Learning to Describe Unknown Phrases with Local and Global Contexts

Shonosuke Ishiwatari[†] **Hiroaki Hayashi**[‡] **Naoki Yoshinaga**[§] **Graham Neubig**[‡]
Shoetsu Sato[†] **Masashi Toyoda**[§] **Masaru Kitsuregawa**[¶§]
† The University of Tokyo ‡ Carnegie Mellon University
§ Institute of Industrial Science, the University of Tokyo ¶ National Institute of Informatics
[†§¶]{ishiwatari,ynaga,shoetsu,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp
[‡]{hiroakih, gneubig}@cs.cmu.edu

Abstract

When reading a text, it is common to become stuck on unfamiliar words and phrases, such as polysemous words with novel senses, rarely used idioms, internet slang, or emerging entities. If we humans cannot figure out the meaning of those expressions from the immediate *local* context, we consult dictionaries for definitions or search documents or the web to find other *global* context to help in interpretation. Can machines help us do this work? Which type of context is more important for machines to solve the problem? To answer these questions, we undertake a task of describing a given phrase in natural language based on its local and global contexts. To solve this task, we propose a neural description model that consists of two context encoders and a description decoder. In contrast to the existing methods for non-standard English explanation (Ni and Wang, 2017) and definition generation (Noraset et al., 2017; Gadetsky et al., 2018), our model appropriately takes important clues from *both* local and global contexts. Experimental results on three existing datasets (including WordNet, Oxford and Urban Dictionaries) and a dataset newly created from Wikipedia demonstrate the effectiveness of our method over previous work.

1 Introduction

When we read news text with emerging entities, text in unfamiliar domains, or text in foreign languages, we often encounter expressions (words or phrases) whose senses we do not understand. In such cases, we may first try to figure out the meanings of those expressions by reading the surrounding words (*local* context) carefully. Failing to do so, we may consult dictionaries, and in the case of polysemous words, choose an appropriate meaning based on the context. Learning novel word senses via dictionary definitions is known to be
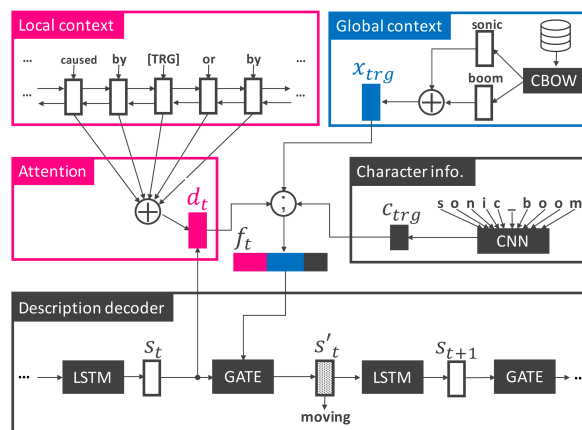

Figure 1: **Lo**cal & **G**lobal **C**ontext-**a**ware **D**escription generator (**LOG-CaD**).

more effective than contextual guessing (Fraser, 1998; Chen, 2012). However, very often, hand-crafted dictionaries do not contain definitions of expressions that are rarely used or newly created. Ultimately, we may need to read through the entire document or even search the web to find other occurances of the expression (*global* context) so that we can guess its meaning.

Can machines help us do this work? Ni and Wang (2017) have proposed a task of generating a definition for a phrase given its local context. However, they follow the strict assumption that the target phrase is newly emerged and there is only a single local context available for the phrase, which makes the task of generating an accurate and coherent definition difficult (perhaps as difficult as a human comprehending the phrase itself). On the other hand, Noraset et al. (2017) attempted to generate a definition of a word from an embedding induced from massive text (which can be seen as global context). This is followed by Gadetsky et al. (2018) that refers to a local context to disambiguate polysemous words by choosing relevant dimensions of their word embeddings. Al-

though these research efforts revealed that both local and global contexts are useful in generating definitions, none of these studies exploited both contexts directly to describe unknown phrases.

In this study, we tackle the task of describing (defining) a phrase when given its local and global contexts. We present **LOG-CaD**, a neural description generator (Figure 1) to directly solve this task. Given an unknown phrase without sense definitions, our model obtains a phrase embedding as its global context by composing word embeddings while also encoding the local context. The model therefore combines both pieces of information to generate a natural language description.

Considering various applications where we need definitions of expressions, we evaluated our method with four datasets including Word-Net (Noraset et al., 2017) for general words, the Oxford dictionary (Gadetsky et al., 2018) for polysemous words, Urban Dictionary (Ni and Wang, 2017) for rare idioms or slang, and a newly-created Wikipedia dataset for entities.

Our contributions are as follows:

- We propose **a general task of defining unknown phrases given their contexts**. This task is a generalization of three related tasks (Noraset et al., 2017; Ni and Wang, 2017; Gadetsky et al., 2018) and involves various situations where we need definitions of unknown phrases (§ 2).

- We propose a method for **generating natural language descriptions for unknown phrases with local and global contexts** (§ 3).

- As a benchmark to evaluate the ability of the models to describe entities, we **build a large-scale dataset** from Wikipedia and Wikidata for the proposed task. We release our dataset and the code[1] to promote the reproducibility of the experiments (§ 4).

- The proposed method achieves **the state-of-the-art performance** on our new dataset and the three existing datasets used in the related studies (Noraset et al., 2017; Ni and Wang, 2017; Gadetsky et al., 2018) (§ 5).

---

[1] https://github.com/shonosuke/ishiwatari-naacl2019

2 Context-aware Phrase Description Generation

In this section, we define our task of describing a phrase in a specific context. Given an undefined phrase $X_{trg} = \{x_j, \cdots, x_k\}$ with its context $X = \{x_1, \cdots, x_I\}$ ($1 \leq j \leq k \leq I$), our task is to output a description $Y = \{y_1, \cdots, y_T\}$. Here, X_{trg} can be a word or a short phrase and is included in X. Y is a definition-like concrete and concise sentence that describes the X_{trg}.

For example, given a phrase "sonic boom" with its context "the shock wave may be caused by *sonic boom* or by explosion," the task is to generate a description such as "sound created by an object moving fast." If the given context has been changed to "this is the first official tour to support the band's latest studio effort, 2009's *Sonic Boom*," then the appropriate output would be "album by Kiss."

The process of description generation can be modeled with a conditional language model as

$$p(Y|X, X_{trg}) = \prod_{t=1}^{T} p(y_t|y_{<t}, X, X_{trg}). \quad (1)$$

3 LOG-CaD: Local & Global Context-aware Description Generator

In this section, we describe our idea of utilizing local and global contexts in the description generation task, and present the details of our model.

3.1 Local & global contexts

When we find an unfamiliar phrase in text and it is not defined in dictionaries, how can we humans come up with its meaning? As discussed in Section 1, we may first try to figure out the meaning of the phrase from the immediate context, and then read through the entire document or search the web to understand implicit information behind the text.

In this paper, we refer to the explicit contextual information included in a given sentence with the target phrase (i.e., the X in Eq. (1)) as "local context," and the implicit contextual information in massive text as "global context." While both local and global contexts are crucial for humans to understand unfamiliar phrases, are they also useful for machines to generate descriptions? To verify this idea, we propose to incorporate both local and global contexts to describe an unknown phrase.

3.2 Proposed model

Figure 1 shows an illustration of our **LOG-CaD** model. Similarly to the standard encoder-decoder model with attention (Bahdanau et al., 2015; Luong and Manning, 2016), it has a context encoder and a description decoder. The challenge here is that the decoder needs to be conditioned not only on the local context, but also on its global context. To incorporate the different types of contexts, we propose to use a gate function similar to Noraset et al. (2017) to dynamically control how the global and local contexts influence the description.

Local & global context encoders We first describe how to model local and global contexts. Given a sentence X and a phrase X_{trg}, a bidirectional LSTM (Gers et al., 1999) encoder generates a sequence of continuous vectors $\boldsymbol{H} = \{\boldsymbol{h}_1 \cdots, \boldsymbol{h}_I\}$ as

$$\boldsymbol{h}_i = \text{Bi-LSTM}(\boldsymbol{h}_{i-1}, \boldsymbol{h}_{i+1}, \boldsymbol{x}_i), \qquad (2)$$

where $\boldsymbol{x}_i$ is the word embedding of word x_i. In addition to the local context, we also utilize the global context obtained from massive text. This can be achieved by feeding a phrase embedding $\boldsymbol{x}_{trg}$ to initialize the decoder (Noraset et al., 2017) as

$$\boldsymbol{y}_0 = \boldsymbol{x}_{trg}. \qquad (3)$$

Here, the phrase embedding $\boldsymbol{x}_{trg}$ is calculated by simply summing up all the embeddings of words that consistute the phrase X_{trg}. Note that we use a randomly-initialized vector if no pre-trained embedding is available for the words in X_{trg}.

Description decoder Using the local and global contexts, a description decoder computes the conditional probability of a description Y with Eq. (1), which can be approximated with another LSTM as

$$\boldsymbol{s}_t = \text{LSTM}(\boldsymbol{y}_{t-1}, \boldsymbol{s}'_{t-1}), \qquad (4)$$
$$\boldsymbol{d}_t = \text{ATTENTION}(\boldsymbol{H}, \boldsymbol{s}_t), \qquad (5)$$
$$\boldsymbol{c}_{trg} = \text{CNN}(X_{trg}), \qquad (6)$$
$$\boldsymbol{s}'_t = \text{GATE}(\boldsymbol{s}_t, \boldsymbol{x}_{trg}, \boldsymbol{c}_{trg}, \boldsymbol{d}_t), \qquad (7)$$
$$p(y_t|y_{<t}, X_{trg}) = \text{softmax}(\boldsymbol{W}_{s'}\boldsymbol{s}'_t + \boldsymbol{b}_{s'}), \qquad (8)$$

where $\boldsymbol{s}_t$ is a hidden state of the decoder LSTM ($\boldsymbol{s}_0 = \vec{0}$), and $\boldsymbol{y}_{t-1}$ is a jointly-trained word embedding of the previous output word y_{t-1}. In what follows, we explain each equation in detail.

Attention on *local* context Considering the fact that the local context can be relatively long (e.g., around 20 words on average in our Wikipedia dataset introduced in Section 4), it is hard for the decoder to focus on important words in local contexts. In order to deal with this problem, the ATTENTION($\cdot$) function in Eq. (5) decides which words in the local context X to focus on at each time step. $\boldsymbol{d}_t$ is computed with an attention mechanism (Luong and Manning, 2016) as

$$\boldsymbol{d}_t = \sum_{i=1}^{T} \alpha_i \boldsymbol{h}_i, \qquad (9)$$
$$\alpha_i = \text{softmax}(\boldsymbol{U}_h \boldsymbol{h}_i^{\text{T}} \boldsymbol{U}_s \boldsymbol{s}_t), \qquad (10)$$

where $\boldsymbol{U}_h$ and $\boldsymbol{U}_s$ are matrices that map the encoder and decoder hidden states into a common space, respectively.

Use of character information In order to capture the surface information of X_{trg}, we construct character-level CNNs (Eq. (6)) following (Noraset et al., 2017). Note that the input to the CNNs is a sequence of words in X_{trg}, which are concatenated with special character "_," such as "sonic_boom." Following Noraset et al. (2017), we set the CNN kernels of length 2-6 and size $10, 30, 40, 40, 40$ respectively with a stride of 1 to obtain a 160-dimensional vector $\boldsymbol{c}_{trg}$.

Gate function to control local & global contexts In order to capture the interaction between the local and global contexts, we adopt a GATE($\cdot$) function (Eq. (7)) which is similar to Noraset et al. (2017). The GATE($\cdot$) function updates the LSTM output $\boldsymbol{s}_t$ to $\boldsymbol{s}'_t$ depending on the global context $\boldsymbol{x}_{trg}$, local context $\boldsymbol{d}_t$, and character-level information $\boldsymbol{c}_{trg}$ as

$$\boldsymbol{f}_t = [\boldsymbol{x}_{trg}; \boldsymbol{d}_t; \boldsymbol{c}_{trg}] \qquad (11)$$
$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z[\boldsymbol{f}_t; \boldsymbol{s}_t] + \boldsymbol{b}_z), \qquad (12)$$
$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r[\boldsymbol{f}_t; \boldsymbol{s}_t] + \boldsymbol{b}_r), \qquad (13)$$
$$\tilde{\boldsymbol{s}}_t = \tanh(\boldsymbol{W}_s[(\boldsymbol{r}_t \odot \boldsymbol{f}_t); \boldsymbol{s}_t] + \boldsymbol{b}_s), \qquad (14)$$
$$\boldsymbol{s}'_t = (1 - \boldsymbol{z}_t) \odot \boldsymbol{s}_t + \boldsymbol{z}_t \odot \tilde{\boldsymbol{s}}_t, \qquad (15)$$

where $\sigma(\cdot)$, $\odot$ and ; denote the sigmoid function, element-wise multiplication, and vector concatenation, respectively. $\boldsymbol{W}_*$ and $\boldsymbol{b}_*$ are weight matrices and bias terms, respectively. Here, the update gate $\boldsymbol{z}_t$ controls how much the original hidden state $\boldsymbol{s}_t$ is to be changed, and the reset gate $\boldsymbol{r}_t$ controls how much the information from $\boldsymbol{f}_t$ contributes to word generation at each time step.
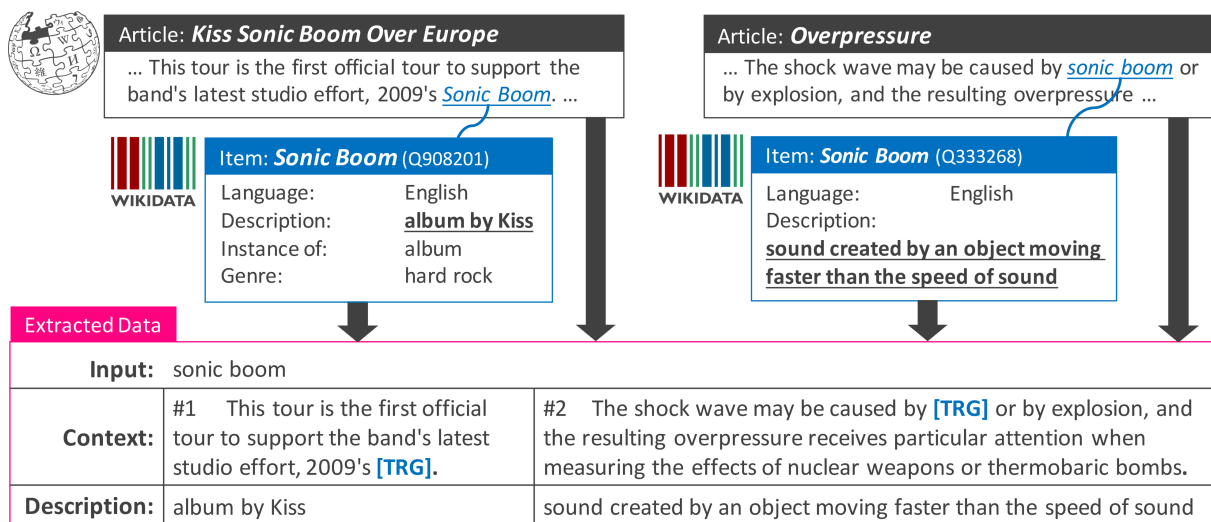
Input:	sonic boom	
Context:	#1 This tour is the first official tour to support the band's latest studio effort, 2009's **[TRG]**.	#2 The shock wave may be caused by **[TRG]** or by explosion, and the resulting overpressure receives particular attention when measuring the effects of nuclear weapons or thermobaric bombs.
Description:	album by Kiss	sound created by an object moving faster than the speed of sound

Figure 2: Context-aware description dataset extracted from Wikipedia and Wikidata.

4   Wikipedia Dataset

Our goal is to let machines describe unfamiliar words and phrases, such as polysemous words, rarely used idioms, or emerging entities. Among the three existing datasets, WordNet and Oxford dictionary mainly target the words but not phrases, thus are not perfect test beds for this goal. On the other hand, although the Urban Dictionary dataset contains descriptions of rarely-used phrases, the domain of its targeted words and phrases is limited to Internet slang.

In order to confirm that our model can generate the description of entities as well as polysemous words and slang, we constructed a new dataset for context-aware phrase description generation from Wikipedia[2] and Wikidata[3] which contain a wide variety of entity descriptions with contexts. The overview of the data extraction process is shown in Figure 2. Each entry in the dataset consists of (1) a phrase, (2) its description, and (3) context (a sentence).

For preprocessing, we applied Stanford Tokenizer[4] to the descriptions of Wikidata items and the articles in Wikipedia. Next, we removed phrases in parentheses from the Wikipedia articles, since they tend to be paraphrasing in other languages and work as noise. To obtain the contexts of each item in Wikidata, we extracted the

sentence which has a link referring to the item through all the first paragraphs of Wikipedia articles and replaced the phrase of the links with a special token **[TRG]**. Wikidata items with no description or no contexts are ignored. This utilization of links makes it possible to resolve the ambiguity of words and phrases in a sentence without human annotations, which is a major advantage of using Wikipedia. Note that we used only links whose anchor texts are identical to the title of the Wikipedia articles, since the users of Wikipedia sometimes link mentions to related articles.

5   Experiments

We evaluate our method by applying it to describe words in WordNet[5] (Miller, 1995) and Oxford Dictionary,[6] phrases in Urban Dictionary[7] and Wikipedia/Wikidata.[8] For all of these datasets, a given word or phrase has an inventory of senses with corresponding definitions and usage examples. These definitions are regarded as ground-truth descriptions.

Datasets   To evaluate our model on the word description task on WordNet, we followed Noraset et al. (2017) and extracted data from WordNet using the `dict-definition`[9] toolkit. Each entry in the data consists of three elements: (1) a word, (2) its definition, and (3) a usage example of the

Corpus	#Phrases	#Entries	Phrase length	Context length	Desc. length
WordNet					
Train	7,938	13,883	1.00	5.81	6.61
Valid	998	1,752	1.00	5.64	6.61
Test	1,001	1,775	1.00	5.77	6.85
Oxford Dictionary					
Train	33,128	97,855	1.00	17.74	11.02
Valid	8,867	12,232	1.00	17.80	10.99
Test	8,850	12,232	1.00	17.56	10.95
Urban Dictionary					
Train	190,696	411,384	1.54	10.89	10.99
Valid	26,876	57,883	1.54	10.86	10.95
Test	26,875	38,371	1.68	11.14	11.50
Wikipedia					
Train	151,995	887,455	2.10	18.79	5.89
Valid	8,361	44,003	2.11	19.21	6.31
Test	8,397	57,232	2.10	19.02	6.94

Table 1: Statistics of the word/phrase description datasets.

Corpus	Domain	Inputs	Cov. emb.
WordNet	General	words	100.00%
Oxford Dictionary	General	words	83.04%
Urban Dictionary	Internet slang	phrases	21.00%
Wikipedia	Proper nouns	phrases	26.79%

Table 2: Domains, expressions to be described, and the coverage of pre-trained embeddings of the expressions to be described.

word. We split this dataset to obtain Train, Validation, and Test sets. If a word has multiple definitions/examples, we treat them as different entries. Note that the words are mutually exclusive across the three sets. The only difference between our dataset and theirs is that we extract the tuples only if the words have their usage examples in WordNet. Since not all entries in WordNet have usage examples, our dataset is a small subset of Noraset et al. (2017).

In addition to WordNet, we use the Oxford Dictionary following Gadetsky et al. (2018), the Urban Dictionary following Ni and Wang (2017) and our Wikipedia dataset described in the previous section. Table 1 and Table 2 show the properties and statistics of the four datasets, respectively.

To simulate a situation in a real application where we might not have access to global context for the target phrases, we did not train domain-specific word embeddings on each dataset. Instead, for all of the four datasets, we use the same

	Global	Local	I-Attn.	LOG-CaD
# Layers of Enc-LSTMs	-	2	2	2
Dim. of Enc-LSTMs	-	600	600	600
Dim. of Attn. vectors	-	300	300	300
Dim. of input word emb.	300	-	300	300
Dim. of char. emb.	160	160	-	160
# Layers of Dec-LSTMs	2	2	2	2
Dim. of Dec-LSTMs	300	300	300	300
Vocabulary size	10k	10k	10k	10k
Dropout rate	0.5	0.5	0.5	0.5

Table 3: Hyperparameters of the models

pre-trained CBOW[10] vectors trained on Google news corpus as global context following previous work (Noraset et al., 2017; Gadetsky et al., 2018). If the expression to be described consists of multiple words, its phrase embedding is calculated by simply summing up all the CBOW vectors of words in the phrase, such as "sonic" and "boom." (See Figure 1). If pre-trained CBOW embeddings are unavailable, we instead use a special **[UNK]** vector (which is randomly initialized with a uniform distribution) as word embeddings. Note that our pre-trained embeddings only cover 26.79% of the words in the expressions to be described in our Wikipedia dataset, while it covers all words in WordNet dataset (See Table 2). Even if no reliable word embeddings are available, all models can capture the character information through character-level CNNs (See Figure 1).

Models We implemented four methods: (1) **Global** (Noraset et al., 2017), (2) **Local** (Ni and Wang, 2017) with CNN, (3) **I-Attention** (Gadetsky et al., 2018), and our proposed model, (4) **LOG-CaD**. The **Global** model is our reimplementation of the best model (S + G + CH) in Noraset et al. (2017). It can access the global context of a phrase to be described, but has no ability to read the local context. The **Local** model is the reimplementation of the best model (dual encoder) in Ni and Wang(2017). In order to make a fair comparison of the effectiveness of local and global contexts, we slightly modify the original implementation by Ni and Wang(2017); as the character-level encoder in the **Local** model, we adopt CNNs that are exactly the same as the other two models instead of the original LSTMs.

The **I-Attention** is our reimplementation of the best model (S + I-Attention) in Gadetsky

---

[10] GoogleNews-vectors-negative300.bin.gz at https://code.google.com/archive/p/word2vec/

Model	WordNet	Oxford	Urban	Wikipedia
Global	24.10	15.05	6.05	44.77
Local	22.34	17.90	9.03	52.94
I-Attention	23.77	17.25	10.40	44.71
LOG-CaD	**24.79**	**18.53**	**10.55**	**53.85**

Table 4: BLEU scores on four datasets.

Model	Annotated score
Local	2.717
LOG-CaD	**3.008**

Table 5: Averaged human annotated scores on Wikipedia dataset.

et al.(2018). Similar to our model, it uses both local and global contexts. Unlike our model, however, it does not use character information to predict descriptions. Also, it cannot directly use the local context to predict the words in descriptions. This is because the **I-Attention** model indirectly uses the local context only to disambiguate the phrase embedding x_{trg} as

$$x'_{trg} = x_{trg} \odot m, \qquad (16)$$

$$m = \sigma(W_m \frac{\sum_{i=1}^{I} \text{FFNN}(h_i)}{I} + b_m). \quad (17)$$

Here, the FFNN($\cdot$) function is a feed-forward neural network that maps the encoded local contexts h_i to another space. The mapped local contexts are then averaged over the length of the sentence X to obtain a representation of the local context. This is followed by a linear layer and a sigmoid function to obtain the soft binary mask m which can filter out the unrelated information included in global context. Finally, the disambiguated phrase embedding x'_{trg} is then used to update the decoder hidden state as

$$s_t = \text{LSTM}([y_{t-1}; x'_{trg}], s_{t-1}). \qquad (18)$$

All four models (Table 3) are implemented with the PyTorch framework (Ver. 1.0.0).[11]

Automatic Evaluation Table 4 shows the BLEU (Papineni et al., 2002) scores of the output descriptions. We can see that the **LOG-CaD** model consistently outperforms the three baselines in all four datasets. This result indicates that using both local and global contexts helps describe the unknown words/phrases correctly. While the

---

[11] http://pytorch.org/

Input:	waste	
Context:	#1	#2
	if the effort brings no compensating gain it is a **waste**	We **waste** the dirty water by channeling it into the sewer
Reference:	useless or profitless activity	to get rid of
Global:	to give a liquid for a liquid	
Local:	a state of being assigned to a particular purpose	to make a break of a wooden instrument
I-Attention:	a person who makes something that can be be be done	to remove or remove the contents of
LOG-CaD:	a source of something that is done or done	to remove a liquid

Table 6: Descriptions for a word in WordNet.

Input:	daniel o'neill	
Context:	#1	#2
	after being enlarged by publisher **daniel o'neill** it was reportedly one of the largest and most prosperous newspapers in the united states.	in 1967 he returned to belfast where he met fellow belfast artist **daniel o'neill**.
Reference:	american journalist	irish artist
Global:	american musician	
Local:	american publisher	british musician
I-Attention:	american musician	american musician
LOG-CaD:	american writer	british musician

Table 7: Descriptions for a phrase in Wikipedia.

I-Attention model also uses local and global contexts, its performance was always lower than the **LOG-CaD** model. This result shows that using local context to predict description is more effective than using it to disambiguate the meanings in global context.

In particular, the low BLEU scores of **Global** and **I-Attention** models on Wikipedia dataset suggest that it is necessary to learn to ignore the noisy information in global context if the coverage of pre-trained word embeddings is extremely low (see the third and fourth rows in Table 2). We suspect that the Urban Dictionary task is too difficult and the results are unreliable considering its extremely low BLEU scores and high ratio of unknown tokens in generated descriptions.

Input:	q			
Context:	#1	#2	#3	#4
	q-lets and co. is a fil-ipino and english infor-mative children 's show on **q** in the philippines .	she was a founding pro-ducer of the cbc radio one show " **q** " .	the q awards are the uk 's annual music awards run by the music maga-zine " **q** " .	charles fraser-smith was an author and one-time mis-sionary who is widely cred-ited as being the inspira-tion for ian fleming 's james bond quartermaster **q** .
Reference:	philippine tv network	canadian radio show	british music magazine	fictional character from james bond
Global:	american rapper			
Local:	television channel	television show show	magazine	american writer
I-Attention:	american rapper	american rapper	american rapper	american rapper
LOG-CaD:	television station in the philippines	television program	british weekly music journalism magazine	[unk] [unk]

Table 8: Descriptions for a word in Wikipedia.

Manual Evaluation To compare the proposed model and the strongest baseline in Table 4 (i.e., the **Local** model), we performed a human evalu-ation on our dataset. We randomly selected 100 samples from the test set of the Wikipedia dataset and asked three native English speakers to rate the output descriptions from 1 to 5 points as: 1) com-pletely wrong or self-definition, 2) correct topic with wrong information, 3) correct but incom-plete, 4) small details missing, 5) correct. The av-eraged scores are reported in Table 5. Pair-wise bootstrap resampling test (Koehn, 2004) for the annotated scores has shown that the superiority of **LOG-CaD** over the **Local** model is statistically significant ($p < 0.01$).

Qualitative Analysis Table 6 shows a word in the WordNet, while Table 7 and Table 8 show the examples of the entities in Wikipedia as examples. When comparing the two datasets, the quality of generated descriptions of Wikipedia dataset is sig-nificantly better than that of WordNet dataset. The main reason for this result is that the size of train-ing data of the Wikipedia dataset is 64x larger than the WordNet dataset (See Table 1).

For all examples in the three tables, the **Global** model can only generate a single description for each input word/phrase because it cannot access any local context. In the Wordnet dataset, only the **I-Attention** and **LOG-CaD** models can suc-cessfully generate the concept of "remove" given the context #2. This result suggests that consid-ering both local and global contexts are essential to generate correct descriptions. In our Wikipedia

dataset, both the **Local** and **LOG-CaD** models can describe the word/phrase considering its local con-text. For example, both the **Local** and **LOG-CaD** models could generate "american" in the descrip-tion for "daniel o'neill" given "united states" in context #1, while they could generate "british" given "belfast" in context #2. A similar trend can also be observed in Table 8, where **LOG-CaD** could generate the locational expressions such as "philippines" and "british" given the different con-texts. On the other hand, the **I-Attention** model could not describe the two phrases, taking into ac-count the local contexts. We will present an anal-ysis of this phenomenon in the next section.

6 Discussion

In this section, we present analyses on how the lo-cal and global contexts contribute to the descrip-tion generation task. First, we discuss how the lo-cal context helps the models to describe a phrase. Then, we analyze the impact of global context un-der the situation where local context is unreliable.

6.1 How do the models utilize *local* contexts?

Local context helps us (1) disambiguate polyse-mous words and (2) infer the meanings of un-known expressions. Can machines also utilize the local context? In this section, we discuss the two roles of local context in description generation.

Considering that the pre-trained word em-beddings are obtained from word-level co-occurrences in massive text, more information is mixed up into a single vector as the more senses the word has. While Gadetsky et al. (2018) de-

(a) Number of senses of the phrase.　　(b) Unknown words ratio in the phrase.　　(c) Length of the local context.
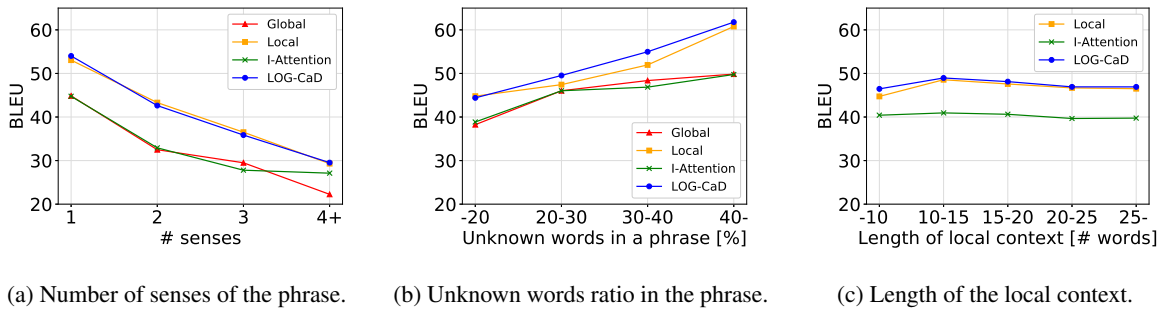
Figure 3: Impact of various parameters of a phrase to be described on BLEU scores of the generated descriptions.

signed the **I-Attention** model to filter out unrelated meanings in the global context given local context, they did not discuss the impact of the number of senses has on the performance of definition generation. To understand the influence of the ambiguity of phrases to be defined on the generation performance, we did an analysis on our Wikipedia dataset. Figure 3(a) shows that the description generation task becomes harder as the phrases to be described become more ambiguous. In particular, when a phrase has an extremely large number of senses, (i.e., #senses ≥ 4), the **Global** model drops its performance significantly. This result indicates that the local context is necessary to disambiguate the meanings in global context.

As shown in Table 2, a large proportion of the phrases in our Wikipedia dataset includes unknown words (i.e., only 26.79% of words in the phrases have their pre-trained embeddings). This fact indicates that the global context in this dataset is not fully reliable. Then our next question is, how does the lack of information from global context affect the performance of phrase description? Figure 3(b) shows the impact of unknown words in the phrases to be described on the performance. As we can see from the result, the advantage of **LOG-CaD** and **Local** models over **Global** and **I-Attention** models becomes larger as the unknown words increases. This result suggests that we need to fully utilize local contexts especially in practical applications where the phrases to be defined have many unknown words. Here, Figure 3(b) also shows a counterintuitive phenomenon that BLEU scores increase as the ratio of unknown words in a phrase increase. This is mainly because unknown phrases tend to be person names such as writers, actors, or movie directors. Since these entities have fewer ambiguities in categories, they can be described in extremely short sentences that are

easy for all four models to decode (e.g., "finnish writer" or "american television producer").

6.2　How do the models utilize *global* contexts?

As discussed earlier, local contexts are important to describe unknown expressions, but how about global contexts? Assuming a situation where we cannot obtain much information from local contexts (*e.g.*, infer the meaning of "boswellia" from a short local context "Here is a boswellia"), global contexts should be essential to understand the meaning. To confirm this hypothesis, we analyzed the impact of the length of local contexts on BLEU scores. Figure 3(c) shows that when the length of local context is extremely short ($l \leq 10$), the **LOG-CaD** model becomes much stronger than the **Local** model. This result indicates that not only local context but also global context help models describe the meanings of phrases.

7　Related Work

In this study, we address a task of describing a given phrase with its context. In what follows, we explain existing tasks that are related to our work.

Our task is closely related to word sense disambiguation (WSD) (Navigli, 2009), which identifies a pre-defined sense for the target word with its context. Although we can use it to solve our task by retrieving the definition sentence for the sense identified by WSD, it requires a substantial amount of training data to handle a different set of meanings of each word, and cannot handle words (or senses) which are not registered in the dictionary. Although some studies have attempted to detect novel senses of words for given contexts (Erk, 2006; Lau et al., 2014), they do not provide definition sentences. Our task avoids these difficulties in WSD by directly generating descriptions for

phrases or words. It also allows us to flexibly tailor a fine-grained definition for the specific context.

Paraphrasing (Androutsopoulos and Malakasiotis, 2010; Madnani and Dorr, 2010) (or text simplification (Siddharthan, 2014)) can be used to rephrase words with unknown senses. However, the target of paraphrase acquisition are words/phrases with no specified context. Although a few studies (Connor and Roth, 2007; Max, 2009; Max et al., 2012) consider sub-sentential (context-sensitive) paraphrases, they do not intend to obtain a definition-like description as a paraphrase of a word.

Recently, Noraset et al. (2017) introduced a task of generating a definition sentence of a word from its pre-trained embedding. Since their task does not take local contexts of words as inputs, their method cannot generate an appropriate definition for a polysemous word for a specific context. To cope with this problem, Gadetsky et al. (2018) proposed a definition generation method that works with polysemous words in dictionaries. They presented a model that utilizes local context to filter out the unrelated meanings from a pre-trained word embedding in a specific context. While their method use local context for disambiguating the meanings that are mixed up in word embeddings, the information from local contexts cannot be utilized if the pre-trained embeddings are unavailable or unreliable. On the other hand, our method can fully utilize the local context through an attentional mechanism, even if the reliable word embeddings are unavailable.

The most related work to this paper is Ni and Wang (2017). Focusing on non-standard English phrases, they proposed a model to generate the explanations solely from local context. They followed the strict assumption that the target phrase was newly emerged and there was only a single local context available, which made the task of generating an accurate and coherent definition difficult. Our proposed task and model are more general and practical than Ni and Wang (2017); where (1) we use Wikipedia, which includes expressions from various domains, and (2) our model takes advantage of global contexts if available.

Our task of describing phrases with its context is a generalization of the three tasks (Noraset et al., 2017; Ni and Wang, 2017; Gadetsky et al., 2018), and the proposed method utilizes both local and global contexts of an expression in question.

8 Conclusions

This paper sets up a task of generating a natural language description for an unknown phrase with a specific context, aiming to help us acquire unknown word senses when reading text. We approached this task by using a variant of encoder-decoder models that capture the given local context with the encoder and global contexts with the decoder initialized by the target phrase's embedding induced from massive text. We performed experiments on three existing datasets and one newly built from Wikipedia and Wikidata. The experimental results confirmed that the local and global contexts complement one another and are both essential; global contexts are crucial when local contexts are short and vague, while the local context is important when the target phrase is polysemous, rare, or unseen.

As future work, we plan to modify our model to use multiple contexts in text to improve the quality of descriptions, considering the "one sense per discourse" hypothesis (Gale et al., 1992). We will release the newly built Wikipedia dataset and the experimental codes for the academic and industrial communities at https://github.com/shonosuke/ishiwatari-naacl2019 to facilitate the reproducibility of our results and their use in various application contexts.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

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 (ICLR)*.

Yuzhen Chen. 2012. Dictionary use and vocabulary learning in the context of reading. *International Journal of Lexicography*, 25(2):216–247.

Michael Connor and Dan Roth. 2007. Context sensitive paraphrasing with a global unsupervised classifier. In *Proceedings of the 18th European Conference on Machine Learning (ECML)*, pages 104–115.

Katrin Erk. 2006. Unknown word sense detection as outlier detection. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, pages 128–135.

Carol A. Fraser. 1998. The role of consulting a dictionary in reading and vocabulary learning. *Canadian Journal of Applied Linguistics*, 2(1-2):73–89.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional generators of words definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Short Papers*, pages 266–271.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language, HLT*, pages 233–237.

Felix A. Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with lstm. In *Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN)*, pages 850–855.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 259–270.

Minh-Thang Luong and Christopher D. Manning. 2016. Achieving open vocabulary neural machine translation with hybrid word-character models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1054–1063.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics*, 36(3):341–387.

Aurélien Max. 2009. Sub-sentencial paraphrasing by contextual pivot translation. In *Proceedings of the 2009 Workshop on Applied Textual Inference*, pages 18–26.

Aurélien Max, Houda Bouamor, and Anne Vilnat. 2012. Generalizing sub-sentential paraphrase acquisition across original signal type of text pairs. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 721–731.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM COMPUTING SURVEYS*, 41(2):1–69.

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 413–417.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, pages 3259–3266.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318.

Advaith Siddharthan. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics*, 165(2):259–298.