-modal data with Creative Commons license[9], for social and sports events. We choose the images that already have news-style descriptions from users and manually confirm they are well-aligned. In total, we collect 2,594 images for evaluation. For each image, we also obtain the tags (30,148 totally) and meta-data, such as EXIF and geotag data, when they are available.

5.2 Models for Comparison

We compare our entity-aware model with the following baselines:

CNN-RNN-raw. We use the model proposed by Vinyals et al. (2015) to train an image captioning model on the raw news image-caption pairs, and apply to Flickr testing data directly.

CNN-RNN-compressed. We use the model proposed by Vinyals et al. (2015) to train a model on the compressed news image-caption pairs.

Text-summarization. We apply SumBasic summarization algorithms (Vanderwende et al., 2007), that is a summarization for multiple documents based on frequency of word and semantic content units, to text documents retrieved by hashtag.

Entity-aware. We apply trained template generator on Flickr testing data, and then fill in the slots with extracted background information.

5.3 Evaluation Metrics

We use three standard image captioning evaluation metrics, BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014), ROUGE (Lin, 2004) and CIDEr (Vedantam et al., 2015), to evaluate the quality of both the generated templates and generated captions. BLEU is a metric based on correlations at the sentence level. METEOR is a metric with recall weighted higher than precision and it takes into account stemming as well as synonym matching. ROUGE is proposed for evaluation of summarization and relies

[6] https://www.reuters.com/

[7] Social Events: concert, festival, parade, protest, ceremony; Sports Events: Soccer, Basketball, Soccer, Baseball, Ice Hockey

[8] https://www.flickr.com/

[9] https://creativecommons.org/licenses/

highly on recall. CIDEr metric downweights the n-grams common in all image captions, which are similar to tf-idf. Since the goal of this task is to generate entity-aware descriptions, we also measure the entity F1 scores for the final captions, where we do fuzzy matching to manually count the overlapped entities between the system output and reference captions. Besides, we do human evaluation with a score in range of 0 to 3 using the criteria as follows: Score 0: generated caption is not related to the ground-truth; Score 1: generated caption is topically-related to the ground-truth, but has obvious errors; Score 2: generated caption is topically-related to the ground-truth, and has overlapped named entities; Score 3: generated caption well describes the image.

5.4 Template Evaluation

Raw: new york giants kicker josh brown kicks a field goal during the first quarter against the san francisco 49ers at metlife stadium.

Generated Coarse Template: <Person> runs with ball against <Location> <Location> in the first half of their <Miscellaneous> football game in <Location>

Generated Fine Template: <Footballteam> kicker <Player> kicks a field goal out of the game against the <Footballteam> at <Organization> stadium

Figure 5: Example of generated template.

Table 2 shows the performances of template generator based on coarse-grained and fine-grained type respectively, and Figure 5 shows an example of the template generated. Coarse templates are the ones after we replace names with these coarse-grained types. Entity Linking classifies names into more fine-grained types, so the corresponding templates are fine templates. The generalization method of replacing the named entities with entity types can reduce the vocabulary size significantly, which reduces the impact of sparse named entities in training data. The template generation achieves close performance with state-of-the-art generic image captioning on MSCOCO dataset (Xu et al., 2015). The template generator based on coarse-grained entity type outperforms the one based on fine-grained entity type for two reasons: (1) fine template relies on EDL, and incorrect linkings import noise; (2) named entities usually has multiple types, but we only choose one during generalization. Thus the caption, '*Bob Dylan performs at the Wiltern Theatre in Los Angeles*', is generalized into '*<Writer> performs at the <Theater> in <Loaction>*', but the correct type for *Bob Dylan* in this context should be Artist.

5.5 Flickr Caption Results

Table 4 shows the comparison between our model and the baselines. The scores are much lower than traditional caption generation tasks such as COCO, because we use the real captions as ground-truth. Our model outperforms all the baselines on all metrics except BLEU-4, where Text-summarization model achieves better score. Generally, the model based on textual features (Text-summarization) has better performance than vision-based models (CNN-RNN-raw and CNN-RNN-compressed). It indicates textual summarization algorithm is more effective when it involves specific knowledge generation. Text-summarization model generates results from documents retrieved by hashtags, so it tends to include some long phrases common in those documents. However the templates generated by our model is based on the language model trained from the news captions, which has different style with Flickr captions. It results in that Text-summarization model achieves better BLEU-4 score. Our model improves CIDEr score more significantly compared with other metrics, because CIDEr downweights the n-grams common in all captions, where more specific information such as named entities contribute more to the scores. The experimental results demonstrate that our model is effective to generate image captions with specific knowledge.

5.6 Analysis

Figure 6 shows some examples of the captions generated by the entity-aware model.

Good Examples: (A) in Figure 6 describes the events in the images well ('*performs*') and include correct, well-placed entities, such as '*Little Dragon*' and '*House of Blues*'.

Approach	Vocabulary	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
Raw-caption	10,979	15.1	11.7	9.9	8.8	8.8	24.2	34.7
Coarse Template	3,533	**46.7**	**36.1**	**29.8**	**25.7**	**22.4**	**43.5**	161.6
Fine Template	3,642	43.0	33.4	27.8	24.3	20.3	39.8	**165.3**

Table 2: Comparison of Template Generator with coarse/fine-grained entity type. Coarse Template is generalized by coarse-grained entity type (name tagger) and fine template is generalized by fine-grained entity type (EDL).

#littledragon#houseofblues
#cleveland#ohio#concert

EXIF: 2017-08-02 20:42:53

(A)

#mlb#baseball
#orioleparkatcamdenyards
#joekelly#bostonredsox

EXIF: 2016-06-01 19:25:01

(B)

#toronto#tiff#tiff17
#tiff2017#raptors#patrick
#patterson#patrickpatterson

EXIF: 2017-09-09 15:38:21

(C)

	Model	Caption
A	CNN-RNN-compressed	jack white from rock band the dead weather performs during the 44th montreux jazz festival in montreux
	Text-summarization	Little Dragon performing at the House of Blues in Cleveland, OH
	Entity-aware(ours)	singer **little dragon** performs at the **house of blues** in **cleveland** August 2 2017
	Human	**little dragon** performing at the **house of blues** in **cleveland, oh**
B	CNN-RNN-compressed	houston astros starting pitcher brett delivers in the second inning against the cleveland indians at progressive field
	Text-summarization	Red Sox at Orioles 6/2/16
	Entity-aware(ours)	**baltimore orioles** starting pitcher **joe kelly** pitches in the first inning against the **baltimore orioles** at **baltimore** June 1 2016
	Human	**joe kelly** of the **boston red sox** pitches in a game against the **baltimore orioles** at **oriole park** at camden yards on june 1, 2016 in **baltimore, maryland**
C	CNN-RNN-compressed	protestors gesture and hold signs during a protest against what demonstrators call police brutality in mckinney , texas .
	Text-summarization	Toronto, Canada September 9, 2017.
	Entity-aware(ours)	supporters of an ban protest outside the **toronto international film festival** in **toronto** September 9 2017
	Human	**patrick patterson** at the premiere of the **carter effect**, 2017 **toronto film festival**

Figure 6: Examples of generated entity-aware caption.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr	F1	Human*
CNN-RNN-raw	7.5	2.7	0.9	0.3	2.6	7.5	1.1	1.2	0.49
CNN-RNN-compress	8.3	3.3	1.1	0.5	3.0	9.2	1.5	2.1	0.50
Text-summarization	9.5	8.0	7.1	**6.5**	9.9	11.9	17.2	35.4	0.59
Entity-aware(ours)	**25.5**	**14.9**	**8.0**	4.7	**11.0**	**21.1**	**29.9**	**39.7**	**0.87**

Table 4: Comparison between our entity-aware model and baseline models on various topics. (* We make human evaluation on 259 images randomly selected.)

Relation Error of Filled Entities: Some of our errors result from ignoring of relations between entities. In Example (B) of Figure 6 our model generate a good template, but connects '*Joe Kelly*', who is actually a pitcher of '*Res Sox*', with '*Baltimore Orioles*' incorrectly. One possible solution is to incorporate relation information when the model fills in the slots with entities.

Template Error: Another category of errors results from wrong templates generated by our model. Examples (C) in Figure 6 is about a film festival, but the model generates a template about protest, which is not related to the image. One potential improvement is to incorporate the information from associated tags, such as the number of tags and the named entity types related to the tags, as features during template caption generation to make generated templates dynamically change according to the context.

6 Related Work

The goal of image captioning is to automatically generate a natural language sentence given an image. (Kulkarni et al., 2013; Yang et al., 2011; Mitchell et al., 2012) perform object recognition in images and fill hand-made templates with the recognized objects. (Kuznetsova et al., 2012, 2014) retrieve similar images, parse associated captions into phrases, and compose them into new sentences. Due to the use of static, handmade templates, these approaches are unable to generate a variety of sentence realizations, which can result in poorly generated sentences and requires one to manually create more templates to extend the generation. Our approach overcomes this by dynamically generating the output.

More recent work utilizes neural networks and applies an encoder-decoder model (Cho et al., 2014). Vinyals et al. (2015) use a CNN to encode images into a fixed size vector representation and a LSTM to decode the image representations into a sentence. Xu et al. (2015) incorporate an attention mechanism (Bahdanau et al., 2015) and attend to the output from a convolutional layer of a CNN to produce the image representations for the decoder. Instead of encoding a whole image as a vector, (Johnson et al., 2016) apply R-CNN object detection (Girshick et al., 2014), match text snippets to the regions of the image detected by the R-CNN, and use a recurrent neural network (RNN) language model, similar to (Vinyals et al., 2015), to generate a description of each region.

The surface realization for state-of-the-art neural approaches is impressive, but, in the context of generating entity-aware captions, these methods fall short as they heavily rely on training data for language modeling. (Tran et al., 2016) leverage face and landmark recognition to generate captions containing named persons, but such large-scale training is difficult. Consequently, OOV words like named entities, which are a quintessential aspect of entity-aware captions because OOV words typically represent entities or events, are difficult to generate due to low training probabilities. Some work has been done to incorporate novel objects into captions (Venugopalan et al., 2017), but this does not address the need to generate entity-aware captions and incorporate contextual information; rather, it gives the ability to generate more fine-grained entities within captions that still lack the necessary context. (Feng and Lapata, 2013) also generates a caption with named entities, but from associated news articles, in which there is much more textual context than our setting. Our approach uses neural networks to generate dynamic templates and then fills in the templates with specific entities. Thus, we are able to combine the sentence variation and surface realization quality of neural language modeling and the capability to incorporate novel words of template-based approaches.

7 Conclusions and Future Work

In this paper we propose a new task which aims to automatically generate entity-aware image descriptions with limited textual information. Experiments on a new benchmark dataset collected from Flickr show that our approach generates more informative captions compared to traditional image captioning methods. Moreover, our two-step approach can easily be applied to other language generation tasks involving specific information.

In the future, we will expand the entity-aware model to incorporate the relations between candidates when the model fills in the slots, which can avoid the cases such as *'Cristiano Ronaldo of Barcelona'*. We will also make further research on context-aware fine-grained entity typing to train a better template generator. Another research direction based on this work is to develop an end-to-end neural architecture to make the model more flexible without generating a template in the middle.

Acknowledgments

This work was supported by the U.S. DARPA AIDA Program No. FA8750-18-2-0014 and U.S. ARL NS-CTA No. W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the third International Conference on Learning Representations*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Marie-Catherine De Marneffe and Christopher D Manning. 2008. Stanford typed dependencies manual. Technical report, Technical report, Stanford University.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*.

Yansong Feng and Mirella Lapata. 2013. Automatic caption generation for news images. *IEEE transactions on pattern analysis and machine intelligence*, 35(4):797–812.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 2014 IEEE conference on computer vision and pattern recognition*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*.

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903.

Polina Kuznetsova, Vicente Ordonez, Alexander C Berg, Tamara L Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*.

Polina Kuznetsova, Vicente Ordonez, Tamara L Berg, and Yejin Choi. 2014. Treetalk: Composition and compression of trees for image descriptions. *Transactions of the Association of Computational Linguistics*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proceedings of the 2014 European Conference on Computer Vision*.

Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*.

Xiaoman Pan, Taylor Cassidy, Ulf Hermjakob, Heng Ji, and Kevin Knight. 2015. Unsupervised entity linking with abstract meaning representation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics–Human Language Technologies*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics–Human Language Technologies Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*.

Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. 2016. Rich image captioning in the wild. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Van-Khanh Tran and Le-Minh Nguyen. 2017. Natural language generation for spoken dialogue system using rnn encoder-decoder networks. In *Proceedings of the 21st Conference on Computational Natural Language Learning*.

Lucy Vanderwende, Hisami Suzuki, Chris Brockett, and Ani Nenkova. 2007. Beyond sumbasic: Task-focused summarization with sentence simplification and lexical expansion. *Information Processing & Management*, 43(6):1606–1618.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2017. Captioning images with diverse objects. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*.

Han Wang, Jin Guang Zheng, Xiaogang Ma, Peter Fox, and Heng Ji. 2015. Language and domain independent entity linking with quantified collective validation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 2015 International Conference on Machine Learning*.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.