Automatic Paraphrasing via Sentence Reconstruction and Back-translation

Abstract
Paraphrase generation plays key roles in NLP tasks such as question answering, machine translation, and information retrieval. In this paper, we propose a novel framework for paraphrase generation. It simultaneously decodes the output sentence using a pretrained wordset-to-sequence model and a back-translation model. We evaluate this framework on Quora, WikiAnswers, MSCOCO and Twitter, and show its advantage over previous state-of-the-art unsupervised methods and distantly-supervised methods by significant margins on all datasets. For Quora and WikiAnswers, our framework even performs better than some strongly supervised methods with domain adaptation. Further, we show that the generated paraphrases can be used to augment the training data for machine translation to achieve substantial improvements.

1 Introduction
The paraphrase of a sentence retains its meaning but makes different choices of words and expressions than the original form. Paraphrase generation plays an important role in many downstream tasks, such as question answering, machine translation, and information retrieval [Hu et al., 2019a].

Most existing parallel datasets for paraphrase generation are domain-specific. Quora and WikiAnswers [Fader et al., 2013] datasets only contain questions; sentences in MSCOCO [Lin et al., 2014] dataset are mostly descriptions for physical objects since they are the captions of images; and PPDB [Ganitkevitch et al., 2013] contains phrases rather than sentences. The performance of a model trained with these domain-specific parallel data declines seriously when it is used in another domain [Li et al., 2019].

Many efforts were made to solve this domain adaptation problem. These efforts are roughly divided into three directions: unsupervised fine-tuning for supervised model, unsupervised methods based on word/phrase replacement, and distantly-supervised methods based on bilingual data. Li et al. [2019] chose to fine-tune the supervised model with non-parallel in-domain data, but the performance of their model decreases a lot when the domain span is large. Liu et al. [2019] and Miao et al. [2019] used unsupervised methods to generate paraphrases, but their models are mostly based on the variation of words and phrases and can hardly change the structure of the whole sentence. Wieting and Gimpel [2017] generated paraphrases with a back-translation model, but the existing translation models are sometimes not very accurate, which also affects the performance of their method. Liu et al. [2020] use bilingual data to train an unsupervised model, but their improvement is mainly brought about by the follow-up supervised fine-tuning.

In this paper, we propose a novel paraphrase generation framework that does not require any parallel paraphrase data and can be applied in any domain. In our framework, two kinds of underlying semantics are extracted from the original sentence and are recombined into a new sentence through a hybrid decoder.

The first kind of underlying semantics is represented by a word set, which is inspired by the Denoising Auto-Encoder (DAE) [Vincent et al., 2008]. A bag of words is a great carrier of information, as it communicates the central idea without syntactic constraints. People can produce different sentences with similar meaning from the same set of words. Table 1 shows an example of such paraphrase sentences. We construct a word set from the original sentence and extend the word set into a complete sentence with a set-to-sequence (set2seq) model, which is adapted from the well-known sequence-to-sequence (seq2seq) model by ignoring the sequential information from the input sequence.

<table>
<thead>
<tr>
<th>word set: (man, sit, bike, bench)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A man is sitting on a bench next to a bike</td>
</tr>
<tr>
<td>A man is sitting on a bench next to a bicycle</td>
</tr>
<tr>
<td>A man sits on a bench by a bike</td>
</tr>
<tr>
<td>Man sitting on a bench near a personal bicycle</td>
</tr>
<tr>
<td>A man is sitting on a bench with a bike</td>
</tr>
</tbody>
</table>

Table 1: An example of paraphrases formed from the same set of words (enclosed in the parentheses).

The second carrier of semantics is the translation of the original sentence into another language. Semantics is preserved but syntactic perturbations are added when the translation is then translated back to the original language. This is known as back-translation [Wieting and Gimpel, 2017].

1Considering that training the back-translation model requires
how can i make money online with free of cost?

Stopwords Removal & Random Synonym Replacement

Set to Sequence (Encoder)

Set to Sequence (Decoder)

Input Token Sequence

Word Set in L1

Output Token Sequence

The next token is “free”

Choose the token with the highest score to be the next one

Vocabulary:

Set2seq Score: 0.53 0.11 0.09 0.02 ... 

Back-trans Score: 0.32 0.19 0.12 0.05 ... 

Weighted Sum: 0.46 0.14 0.10 0.03 ...

Hybrid Decoding

L2 to L1 Translation Model

L2 to L1 Translation Model (Encoder)

L1 to L2 Translation Model

Translation in L2

Figure 1: Our Paraphrasing Framework

The above two types of semantics are complementary. The back-translation makes up for the missing information in the set2seq model, such as sequential information. The set2seq model gives the back-translation model some lexical hints, and makes the translation result more accurate. We thus integrate the decoding parts of the set2seq model and the back-translation model to jointly generate paraphrases.

We evaluate our framework on four paraphrasing datasets, namely Quora, WikiAnswers, MSCOCO, and Twitter [Lan et al., 2017], and achieve the state-of-the-art accuracies compared to existing models trained with non-parallel data.

We also train the set2seq model on a big common-domain dataset and test it on these four datasets, and still obtain decent results. We call the set2seq model trained from the big common-domain dataset “set2seq-common”, and can apply it to any domain when there is no in-domain data to train a set2seq model.

Finally, we propose an application of our paraphrase generator: to augment the training data of a neural machine translation (NMT) model between low-resource languages and English. We paraphrase the English sentences in the parallel training pairs with set2seq-common and improve the BLEU score of X-to-English translation by 1.53 to 2.17, where X is a low-resource language.

In summary, the main contributions of this work are:

• We are the first to apply the set2seq model to the task of paraphrase generation by combining it with a back-translation model through a hybrid decoder.

• The framework proposed by us achieve state-of-the-art accuracies on four benchmark datasets compared with existing methods.

• We apply our method to augment the training data of low-resource translation tasks and obtain significant improvement in translation quality.

2 Approach

In this section, we describe our framework. We first give an overview and then describe the detailed components of the framework.

2.1 Overview

The set2seq model and the two translation models used in back-translation are trained separately, and our framework is designed for use during inference time only. Figure 1 shows the architecture for our framework, which is divided into two major components and two major phases. The two components are sentence reconstruction based on word set, and back-translation. The two phases are information extraction and paraphrase generation.

Suppose the original sentence is in language $L_1$ and the back-translation is via language $L_2$. During information extraction phase, given an input sequence of tokens $X = [x_1, x_2, \cdots]$, we process it in two different approaches to extract two different representations of the underlying semantics: a word set and a translation in language $L_2$. For the former, we construct a word set $WS = \{w_1, w_2, \cdots\}$. For the latter, we use a $L_1$-$L_2$ translation model to get a sequence of translated tokens $Z = [z_1, z_2, \cdots]$ in $L_2$.

In the paraphrase generation phase, we employ a hybrid decoder which takes inputs from two separate encoders, one

intensive computing resources, we will open-source these models and the codes when the paper is published.
We use the word set constructor to extract a word set from the original sentence. To ensure accuracy and diversity of sentences generated from the word set, the word set constructor tries to strike a balance between both content preservation and lexical variation.

For content preservation, we could select informative words from the original sentence by either removing stopwords or retaining high-IDF words to build the keywords set $WS$, which will be passed to the next stage. Here, we choose to remove stopwords, the reason for which will be explained in the result analysis in Section 3.5.

To increase the lexical diversity of the generated paraphrase, each word in $WS$ is randomly replaced with one of its synonyms using WordNet [Miller, 1995], and optionally, itself. This process is known as “random replacement”. We obtain $WS$ after this step. BERT based methods, instead of WordNet, can also be used to generate synonyms. They are not used because: i) we have to generate synonyms for every single word in the training set, and it is too computational expensive if we use BERT; and ii) WordNet is good enough for generating high-quality word sets.

2.2 Word Set Constructor

We use the word set constructor to extract a word set from the original sentence. To ensure accuracy and diversity of sentences generated from the word set, the word set constructor tries to strike a balance between both content preservation and lexical variation.

For content preservation, we could select informative words from the original sentence by either removing stopwords or retaining high-IDF words to build the keywords set $WS$, which will be passed to the next stage. Here, we choose to remove stopwords, the reason for which will be explained in the result analysis in Section 3.5.

To increase the lexical diversity of the generated paraphrase, each word in $WS$ is randomly replaced with one of its synonyms using WordNet [Miller, 1995], and optionally, itself. This process is known as “random replacement”. We obtain $WS$ after this step. BERT based methods, instead of WordNet, can also be used to generate synonyms. They are not used because: i) we have to generate synonyms for every single word in the training set, and it is too computational expensive if we use BERT; and ii) WordNet is good enough for generating high-quality word sets.

2.3 Set-to-Sequence

A set2seq model consists of an encoder and a decoder, similar to a seq2seq model. However, instead of taking a sequence as the input, the input of a set2seq model is a set of tokens with no sequential information.

To train a set2seq model, we prevent the encoder to do serial processing for the input set. RNN-based models are inappropriate for this purpose due to their recurrent nature. Therefore we use a transformer-based model. In transformer, the sequential information of the input sequence is captured in the position encoding. We use a transformer but omit the position encoding in the encoder as the set2seq model.

We train set2seq with word set $WS$ as the input and original sentence $X$ as the output. This training data is automatically created and thus the training process is considered self-supervised. Specifically, given a set of words $WS = \{w_1, w_2, \cdots\}$, the set2seq model does the following steps in a single layer while encoding:

\[
\overline{T}_i = \text{LayerNorm}(\text{MultiAttn}(h_i)) + h_i \\
h_{i+1} = \text{LayerNorm}(\text{FF}(h_i)) + \overline{T}_i,
\]

where $h_{i+1}$ is the output of layer $i$ and $h_0$ is the embedding of tokens in $WS$.

2.4 Hybrid Decoding

A hybrid decoder can take the hidden states of multiple encoders as input and generate a single output sequence based on the information from all hidden states.

As we mentioned before, we divide the framework into two components, the set2seq model and the back-translation model, and obtain two hidden states $H_{ws}$ and $H_{bt}$.

Assume that our vocabulary is $V = \{v_1, v_2, \cdots, v_D\}$ with $D$ different tokens. In decoding step $t$, the decoder of the set2seq and the $L_2$-$L_1$ translation model can give the probability of $v$ being the next token individually. Supposing we already generated $t-1$ tokens $y_1, y_2, \cdots, y_{t-1}$, the next token $y_t$ to be generated is given by the following equation:

\[
y_t = \arg \max_{v \in V} \left( P_{bt}(v | y_{1:t-1}, H_{bt}) + \lambda \cdot P_{ws}(v | y_{1:t-1}, H_{ws}) \right)
\]

Here $P_{ws}$ and $P_{bt}$ are the probabilities of $v_t$ being the next token calculated by the decoder of the set2seq model and the $L_2$-$L_1$ translation model respectively, and $\lambda$ is the hyper-parameter to balance the weight between the two probabilities.

3 Experimental Results

In this section, we first introduce the experimental setup, including datasets, baselines, evaluation metrics, and implementation details. Then, we show the results the competing methods. Finally, we analyze the results from different aspects.

3.1 Datasets

We evaluate our framework on four different datasets, namely Quora, WikiAnswers, MSCOCO, and Twitter. Following Liu et al. [2019], we randomly choose 20K parallel paraphrase pairs as the test set and 3K parallel paraphrase pairs as the validation set for Quora, WikiAnswers, and MSCOCO.

Training with In-domain Data. We randomly sample the remaining parallel paraphrase pairs and pick one sentence from each pair to construct the non-parallel training data. The number of selected sentences is the same as the work by Liu et al. [2019], which is 400K for Quora, which is 500K for WikiAnswers, 320K for MSCOCO and 110K for Twitter.

Training with Common-Domain Data. When there is no sufficient available target-domain non-parallel data, it is hard to train unsupervised models or fine-tune supervised models in the target-domain. Our solution is to train the set2seq model with a big common-domain dataset and apply it to the target-domain. We name the model “set2seq-common”. We test the performance of our framework with set2seq-common on four datasets to show the generality of our framework. Further, we apply set2seq-common in the Application section.

3.2 Baselines and Evaluation Metrics

We compare our framework with five unsupervised/distantly-supervised methods and four supervised methods with domain adaptation. We re-produce ParaNMT [Wieting and Gimpel, 2017] and ParaBank [Hu et al., 2019b] using our translation models, and take the results from Liu et al. [2019] and Liu et al. [2020] for other baselines. For a fair comparison, we keep the scripts for data pre-processing and evaluation from UPSA.

\footnote{https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs}
On the Quora dataset, we even use the same train-test split as UPSA.$^3$

Unsupervised and distantly-supervised methods. The current state-of-the-art unsupervised method is Unsupervised Paraphrasing by Simulated Annealing (UPSA) [Liu et al., 2019], which is also our main target of comparison. The other unsupervised methods is CGMH [Miao et al., 2019]. Distantly-supervised baselines are the unsupervised part by Liu et al. [Liu et al., 2020], ParaNMT and ParaBank(-3$^2$IDF). Note that ParaNMT used back-translation to generate paraphrases, so it can be viewed as “back-translation only”.

Supervised methods with domain adaptation. Decomposable Neural Paraphrase Generation (DNPG) [Li et al., 2019] is the current state-of-the-art method for supervised paraphrase generation. Other baselines are Pointer-generator [See et al., 2017], Transformer [Vaswani et al., 2017] with copy mechanism, and MTL[Domhan and Hieber, 2017] with copy mechanism.

Evaluation metrics. For fair comparisons, we take the same evaluation metrics as in UPSA and DNPG, which are iBLEU [Sun and Zhou, 2012], BLEU [Papineni et al., 2002] and ROUGE [Lin, 2004] scores. BLEU and ROUGE scores are common evaluation metrics for NLG tasks while iBLEU is especially designed for paraphrase generation tasks. It penalizes the similarity between paraphrase and the original sentence. Suppose the input sentence is $src$, the output paraphrase is $out$, and the ground truth paraphrase is $trg$, we calculate iBLEU as follows:

$$iBLEU = \alpha \cdot BLEU(out, trg) - (1 - \alpha) \cdot BLEU(out, src)$$ (4)

BLEU and ROUGE only consider the accuracy but ignore the diversity of generated paraphrases, while iBLEU considers both. So we use iBLEU as our main evaluation metric. We set $\alpha = 0.9$, same as other baselines.

3.3 Implementation and Training Details
To be consistent with the pre-processing of UPSA and DNPG, we convert the input words into lower-case and truncate all sentences to up to 20 words. For the convenience of hybrid decoding, we learn a shared byte-pair encoding (BPE, [Sennrich et al., 2016]) with size 50k from the training data for translation models, and use a 30k vocabulary for all models. The hyper-parameter $\lambda$ in the hybrid decoder is set to 0.5 for all datasets after experimenting with difference choices.

For the translation models in back-translation, we train them with the WMT17$^4$ zh-en dataset [Ziemski et al., 2016] with a standard transformer for 3 days on two GTX-2080 GPUs. We reuse these translation models for ParaNMT and ParaBank. For the set2seq-common model, we use the news-crawl-2016 English monolingual data from WMT17 and train 1.5 days with a standard transformer. For the domain-specific set2seq models, we use a 2-layer transformer with 300 embedding size, 256 units, 1024 feed-forward dimensions for all layers to train them. The training lasts 3 hours on a single GTX-2080 GPU. Set2seq is a lightweight model with 31M parameters, 3.7M parameters for multi-head attention layers, only one-third of a standard transformer.

To calculate iBLEU and BLEU, four references are used for MSCOCO, five for WikiAnswers, and one for other datasets. Some test cases in WikiAnswers may have fewer than 5 references. For ROUGE scores, we take the average score against all references.

3.4 Results
Table 2 presents our experimental results. We compare three different models with the previous methods, namely set2seq, set2seq-common+BT, and set2seq+BT, where BT stands for back-translation. We show the set2seq alone here to demonstrate that the useful information comes not only from the translation, since the set2seq model alone can already outperform almost all competitors. Our framework outperforms all existing unsupervised methods, distantly-supervised methods, and supervised methods with domain adaptation.

3.5 Analysis
Datasets. Due to the domain-specific differences between four datasets, it is understandable that scores on all metrics vary a lot across different datasets.

Paraphrases from MSCOCO are descriptions of images, the set2seq model fits this dataset quite well since the process of generating paraphrases are similar: one extends information from a static picture; the other extends from a word set.

Lack of training data for Twitter leads to insufficient training of most models. Models using back-translation perform extraordinary well since they have adequate information. Besides, set2seq-common+BT achieves an excellent result, which shows the advantages of the set2seq-common model compared with the set2seq model trained with insufficient in-domain data.

Ablation Study. Table 3 shows the result of the ablation study on the Quora dataset, where BLEU$_{\text{ref}}$ is the BLEU between reference and output, the higher the better and BLEU$_{\text{src}}$ is the BLEU between source sentence and output, the lower the better.

We demonstrate that removing stopwords is better than keeping high-IDF words. For high-IDF words, we keep the top $k\%$ high-IDF words in the original sentence. We set $k = 50$, the best from $\{30, 40, 50, 60, 70\}$ by empirics. We also tried TextRank [Mihalcea and Tarau, 2004] to score words and get similar results with IDF scores.

Removing random replacement and adding position encoding can both give high BLEUs between reference sentences and output paraphrases, but substantially reduce the diversity of the generated sentences.

Human Evaluation. We choose 100 sentences from Quora and ask 3 human annotators to score the results from different methods blindly on a scale of 1 to 5 according to fluency and accuracy (the higher the better). Fluency measures whether the paraphrase conforms to grammar and common sense; accuracy measures whether the paraphrase has the same meaning as the original sentence though in a different expression.

We can see that word/phrase based methods have bad performances on fluency since their language model is trained.

---

$^3$ https://github.com/anonymity-person/UPSA

$^4$ http://statmt.org/wmt17/translation-task.html
Table 2: Evaluation results on Quora, WikiAnswers, MSCOCO and Twitter. The comparison with supervised + domain adapted methods is only on Quora and WikiAnswers because results of current SOTA method (DNPG) are only available on these two datasets. The previous highest scores are marked with the underlines and the present highest scores are marked with the bold font. The supervised method DNPG (SOTA) is shown here only for reference.

<table>
<thead>
<tr>
<th>Model</th>
<th>iBLEU</th>
<th>BLEU</th>
<th>R-1</th>
<th>R-2</th>
<th>iBLEU</th>
<th>BLEU</th>
<th>R-1</th>
<th>R-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>DNPG (SOTA)</td>
<td>18.01</td>
<td>25.03</td>
<td>63.73</td>
<td>37.75</td>
<td>34.15</td>
<td>41.64</td>
<td>57.32</td>
</tr>
<tr>
<td>Supervised + Domain-Adapted</td>
<td>Pointer-generator</td>
<td>5.04</td>
<td>6.96</td>
<td>41.89</td>
<td>12.77</td>
<td>21.87</td>
<td>27.94</td>
<td>54.10</td>
</tr>
<tr>
<td></td>
<td>Transformers+Copy</td>
<td>6.17</td>
<td>8.15</td>
<td>44.89</td>
<td>14.79</td>
<td>23.25</td>
<td>29.22</td>
<td>53.33</td>
</tr>
<tr>
<td></td>
<td>MTL+Copy</td>
<td>7.22</td>
<td>9.83</td>
<td>47.08</td>
<td>19.03</td>
<td>21.87</td>
<td>30.78</td>
<td>54.10</td>
</tr>
<tr>
<td></td>
<td>DNPG</td>
<td>10.39</td>
<td>16.98</td>
<td>56.01</td>
<td>28.61</td>
<td>25.60</td>
<td>35.12</td>
<td>56.17</td>
</tr>
<tr>
<td>Unsupervised</td>
<td>CGMH</td>
<td>9.94</td>
<td>15.73</td>
<td>48.73</td>
<td>26.12</td>
<td>20.05</td>
<td>26.45</td>
<td>43.31</td>
</tr>
<tr>
<td></td>
<td>UPSA</td>
<td>12.02</td>
<td>18.18</td>
<td>56.51</td>
<td>30.69</td>
<td>24.84</td>
<td>32.39</td>
<td>54.12</td>
</tr>
<tr>
<td>Distantly-Supervised</td>
<td>Liu et al. [2020]</td>
<td>9.90</td>
<td>15.03</td>
<td>52.65</td>
<td>23.18</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ParaNMT (back-translation)</td>
<td>10.69</td>
<td>15.75</td>
<td>52.28</td>
<td>25.12</td>
<td>14.94</td>
<td>20.01</td>
<td>30.55</td>
</tr>
<tr>
<td></td>
<td>ParaBank</td>
<td>9.92</td>
<td>14.71</td>
<td>50.03</td>
<td>23.80</td>
<td>13.14</td>
<td>17.56</td>
<td>28.97</td>
</tr>
<tr>
<td></td>
<td>set2seq (ours)</td>
<td>13.54</td>
<td>20.85</td>
<td>58.27</td>
<td>32.59</td>
<td>25.98</td>
<td>35.12</td>
<td>56.17</td>
</tr>
<tr>
<td></td>
<td>set2seq-common+BT (ours)</td>
<td>12.60</td>
<td>18.85</td>
<td>57.13</td>
<td>31.19</td>
<td>25.04</td>
<td>32.39</td>
<td>54.12</td>
</tr>
<tr>
<td></td>
<td>set2seq+BT (ours)</td>
<td>14.66</td>
<td>22.53</td>
<td>59.98</td>
<td>34.09</td>
<td>28.27</td>
<td>37.42</td>
<td>56.71</td>
</tr>
</tbody>
</table>

Table 3: Ablation Study on Quora.

Table 4: Results for Human Evaluation (Mean and Kappa).
Regarding augmentation, we make 10 copies of the original sentences, construct 10 word sets with different seeds in random replacement from the 10 copies and generate 10 paraphrases with set2seq-common+BT. To increase the diversity of the results, we use random sampling [Edunov et al., 2018] during decoding. We take the 10 copies and 10 paraphrases as the augmented data.

For the set2seq-common model, since the sentences in the NMT training set is longer, we truncate all sentences to 50 words instead of 20 during the training stage and do not truncate any sentences during the inference stage.

### 4.2 Experimental Setup and Results

We experiment on German-English (de-en), Chinese-English (zh-en), and Russian-English (ru-en) translation pairs. For the training data, we obtain the de-en data from WMT17-europarl3 [Koehn, 2005], and the ru-en data from WMT17 news-commentary and zh-en data from LDC [Liberman, 2002; Huang et al., 2002]. The reason for not using zh-en data from WMT17 is that we are already using the zh-en pairs from WMT17 to train the translation models. For test sets, there are 3004 pairs for de-en, 2000 pairs for zh-en and 3000 pairs for ru-en from the WMT17 news-test.

For each language, we learn a shared BPE of size 50,000 and extract vocabulary of up to 50,000 from the training set for both English and the target language with the shared BPE. We train translation models with a standard transformer-base model [Vaswani et al., 2017]. For the result of each model, we take the average of test results from 5 checkpoints after convergence.

Table 5 shows the result. Paraphrase augmentation improves the model trained with original data pairs by anywhere from 1.53 to 2.17 BLEU.

When producing paraphrases, our methods do use additional data, such as monolingual English data and Chinese English translation data. It is conceivable that there exists other advanced NMT methods that use these data in different ways. However, the purpose of this section is to show the effectiveness of our long-sentence paraphrase generation methods, since only accurate and diverse paraphrases can be used as good translation pairs and subsequently train good translators.

<table>
<thead>
<tr>
<th>Language</th>
<th>Size</th>
<th>Orig. Pairs</th>
<th>Augmented</th>
</tr>
</thead>
<tbody>
<tr>
<td>De-En</td>
<td>150k</td>
<td>12.89</td>
<td>15.06</td>
</tr>
<tr>
<td></td>
<td>300k</td>
<td>15.67</td>
<td>17.20</td>
</tr>
<tr>
<td>Zh-En</td>
<td>150k</td>
<td>10.21</td>
<td>11.99</td>
</tr>
<tr>
<td></td>
<td>300k</td>
<td>12.10</td>
<td>14.07</td>
</tr>
<tr>
<td>Ru-En</td>
<td>150k</td>
<td>16.88</td>
<td>18.55</td>
</tr>
<tr>
<td></td>
<td>300k</td>
<td>19.30</td>
<td>21.09</td>
</tr>
</tbody>
</table>

Table 5: BLEU scores of translating three languages into English; each task is trained with 150k/300k original pairs and 3M/6M pairs after data-augmentation.

### 5 Related Work

We show the relevant work of paraphrase generation from the aspects of supervised, distance-supervised, and unsupervised methods.

For supervised methods, Prakash et al. [2016] proposed “stacked residual LSTM” as the earliest deep-learning method in this topic, seq2seq models like transformer [Vaswani et al., 2017] and MTL [Domhan and Hieber, 2017] outperformed many methods due to the advantages of their model structures. We include these well-known methods in our baseline. Li et al. [2019] proposed the current state-of-the-art method DNPG and revealed the disadvantage of supervised methods when it comes to domain adaptation. Other methods include VAE-SVG [Gupta et al., 2018] and transformer-pb [Wang et al., 2019], but these methods perform worse than DNPG and have no discussion about domain adaptation, so we do not include them in our baselines.

For distance-supervised methods, Wieting and Gimpel [2017] created a 50M parallel dataset for paraphrases with back-translation, Hu et al. [2019b] used lexically-constrained to improve the diversity of generated paraphrase, and their work is proved to be useful for many downstream tasks like Natural Language Inference [Hu et al., 2019a]. Liu et al. [2020] also use bilingual data to generate paraphrase without parallel data. However, their focus is on the supervised fine-tuning part. Their method do not performs well without the fine-tuning part.

For unsupervised methods, Miao et al. [2019] used Metropolis-Hastings Sampling to generate paraphrases, Liu et al. [2019] generated paraphrases with Simulated Annealing, both of them were the best at their times. We compare our framework with these two methods to show changes on the sentential level are more reliable than changes on the lexical level. Siddique et al. [2020] proposed a method for paraphrasing with deep reinforcement learning. However, we do not regard it as a baseline since their results are not convincing enough for the following two reasons:

- By iBLEU in (4), their $BLEU_{src}$ is 28.04 on Quora, and 91.98 on WikiAnswers, which shows a very large and abnormal disparity.
- The authors provided us with their test set on Quora, where the BLEU score between source sentences and references is 69.75. However, the score should be around 25 if the test samples are randomly selected from Quora. With their test set, it is easier to generate paraphrases similar to references.

### 6 Conclusion

In this paper, we proposed a novel framework for automatical paraphrase generation without parallel training data. It outperforms most existing unsupervised and distantly supervised methods. While the results are positive, some questions remain. Can we find more underlying semantics to represent the input sentence? Can we replace the back-translation model with a lighter-weight model? We plan to look into these questions in the future and generate better paraphrases.
References


