Learning to Describe Phrases with Local and Global Contexts

Shonosuke Ishiwatari† Hiroaki Hayashi‡ Naoki Yoshinaga§
Graham Neubig‡ 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, 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. At first, we attempt to figure out the meaning of those expressions from their context, and ultimately we may consult a dictionary for their definitions. However, rarely-used senses or emerging entities are not always covered by the hand-crafted definitions in existing dictionaries, which can cause problems in text comprehension. This paper undertakes a task of describing (or defining) a given expression (word or phrase) based on its usage context, and presents a novel neural-network generator for expressing its meaning as a natural language description. Experimental results on four datasets (including WordNet, Oxford and Urban Dictionaries, and Wikipedia) demonstrate the effectiveness of our method over previous methods for definition generation (Noraset et al., 2017; Gadetsky et al., 2018) and non-standard English explanation (Ni and Wang, 2017).

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 are unsure of. In such cases, we may first try to examine other usages of the same expression in the text, in order to infer its meaning from this context. Failing to do so, we may consult a dictionary, and in the case of polysemous words, choose an appropriate meaning based on the context. Acquiring novel word senses via dictionary definitions is known to be more effective than contextual guessing (Fraser, 1998; Chen, 2012). However, very often, hand-crafted dictionaries do not contain definitions for rare or novel phrases/words, and we eventually give up on understanding them completely, leaving us with only a shallow reading of the text.

There are several natural language processing (NLP) tasks that can roughly address this problem of unfamiliar word senses, all of which are incomplete in some way. Word sense disambiguation (WSD) can basically only handle words (or senses) that are registered in a dictionary a priori. Paraphrasing can suggest other ways of describing a word while keeping its meaning, but those paraphrases are generally context-insensitive and may not be sufficient for understanding.

To address this problem, Ni and Wang (2017) has proposed a task of describing a phrase in a given context. However, they follow the strict assumption that the target phrase is unknown 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 its word embedding induced from massive text, followed by Gadetsky et al. (2018) that refers
to a local context to define a polysemous word with a local context by choosing relevant dimensions of their embeddings. Although these research efforts revealed that both local and global contexts of words are useful in generating their definitions, none of these studies exploited both local and global contexts directly.

In this study, we tackle a task of describing (defining) a phrase when given its local context as \( \text{(Ni and Wang, 2017)} \), while allowing access to other usage examples via word embeddings trained from massive text (global contexts) \( \text{(Noraset et al., 2017; Gadetsky et al., 2018)} \). We present LOG-Cad, a neural network-based description generator (Figure 1) to directly solve this task. Given a word with its context, our generator takes advantage of the target word’s embedding, pre-trained from massive text (global contexts), while also encoding the given local context, combining both to generate a natural language description. 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 crucial when the target phrase is polysemous, rare, or unseen.

Considering various contexts where we need definitions of phrases, we evaluated our method with four datasets including WordNet \( \text{(Noraset et al., 2017)} \) for general words, the Oxford dictionary \( \text{(Gadetsky et al., 2018)} \) for polysemous words, Urban Dictionary \( \text{(Ni and Wang, 2017)} \) for rare idioms or slangs, and a newly-created Wikipedia dataset for entities. Experimental results demonstrate the effectiveness of our method against the three baselines stated above \( \text{(Noraset et al., 2017; Ni and Wang, 2017; Gadetsky et al., 2018)} \).

Our contributions are as follows:

- **Empirical results are strong**: this method achieves the state-of-the-art performance for our new dataset and the three existing datasets used in the related studies \( \text{(Noraset et al., 2017; Ni and Wang, 2017; Gadetsky et al., 2018)} \). We will release the dataset to the public as well as all of the code to promote reproducibility of the experiments.

2 Context-aware Description Generation

In what follows, we define our task of describing a phrase or word in a specific context. Given expression \( X_{trg} \) with its context \( X = \{x_1, \ldots, x_T\} \), our task is to output a description \( Y = \{y_1, \ldots, y_T\} \). Here, \( X_{trg} \) can be a single word or a short phrase and is included in \( X \). \( Y \) is a definition-like concrete and concise phrase/sentence that describes the expression \( X_{trg} \).

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

\[
p(Y|X, X_{trg}) = \prod_{t=1}^{T} p(y_t|y_1, \ldots, y_{t-1}, X, X_{trg}).
\]

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

We propose LOG-Cad, a neural model that generates the description of a given phrase or word by using its local and global contexts. In the rest of this section, we first describe our idea of utilizing local and global contexts in the description generation task, then present the details of our model.

Local & Global Contexts for Description Generation

In this paper, we refer to the explicit contextual information included in a single sentence as “local context,” and the implicit contextual information in the word/phrase embedding trained in an unsupervised manner on large-scale corpora as “global context.” Previous work
on the definition generation task (Noraset et al., 2017) has shown that global contexts can be useful clues when generating definitions of unknown words. The intuition behind their method is that words with similar meanings tend to have similar definitions in a dictionary. This can be seen as an extension of the Distributional Hypothesis (Harris, 1954; Firth, 1957), which states words that share semantic meanings tend to appear in similar contexts. Additionally, work on the WSD task (Navigli, 2009), novel sense detection (Erk, 2006; Lau et al., 2014), and the non-standard word explanation task (Ni and Wang, 2015) have revealed that local contexts surrounding the word can help disambiguate its sense. Based on these studies, we propose to incorporate both local and global contexts to describe an unknown expression.

**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 consists of two modules: 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 (Noraset et al., 2017) to dynamically control how the global and local contexts influence the generation of the description.

We use bi-directional and uni-directional LSTMs (Hochreiter and Schmidhuber, 1997) as our context encoder and description decoder (Figure 1), respectively. Given a sentence $X$ and a phrase $X_{trg}$, the context encoder generates a sequence of continuous vectors $H = \{h_1, \ldots, h_T\}$ as

$$h_i = \text{Bi-LSTM}(h_{i-1}, x_i), \quad (2)$$

where $x_i$ denotes the word embedding of word $x_i$. Then, the description decoder computes the conditional probability of a description $Y$ with Eq. (1), which can be approximated with another LSTM as

$$s_t = \text{LSTM}(y_{t-1}, s'_{t-1}), \quad (3)$$

$$d_t = \text{ATTENTION}(H, s_t), \quad (4)$$

$$c_{trg} = \text{CNN}(X_{trg}), \quad (5)$$

$$s'_t = \text{GATE}(s_t, x_{trg}, c_{trg}, d_t), \quad (6)$$

$$p(y_t | y_{<t}, X_{trg}) = \text{softmax}(W_s s'_t + b_s), \quad (7)$$

where $s_t$ is a hidden state of the decoder LSTM, and $y_{t-1}$ is a jointly-trained word embedding of the previous output word $y_{t-1}$.

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

$$d_t = \sum_{i=1}^{T} \alpha_i h_i, \quad (8)$$

$$\alpha_i = \text{softmax}(U_h h_i^T U_s s_t), \quad (9)$$

where $U_h$ and $U_s$ are matrices that map the encoder and decoder hidden states into a common space, respectively.

In order to capture prefixes and suffixes in $X_{trg}$, we construct character-level CNNs (Eq. (5)) 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 kernels of length 2-6 and size 10, 30, 40, 40, 40 respectively with a stride of 1 to obtain a 160-dimensional vector $c_{trg}$.

In addition to the local context and the character-information, we also utilize the global context obtained from massive text. We achieve this by two different strategies proposed by Noraset et al. (2017). First, we feed phrase embedding $x_{trg}$ to initialize the decoder as

$$y_0 = x_{trg}. \quad (10)$$

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

As described in the previous section, we use both local and global contexts. In order to capture the interaction between two contexts and the description decoder, we adopt a GATE(·) function (Eq. (6)) that updates the LSTM output $s_t$ to $s'_t$ depending on the global context $x_{trg}$, local
Table 1: Statistics of the word/phrase description datasets.

| Corpus | # Phrases | # Entries | Length of Context | Length of Description |
|--------|-----------|-----------|-------------------|-----------------------|
| WordNet |          |           |                   |                       |
| Train  | 7,938     | 13,883    | 5.81              | 6.61                  |
| Valid  | 998       | 1,752     | 5.64              | 6.61                  |
| Test   | 1,001     | 1,775     | 5.77              | 6.85                  |
| Oxford Dictionary | |           |                   |                       |
| Train  | 33,128    | 97,855    | 17.74             | 11.02                 |
| Valid  | 8,867     | 12,232    | 17.80             | 10.99                 |
| Test   | 8,850     | 12,232    | 17.56             | 10.95                 |
| Urban Dictionary | |           |                   |                       |
| Train  | 190,696   | 411,384   | 10.89             | 10.99                 |
| Valid  | 26,876    | 57,883    | 10.86             | 10.95                 |
| Test   | 26,875    | 38,371    | 11.14             | 11.50                 |
| Wikipedia | |           |                   |                       |
| Train  | 151,995   | 887,455   | 18.79             | 5.89                  |
| Valid  | 8,361     | 44,003    | 19.21             | 6.31                  |
| Test   | 8,390     | 13,883    | 19.02             | 6.94                  |

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

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

context \(d_t\), and character-level information \(c_{trg}\) as

\[
f_t = [x_{trg}; d_t; c_{trg}]
\]

(11)

\[
z_t = \sigma(W_z[f_t; s_t] + b_z),
\]

(12)

\[
r_t = \sigma(W_r[f_t; s_t] + b_r),
\]

(13)

\[
s_{t} = \tanh(W_s[r_t; f_t; s_t] + b_s),
\]

(14)

\[
s'_t = (1 - z_t) \odot s_t + z_t \odot \tilde{s}_t,
\]

(15)

where \(\sigma(\cdot)\), \(\odot\) and \(;\) denote sigmoid function, element-wise multiplication, and vector concatenation, respectively. \(W\)s and \(b\)s are weight matrices and bias terms. Here, the update gate \(z_t\) controls how much the original hidden state \(s_t\) is to be changed, and the reset gate \(r_t\) controls how much the information from \(f_t\) contributes to word generation at each time step.

4 Wikipedia dataset

One of our goals is to describe infrequent/rare words and phrases such as proper nouns in a variety of domains depending on their surrounding context. However, among the three existing datasets, WordNet and Oxford dictionary mainly target the descriptions of relatively common words, and thus are non-ideal test beds for this goal. On the other hand, although the Urban Dictionary dataset contains descriptions of rarely-used phrases as well, the domain of its targeted words and phrases is limited to Internet slang.

Therefore, in order to confirm that our model can generate the description of rarely-used phrases as well as words, we constructed a new dataset for context-aware phrase description generation from Wikipedia\(^1\) and Wikidata\(^2\) which contain a wide variety of entity descriptions with contexts. Table 1 and Table 2 show the properties and statistics of the new dataset and the three existing datasets, respectively.

The overview of the data extraction process is shown in Figure 2. Similarly to the WordNet dataset, each entry in the dataset consists of (1) a phrase, (2) its description, and (3) context (a sentence). For preprocessing, we applied Stanford Tokenizer\(^3\) 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 one of the major advantages 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 (Miller, 1995) and Oxford Dictionary,\(^4\) phrases in Urban Dictionary\(^5\) and Wikidata.
Figure 2: Context-aware description dataset extracted from Wikipedia and Wikidata.

Datasets To evaluate our model on the word description task on WordNet, we followed Noraset et al. (2017) and extracted data from WordNet\(^7\) using the dict-definition\(^8\) toolkit. Each entry in the data consists of three elements: (1) a word, (2) its definition, and (3) a usage example of the 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) (see Table 1).

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.

In order to control the experiments on four datasets, we use the same pre-trained CBOW\(^9\) vectors as global context following (Noraset et al., 2017). 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 including three baselines: (1) Global, (2) Local, (3) I-Attention, and our proposed model, (4) LOG-CaD. The Global model is our reimplementation of the strongest model (S + G + CH) in (Noraset et al., 2017). It can access the embedding (global context) of the phrase to be described,
but has no ability to read the usage examples (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 of (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 et al., 2018). Similar to our model, it uses both local and global contexts. Unlike our model, however, their model 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 filter out unrelated information in phrase embeddings. All four models (Table 3) are implemented with the PyTorch framework.\(^{10}\)

### Results

Table 4 shows the performance of the models. 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 words correctly. While the **I-Attention** model also uses local and global contexts, its performance was always lower than the **LOG-CaD** model. This phenomenon 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.

Table 5 through Table 8 show the input/output examples of the words/phrases in WordNet and Wikipedia datasets, respectively. When comparing the two datasets shown in the two tables, the quality of generated descriptions of Wikipedia dataset is significantly better than that of WordNet dataset. The main reason for this result is that the size of training data of the Wikipedia dataset is 64x larger than the WordNet dataset (Table 1).

For all examples in both datasets in Table 5 through Table 8, the **Global** model can only generate a single description for each input phrase because it cannot access any local context. In the Wikipedia dataset, both the **Local** and **LOG-CaD** models can describe the word/phrase considering its local context. For example, both the **Local** and **LOG-CaD** models could generate “american” in the description for “daniel o’neill” given “united states” in Context #1, while they could generate “british” given “belfast” in Context #2. On the other hand, the **I-Attention** model could not describe the two phrases, taking into account the local contexts. We will present an analysis of this phenomenon in the next section.

### 6 Discussion

In this section, we present some analyses on how the local and global contexts contribute to the description generation task. First, we discuss how the local context helps the models to describe a phrase. Then, we analyze the impact of global context under the situation where local context is unreliable.

| 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.

| Input: daniel o’neill |
|-----------------------|
| Context: |
| #1                  | after being enlarged by publisher daniel o’neill it was reportedly one of the largest and most prosperous newspapers in the united states. |
| #2                  | 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 5: The generated descriptions for “daniel o’neill” in Wikipedia.
Input: q

Context:

#1 q-lets and co. is a filipino and english informative children’s show on q in the philippines.

#2 she was a founding producer of the cbc radio one show “q”.

#3 the q awards are the uk’s annual music awards run by the music magazine “q”.

#4 charles fraser-smith was an author and one-time missionary who is widely credited as being the inspiration 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

LOG-CaD:

television station in the philippines

television program
british weekly music journalism magazine
[unk] [unk]

Table 6: The generated descriptions for “q” in Wikipedia.

Input: gracious

Context:

#1 gracious even to unexpected visitors

#2 thanks to the gracious gods

Reference:

characterized by charm, good taste, and generosity of spirit
disposed to bestow favors

Global:

gracious and gracious

Local:

to or relating to or characteristic of a profession

I-Attention:

a feeling of thankfulness

LOG-CaD:

to be given to a particular purpose

Table 7: The generated descriptions for “gracious” in WordNet.

Input: waste

Context:

#1 if the effort brings no compensating gain it is a waste

#2 We waste the dirty water by channeling it into the sewer

Reference:

useless or profitless activity
to give a liquid for a liquid

to make a break of a wooden instrument purpose

Global:

to give a liquid for a liquid

Local:

a state of being assigned to a particular purpose

I-Attention:

a person who makes something that can be done

LOG-CaD:

a source of something that is done or done

Table 8: The generated descriptions for “waste” in WordNet.

6.1 How do the models utilize local contexts?

Local context helps us (1) disambiguate polysemyous words and (2) infer the meanings of unknown expressions. In this section, we will discuss the two roles of local context.

Considering that the pre-trained word embeddings are obtained from word-level co-occurrences in a massive text, more information is mixed up into a single vector as the more senses the word has. While Gadetsky et al. (2018) designed the I-Attention model to filter out unrelated meanings in the global context given local context, they did not discuss the impact 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 have their pre-trained embeddings). This fact indicates
that the global context in the dataset is extremely noisy. 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 \textit{LOG-CaD} and \textit{Local} models over \textit{Global} and \textit{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.

6.2 How do the models utilize \textit{global} contexts?

As discussed earlier, local contexts are important to describe 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 \textit{LOG-CaD} model becomes much stronger than the \textit{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/word with its context. In what follows, we explain several tasks that are related to our task.

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 with their contexts. 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 (or phrases) with no specified context. Although several 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) have proposed a definition generation method that works with polysemous words in dictionaries. They present 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 only for dis-
ambiguating 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.

Focusing on non-standard English words (or phrases), Ni and Wang (2017) generated their explanations solely from sentences with those words. Their model does not take advantage of global contexts (word embeddings induced from massive text) as was used in Noraset et al. (2017).

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

8 Conclusions

This paper sets up a task of generating a natural language description for a word/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 by an encoder and global contexts by the target word’s embedding induced from massive text. Experimental results on three existing datasets and one novel dataset built from Wikipedia dataset confirmed that the use of both local and global contexts is the key to generating appropriate context-sensitive description in various situations.

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).

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.

John R. Firth. 1957. A synopsis of linguistic theory. Studies in Linguistic Analysis, pages 1–32.

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), 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.

Zellig S. Harris. 1954. Distributional structure. Word, 10:146–162.

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

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-sentential 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 Comput. Surv., 41(2):10:1–10:69.

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.

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