Unsupervised Paraphrasing by Simulated Annealing

Xianggen Liu[1] **Lili Mou**[2] **Fandong Meng**[3] **Hao Zhou**[4] **Jie Zhou**[3] **Sen Song**[1]
[1]Laboratory for Brain and Intelligence and Department of Biomedical Engineering, Tsinghua University
[2]Department of Computing Science, University of Alberta; Alberta Machine Intelligent Institute (Amii)
[3]Pattern Recognition Center, WeChat AI, Tencent Inc, [4]ByteDance AI Lab
liuxg16@mails.tsinghua.edu.cn, doublepower.mou@gmail.com
{fandongmeng,withtomzhou}@tencent.com
zhouhao.nlp@bytedance.com, songsen@tsinghua.edu.cn

Abstract

We propose UPSA, a novel approach that accomplishes Unsupervised Paraphrasing by Simulated Annealing. We model paraphrase generation as an optimization problem and propose a sophisticated objective function, involving semantic similarity, expression diversity, and language fluency of paraphrases. UPSA searches the sentence space towards this objective by performing a sequence of local edits. We evaluate our approach on various datasets, namely, Quora, Wikianswers, MSCOCO, and Twitter. Extensive results show that UPSA achieves the state-of-the-art performance compared with previous unsupervised methods in terms of both automatic and human evaluations. Further, our approach outperforms most existing domain-adapted supervised models, showing the generalizability of UPSA.[1]

1 Introduction

Paraphrasing aims to restate one sentence as another with the same meaning, but different wordings. It constitutes a corner stone in many NLP tasks, such as question answering (Mckeown, 1983), information retrieval (Knight and Marcu, 2000), and dialogue systems (Shah et al., 2018). However, automatically generating accurate and different-appearing paraphrases is a still challenging research problem, due to the complexity of natural language.

Conventional approaches (Prakash et al., 2016; Gupta et al., 2018) model the paraphrase generation as a supervised encoding-decoding problem, inspired by machine translation systems. Usually, such models require massive parallel samples for training. In machine translation, for example, the WMT 2014 English-German dataset contains 4.5M sentence pairs (Neidert et al., 2014).

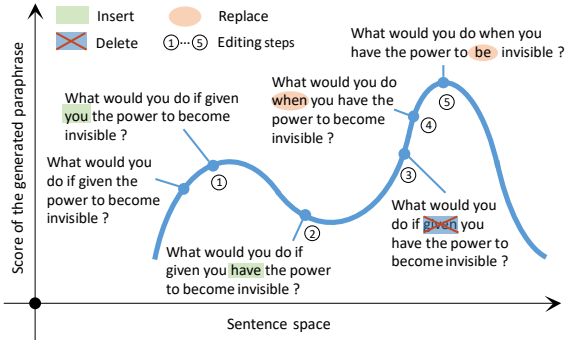

Figure 1: UPSA generates a paraphrase by a series of editing operations (i.e., insertion, replacement, and deletion). At each step, UPSA proposes a candidate modification of the sentence, which is accepted or rejected according to a certain acceptance rate (only accepted modifications are shown). Although sentences are discrete, we make an analogue in the continuous real x-axis where the distance of two sentences is roughly given by the number of edits.

However, the training corpora for paraphrasing are usually small. The widely-used Quora dataset[2] only contains 140K pairs of paraphrases; constructing such human-written paraphrase pairs is expensive and labor-intensive. Further, existing paraphrase datasets are domain-specific: the Quora dataset only contains question sentences, and thus, supervised paraphrase models do not generalize well to new domains (Li et al., 2019). On the other hand, researchers synthesize pseudo-paraphrase pairs by clustering news events (Barzilay and Lee, 2003), crawling tweets of the same topic (Lan et al., 2017), or translating bi-lingual datasets (Wieting and Gimpel, 2017), but these methods typically yield noisy training sets, leading to low paraphrasing performance (Li et al., 2018).

As a result, unsupervised methods would largely benefit paraphrase generation as no parallel data are

[1]Code and data available at: https://github.com/Liuxg16/UPSA

[2]https://www.kaggle.com/c/quora-question-pairs

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 302–312
July 5 - 10, 2020.

needed. With the help of deep learning, researchers are able to generate paraphrases by sampling from a neural network-defined probabilistic distribution, either in a continuous latent space (Bowman et al., 2016; Bao et al., 2019) or directly in the word space (Miao et al., 2019). However, the meaning preservation and expression diversity of those generated paraphrases are less "controllable" in such probabilistic sampling procedures.

To this end, we propose a novel approach to *Unsupervised Paraphrasing by Simulated Annealing* (UPSA). Simulated annealing (SA) is a stochastic searching algorithm towards an objective function, which can be flexibly defined. In our work, we design a sophisticated objective function, considering semantic preservation, expression diversity, and language fluency of paraphrases. SA searches towards this objective by performing a sequence of local editing steps, namely, word replacement, insertion, deletion, and copy. For each step, UPSA first proposes a potential editing, and then accepts or rejects the proposal based on sample quality. In general, a better sentence (higher scored in the objective) is always accepted, while a worse sentence is likely to be rejected, but could also be accepted (controlled by an annealing temperature) to explore the search space in a less greedy fashion. At the beginning, the temperature is usually high, and worse sentences are more likely to be accepted, pushing SA outside a local optimum. The temperature is cooled down as the optimization proceeds, making the model better settle down to some optimum. Figure 1 illustrates how UPSA searches an optimum in unsupervised paraphrase generation.

We evaluate the effectiveness of our model on four paraphrasing datasets, namely, Quora, Wikianswers, MSCOCO, and Twitter. Experimental results show that UPSA achieves a new state-of-the-art unsupervised performance in terms of both automatic metrics and human evaluation.

In summary, our contributions are as follows:

- We propose the novel UPSA framework that addresses Unsupervised Paraphrasing by Simulated Annealing.
- We design a searching objective function for paraphrasing that not only considers language fluency and semantic similarity, but also explicitly models expression diversity between a paraphrase and the input.
- We propose a copy mechanism as one of our search actions of simulated annealing to address rare words.
- We achieve the state-of-the-art performance on four benchmark datasets compared with previous unsupervised paraphrase generators, largely reducing the performance gap between unsupervised and supervised paraphrasing. We outperform most domain-adapted paraphrase generators, and even a supervised one on the Wikianswers dataset.

2 Related Work

In early years, paraphrasing was typically accomplished by exploiting linguistic knowledge (Mckeown, 1983; Ellsworth and Janin, 2007; Narayan et al., 2016) and statistical machine translation methods (Quirk et al., 2004; Dolan et al., 2004). Recently, deep neural networks have become a prevailing approach to text generation, where paraphrasing is often formulated as a supervised encoding-decoding problem, for example, using stacked residual LSTM (Prakash et al., 2016) and the Transformer model (Wang et al., 2019).

Unsupervised paraphrasing is an emerging research direction in the field of NLP. The variational autoencoder (VAE) can be intuitively applied to paraphrase generation in an unsupervised fashion, as we can sample sentences from a learned latent space (Bowman et al., 2016; Zhang et al., 2019; Bao et al., 2019). But the generated sentences are less controllable and suffer from the error accumulation problem in VAE's decoding phase (Miao et al., 2019). Roy and Grangier (2019) introduce an unsupervised model based on vector-quantized autoencoders (Van den Oord et al., 2017). But their work mainly focuses on generating sentences for data augmentation instead of paraphrasing itself.

Miao et al. (2019) use Metropolis–Hastings sampling (1953) for constrained sentence generation, achieving the state-of-the-art unsupervised paraphrasing performance. The main difference between their work and ours is that UPSA imposes the annealing temperature into the sampling process for better convergence to an optimum. In addition, we define our searching objective involving not only semantic similarity and language fluency, but also the expression diversity; we further propose a copy mechanism in our searching process.

Recently, a few studies have applied editing-based approaches to sentence generation. Guu et al. (2018) propose a heuristic delete-retrieve-generate component for a supervised sequence-to-sequence

(Seq2Seq) model. Dong et al. (2019) learn the deletion and insertion operations for text simplification in a supervised way, where their groundtruth operations are obtained by some dynamic programming algorithm. Our editing operations (insertion, deletion, and replacement) are the search actions of unsupervised simulated annealing.

Regarding discrete optimization/searching, a naïve approach is by hill climbing (Edelkamp and Schroedl, 2011; Schumann et al., 2020; Kumar et al., 2020), which is in fact a greedy algorithm. In NLP, beam search (BS, Tillmann et al. 1997) is widely applied to sentence generation. BS maintains a k-best list in a partially greedy fashion during left-to-right (or right-to-left) decoding (Anderson et al., 2017; Zhou and Rush, 2019). By contrast, UPSA is local search with distributed edits over the entire sentence. Moreover, UPSA is able to make use of the original sentence as an initial state of searching, whereas BS usually works in the decoder of a Seq2Seq model and is not applicable to unsupervised paraphrasing.

3 Approach

In this section, we present our novel UPSA framework that uses simulated annealing (SA) for unsupervised paraphrasing. In particular, we first present the general SA algorithm and then design our searching objective and searching actions (i.e., candidate sentence generator) for paraphrasing.

3.1 The Simulated Annealing Algorithm

Simulated Annealing (SA) is an effective and general metaheuristic of searching, especially for a large discrete or continuous space (Kirkpatrick et al., 1983).

Let $\mathcal{X}$ be a (huge) search space of sentences, and $f(\mathrm{x})$ be an objective function. The goal is to search for a sentence x that **maximizes** $f(\mathrm{x})$. At a searching step t, SA keeps a current sentence x_t, and proposes a new candidate x_* by local editing. If the new candidate is better scored by f, i.e., $f(\mathrm{x}_*) > f(\mathrm{x}_t)$, then SA accepts the proposal. Otherwise, SA tends to reject the proposal x_*, but may still accept it with a small probability $e^{\frac{f(\mathrm{x}_*)-f(\mathrm{x}_t)}{T}}$, controlled by an annealing temperature T. In other words, the probability of accepting the proposal is

$$p(\text{accept}|\mathrm{x}_*, \mathrm{x}_t, T) = \min\left(1, e^{\frac{f(\mathrm{x}_*)-f(\mathrm{x}_t)}{T}}\right). \quad (1)$$

If the proposal is accepted, $\mathrm{x}_{t+1} = \mathrm{x}_*$, or otherwise, $\mathrm{x}_{t+1} = \mathrm{x}_t$.

Inspired by the annealing in chemistry, the temperature T is usually high at the beginning of searching, leading to a high acceptance probability even if x_* is worse than x_t. Then, the temperature is decreased gradually as the search proceeds. In our work, we adopt the linear annealing schedule, given by $T = \max(0, T_{\text{init}} - C \cdot t)$, where T_{init} is the initial temperature and C is the decreasing rate.

The high initial temperature of SA makes the algorithm less greedy compared with hill climbing, whereas the decreasing of temperature enables the algorithm to better settle down to a certain optimum.

Theoretically, simulated annealing is guaranteed to converge to the global optimum in a finite search space if the proposal and the temperature satisfy some mild conditions (Granville et al., 1994). Although such convergence may be slower than exhaustive search and the sentence space is, in fact, potentially infinite, simulated annealing is still a widely applied search algorithm, especially for discrete optimization. Readers may refer to Hwang (1988) for details of the SA algorithm.

3.2 Objective Function

Simulated annealing maximizes an objective function, which can be flexibly specified in different applications. In particular, our UPSA objective $f(\mathrm{x})$ considers multiple aspects of a candidate paraphrase, including semantic preservation f_{sem}, expression diversity f_{exp}, and language fluency f_{flu}. Thus, our searching objective is to maximize

$$f(\mathrm{x}) = f_{\text{sem}}(\mathrm{x}, \mathrm{x}_0) \cdot f_{\text{exp}}(\mathrm{x}, \mathrm{x}_0) \cdot f_{\text{flu}}(\mathrm{x}), \quad (2)$$

where x_0 is the input sentence.

Semantic Preservation. A paraphrase is expected to capture all the key semantics of the original sentence. Thus, we leverage the cosine function of keyword embeddings to measure if the key focus of the candidate paraphrase is the same as the input. Specifically, we extract the keywords of the input sentence x_0 by the Rake system (Rose et al., 2010) and embed them by GloVE (Pennington et al., 2014). For each keyword, we find the closest word in the candidate paraphrase x_* in terms of the cosine similarity. Our keyword-based semantic preservation score is given by the lowest cosine similarity among all the keywords, i.e., the least matched keyword:

$$f_{\text{sem,key}}(\mathrm{x}_*, \mathrm{x}_0) = \min_{e \in \text{keywords}(\mathrm{x}_0)} \max_j \{\cos(\boldsymbol{w}_{*,j}, \boldsymbol{e})\}, \quad (3)$$

where $w_{*,j}$ is the jth word in the sentence x_*; e is an extracted keyword of x_0. Bold letters indicate embedding vectors.

In addition to keyword embeddings, we also adopt a sentence-level similarity function, based on Sent2Vec embeddings (Pagliardini et al., 2017). Sent2Vec learns n-gram embeddings and computes the average of n-grams embeddings as the sentence vector. It has been shown to be significant improvements over other unsupervised sentence embedding methods in similarity evaluation tasks (Pagliardini et al., 2017). Let $\mathbf{x}_*$ and $\mathbf{x}_0$ be the Sent2Vec embeddings of the candidate paraphrase and the input sentence, respectively. Our sentence-based semantic preservation scoring function is $f_{\text{sim,sen}}(\mathrm{x}_*, \mathrm{x}_0) = \cos(\mathbf{x}_*, \mathbf{x}_0)$.

To sum up, the overall semantic preservation scoring function of UPSA is given by

$$f_{\text{sem}}(\mathrm{x}_*, \mathrm{x}_0) = f_{\text{sem,key}}(\mathrm{x}_*, \mathrm{x}_0)^P \cdot f_{\text{sem,sen}}(\mathrm{x}_*, \mathrm{x}_0)^Q, \tag{4}$$

where P and Q are hyperparameters, balancing the importance of the two factors. Here, we use power weights because the scoring functions are multiplicative.

Expression Diversity. The expression diversity scoring function computes the lexical difference of two sentences. We adopt a BLEU-induced function to penalize the repetition of the words and phrases in the input sentence:

$$f_{\text{exp}}(\mathrm{x}_*, \mathrm{x}_0) = (1 - \text{BLEU}(\mathrm{x}_*, \mathrm{x}_0))^S, \tag{5}$$

where the BLEU score (Papineni et al., 2002) computes a length-penalized geometric mean of n-gram precision ($n = 1, \cdots, 4$). S coordinates the importance of $f_{\text{exp}}(\mathrm{x}_t, \mathrm{x}_0)$ in the objective function (2).

Language Fluency. Despite semantic preservation and expression diversity, the candidate paraphrase should be a fluent sentence by itself. We use a separately trained (forward) language model (denoted as $\overrightarrow{\text{LM}}$) to compute the likelihood of the candidate paraphrase as our fluency scoring function:

$$f_{\text{flu}}(\mathrm{x}_*) = \prod_{k=1}^{k=l_*} p_{\overrightarrow{\text{LM}}}(w_{*,k} | w_{*,1}, \ldots, w_{*,k-1}), \tag{6}$$

where l_* is the length of x_* and $w_{*,1}, \ldots, w_{*,l}$ are words of x_*. Here, we use a dataset-specific language model, trained on non-parallel sentences. Notice that a weighting hyperparameter is not needed for f_{flu}, because the relative weights of different factors in Eqn. (2) are given by the powers in $f_{\text{sem,key}}$, $f_{\text{sem,sen}}$, and f_{exp}.

3.3 Candidate Sentence Generator

As mentioned, simulated annealing proposes a candidate sentence, given by different search actions. Since each action yields a new sentence x_* from x_t, we call it a *candidate sentence generator*. While the proposal of candidate sentences does not affect convergence in theory (if some mild conditions are satisfied), it may largely influence the efficiency of SA searching.

In our work, we mostly adopt the word-level editing in Miao et al. (2019) as our searching actions, but we differ in sampling distributions and further propose a copy mechanism for editing.

At each step t, the candidate sentence generator randomly samples an editing position k and an editing operation namely, replacement, insertion, and deletion. For replacement and insertion, the candidate sentence generator also samples a candidate word. Let the current sentence be $\mathrm{x}_t = (w_{t,1}, \ldots, w_{t,k-1}, w_k, w_{t,k+1} \ldots, w_{t,l_t})$. If the replacement operation proposes a candidate word w_* for the kth step, the resulting candidate sentence becomes $\mathrm{x}_* = (w_{t,1}, \ldots, w_{t,k-1}, w_*, w_{t,k+1} \ldots, w_{t,l_t})$. The insertion operation works similarly.

Here, the candidate word is sampled from a probabilistic distribution, induced by the objective function (2):

$$p(w_*|\cdot) = \frac{f_{\text{sim}}(\mathrm{x}_*, \mathrm{x}_0) \cdot f_{\text{exp}}(\mathrm{x}_*, \mathrm{x}_0) \cdot f_{\text{flu}}(\mathrm{x}_*)}{Z}, \tag{7}$$

$$Z = \sum_{w_* \in \mathcal{W}} f_{\text{sim}}(\mathrm{x}_*, \mathrm{x}_0) \cdot f_{\text{exp}}(\mathrm{x}_*, \mathrm{x}_0) \cdot f_{\text{flu}}(\mathrm{x}_*), \tag{8}$$

where $\mathcal{W}$ is the sampling vocabulary; Z is known as the normalizing factor (noticing our scoring functions are nonnegative). We observe that sampling from such objective-induced distribution typically yields a meaningful candidate sentence, which enables SA to explore the search space more efficiently.

It is also noted that sampling a word from the entire vocabulary involves re-evaluating (2) for each candidate word, and therefore, we also follow Miao et al. (2019) and only sample from the top-K words given by jointly considering a forward language

Algorithm 1 UPSA

1: **Input**: Original sentence x_0
2: **for** $t \in \{1, \ldots, N\}$ **do**
3: $T = \max\{T_{\text{init}} - C \cdot t, 0\}$
4: Randomly choose an editing operation and a position k
5: Obtain a candidate x_* by candidate sentence generator
6: Compute the accepting probability p_{accept} by Eqn. (1)
7: With probability p_{accept}, $\mathrm{x}_{t+1} = \mathrm{x}_*$
8: With probability $1 - p_{\text{accept}}$, $\mathrm{x}_{t+1} = \mathrm{x}_t$
9: **end for**
10: **return** x_τ s.t. $\tau = \text{argmax}_{\tau \in \{1,\ldots,N\}} f(\mathrm{x}_\tau)$

model and backward language model. The replacement operator, for example, suggests the top-K words vocabulary by

$$\mathcal{W}_{t,\text{replace}} = \text{top-}K_{w_*}\Big[p_{\overrightarrow{\text{LM}}}(w_{t,1}, \ldots, w_{t,k-1}, w_*) \cdot p_{\overleftarrow{\text{LM}}}(w_*, w_{t,k+1}, \ldots, w_{t,l_t})\Big]. \tag{9}$$

For word insertion, the top-K vocabulary $\mathcal{W}_{t,\text{insert}}$ is computed in a similar way (except that the position of w_* is slightly different). Details are not repeated. In our experiments, K is set to 50.

Copy Mechanism. We observe that name entities and rare words are sometimes deleted or replaced during SA stochastic sampling. They are difficult to be recovered because they usually have a low language model-suggested probability.

Therefore, we propose a copy mechanism for SA sampling, inspired by that in Seq2Seq learning (Gu et al., 2016). Specifically, we allow the candidate sentence generator to copy the words from the original sentence x_0 for word replacement and insertion. This is essentially enlarging the top-K sampling vocabulary with the words in x_0, given by

$$\widetilde{\mathcal{W}}_{t,\text{op}} = \mathcal{W}_{t,\text{op}} \cup \{w_{0,1}, \ldots, w_{0,l_0}\} \tag{10}$$

where op $\in$ {replace,insert}. Thus, $\widetilde{\mathcal{W}}_{t,\text{op}}$ is the actual vocabulary from which SA samples the word w_* for replacement and insertion operation.

While such vocabulary reduces the proposal space, it works well empirically because other low-ranked candidate words are either irrelevant or make the sentence disfluent; they usually have low objective scores, and are likely to be rejected even if sampled.

3.4 Overall Optimization Process

We summarize our UPSA algorithm in Algorithm 1.

Given an input x_0, UPSA searches from the sentence space to maximize our objective $f(\mathrm{x})$, which involves semantic preservation, expression diversity, and language fluency. UPSA starts from x_0 itself. For each step, it randomly selects a search action (namely, word insertion, deletion, and replacement) at a position k (Line 4); if insertion or replacement is selected, UPSA also proposes a candidate word, so that a candidate paraphrase x_* is formed (Line 5). Then, UPSA computes an acceptance rate p_{accept} based on the increment of f and the temperature T (Line 6). The candidate sentence x_{t+1} for the next step becomes x_t if the proposal is accepted, or remains x_t if the proposal is rejected. Until the maximum searching iterations, we choose the sentence x_τ that yields the highest score.

4 Experiments

4.1 Datasets

Quora. The Quora question pair dataset (Footnote 2) contains 140K parallel paraphrases and additional 260K pairs of non-parallel sentences. We follow the unsupervised setting in Miao et al. (2019), where 3K and 20K pairs are used for validation and test, respectively.

Wikianswers. The original Wikianswers dataset (Fader et al., 2013) contains 2.3M pairs of question paraphrases from the Wikianswers website. Since our model only involves training a language model, we randomly selected 500K non-parallel sentences for training. For evaluation, we followed the same protocol as Li et al. (2019) and randomly sampled 5K for validation and 20K for testing. Although the exact data split in previous work is not available, our results are comparable to previous ones in the statistical sense.

MSCOCO. The MSCOCO dataset contains 500K+ paraphrases pairs for ∼120K image captions (Lin et al., 2014). We follow the standard split (Lin et al., 2014) and the evaluation protocol in Prakash et al. (2016) where only image captions with fewer than 15 words are considered, since some captions are extremely long (e.g., 60 words).

Twitter. The Twitter URL paraphrasing corpus (Lan et al., 2017) is originally constructed for paraphrase identification. We follow the standard train/test split, but take 10% of the training data as the validation set. The remaining samples are used to train our language model. For the test set, we only consider sentence pairs that are labeled as "paraphrases." This results in 566 test cases.

4.2 Competing Methods and Metrics

Unsupervised paraphrasing is an emerging research topic. We would compare UPSA with recent discrete and continuous sampling-based paraphrase generators, namely, VAE, Lag VAE (He et al., 2019), and CGMH. Early work on unsupervised paraphrasing typically adopts rule-based methods (Mckeown, 1983; Barzilay and Lee, 2003). Their performance could not be verified on the above datasets, since the extracted rules are not available. Therefore, we are unable to compare them in this paper. Also, rule-based systems usually do not generalize well to different domains. In the following, we describe our competing methods:

VAE. We train a variational autoencoder (VAE) with two-layer, 300-dimensional LSTM units. The VAE is trained with non-parallel corpora by maximizing the variational lower bound of log-likelihood; during inference, sentences are sampled from the learned variational latent space (Bowman et al., 2016).

Lag VAE. He et al. (2019) propose to aggressively optimize the inference process of VAE with more updates to address the posterior collapse problem (Chen et al., 2017). This method has been reported to be the state-of-the-art VAE. We adopted the published source code and generated paraphrases for comparison.

CGMH. Miao et al. (2019) use Metropolis–Hastings sampling in the word space for constrained sentence generation. It is shown to outperform latent space sampling as in VAE, and is the state-of-the-art unsupervised paraphrasing approach. We also adopted the published source code and generated paraphrases for comparison.

We further compare UPSA with supervised Seq2Seq paraphrase generators: ResidualLSTM (Prakash et al., 2016), VAE-SVG-eq (Gupta et al., 2018), Pointer-generator (See et al., 2017), the Transformer (Vaswani et al., 2017), and the decomposable neural paraphrase generator (DNPG, Li et al., 2019). DNPG has been reported as the state-of-the-art supervised paraphrase generator.

To better compare UPSA with all paraphrasing settings, we also include domain-adapted supervised paraphrase generators that are trained in a source domain but tested in a target domain, including shallow fusion (Gulcehre et al., 2015) and multi-task learning (MTL, Domhan and Hieber 2017).

We adopt BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) scores as automatic metrics to evaluate model performance. Sun and Zhou (2012) observe that BLEU and ROUGE could not measure the diversity between the generated and the original sentences, and propose the iBLEU variant by penalizing by the similarity with the original sentence. Therefore, we regard the iBLEU score as our major metric, which is also adopted in Li et al. (2019). In addition, we also conduct human evaluation in our experiments (detailed later).

4.3 Implementation Details

Our method involves unsupervised language modeling (forward and backward), realized by two-layer LSTM with 300 hidden units and trained specifically on each dataset with non-parallel sentences.

For hyperparameter tuning, we applied a grid search procedure on the validation set of the Quora dataset using the iBLEU metric. The power weights P, Q, and S in the objective were 8, 1, and 1, respectively, chosen from $\{0.5, 1, 2, \ldots, 8\}$.

The initial temperature T_{init} was chosen from $\{0.5, 1, 3, 5, 7, 9\} \times 10^{-2}$ and set to $T_{\text{init}} = 3 \times 10^{-2}$ by validation. The magnitude of T_{init} appears small here, but is in fact dependent on the scale of the objective function. The annealing rate C was set to $\frac{T_{\text{init}}}{\#\text{Iteration}} = 3 \times 10^{-4}$, where our number of iterations (#Iteration) was 100.

We should emphasize that all SA hyperparameters were validated only on the Quora dataset, and we did not perform any tuning on the other datasets (except the language model). This shows the robustness of our UPSA model and its hyperparameters.

4.4 Results

Table 1 presents the performance of all competing methods on the Quora and Wikianswers datasets. The unsupervised methods are only trained on the non-parallel sentences. The supervised models were trained on 100K paraphrase pairs for Quora and 500K pairs for Wikianswers. The domain-adapted supervised methods are trained on one dataset (Quora or Wikianswers), adapted using non-parallel text on the other (Wikianswers or Quora), and eventually tested on the latter domain (Wikianswers or Quora).

We observe in Table 1 that, among unsupervised approaches, VAE and Lag VAE achieve the worst performance on both datasets, indicating that paraphrasing by latent space sampling is worse than word editing. We further observe that UPSA yields significantly better results than CGMH: the iBLEU score of UPSA is higher than that of CGMH by 2–5

	Model	Quora				Wikianswers			
		iBLEU	BLEU	Rouge1	Rouge2	iBLEU	BLEU	Rouge1	Rouge2
Supervised	ResidualLSTM	12.67	17.57	59.22	32.40	22.94	27.36	48.52	18.71
	VAE-SVG-eq	15.17	20.04	59.98	33.30	26.35	32.98	50.93	19.11
	Pointer-generator	16.79	22.65	61.96	36.07	31.98	39.36	57.19	25.38
	Transformer	16.25	21.73	60.25	33.45	27.70	33.01	51.85	20.70
	Transformer+Copy	17.98	24.77	63.34	37.31	31.43	37.88	55.88	23.37
	DNPG	**18.01**	**25.03**	**63.73**	**37.75**	**34.15**	**41.64**	**57.32**	**25.88**
Supervised + Domain-adapted	Pointer-generator	5.04	6.96	41.89	12.77	21.87	27.94	53.99	20.85
	Transformer+Copy	6.17	8.15	44.89	14.79	23.25	29.22	53.33	21.02
	Shallow fusion	6.04	7.95	44.87	14.79	22.57	29.76	53.54	20.68
	MTL	4.90	6.37	37.64	11.83	18.34	23.65	48.19	17.53
	MTL+Copy	7.22	9.83	47.08	19.03	21.87	30.78	54.10	21.08
	DNPG	10.39	16.98	56.01	28.61	25.60	35.12	56.17	23.65
Unsupervised	VAE	8.16	13.96	44.55	22.64	17.92	24.13	31.87	12.08
	Lag VAE	8.73	15.52	49.20	26.07	18.38	25.08	35.65	13.21
	CGMH	9.94	15.73	48.73	26.12	20.05	26.45	43.31	16.53
	UPSA	12.03	18.21	59.51	32.63	24.84	32.39	54.12	21.45

Table 1: Performance on the Quora and Wikianswers datasets. The best scores within the same training setting are underlined. The results of supervised learning and domain-adapted supervised methods are quoted from Li et al. (2019). We run experiments for all unsupervised methods and use the same evaluation script with Li et al. (2019) for a fair comparison. The results of CGMH in this table is slightly different from Miao et al. (2019), because Miao et al. (2019) use corpus-level BLEU, while Li et al. (2019) and our paper use sentence-level BLEU.

Model	MSCOCO				Twitter			
	iBLEU	BLEU	Rouge1	Rouge2	iBLEU	BLEU	Rouge1	Rouge2
VAE	7.48	11.09	31.78	8.66	2.92	3.46	15.13	3.40
Lag VAE	7.69	11.63	32.20	8.71	3.15	3.74	17.20	3.79
CGMH	7.84	11.45	32.19	8.67	4.18	5.32	19.96	5.44
UPSA	**9.26**	**14.16**	**37.18**	**11.21**	**4.93**	**6.87**	**28.34**	**8.53**

Table 2: Performances on MSCOCO and Twitter.

points. This shows that paraphrase generation is better modeled as an optimization process, instead of sampling from a distribution.

It is curious to see how our unsupervised paraphrase generator is compared with supervised ones, should large-scale parallel data be available. Admittedly, we see that supervised approaches generally outperform UPSA, as they can learn from massive parallel data. Our UPSA nevertheless achieves comparable results with the recent ResidualLSTM model (Prakash et al., 2016), reducing the gap between supervised and unsupervised paraphrasing.

In addition, our UPSA could be easily applied to new datasets and new domains, whereas the supervised setting does not generalize well. This is shown by a domain adaptation experiment, where a supervised model is trained on one domain but tested on the other. We notice in Table 1 that the performance of supervised models (e.g., Transformer+Copy) decreases drastically on out-of-domain sentences, even if both Quora and Wikianswers are question sentences. The performance is supposed to decrease further if the source and target domains are more different. UPSA outperforms all supervised domain-adapted paraphrase generators (except DNPG on the Wikianswers dataset).

Table 2 shows model performance on MSCOCO and Twitter corpora. These datasets are less used for paraphrase generation than Quora and Wikianswers, and thus we could only compare unsupervised approaches by running existing code bases. Again, we see the same trend as Table 1: UPSA achieves the best performance, CGMH second, and VAEs worst. It is also noted that the Twitter corpus yields lower iBLEU scores for all models, largely due to the noise of Twitter utterances (Lan et al., 2017). However, the consistent results demonstrate that UPSA is robust and generalizable to different domains (without hyperparameter re-tuning).

Human Evaluation. We also conducted human

Model	Relevance		Fluency	
	Mean Score	Agreement	Mean Score	Agreement
VAE	2.65	0.41	3.23	0.51
Lag VAE	2.81	0.45	3.25	0.48
CGMH	3.08	0.36	3.51	0.49
UPSA	**3.78**	0.55	**3.66**	0.53

Table 3: Human evaluation on the Quora dataset.

evaluation on the generated paraphrases. Due to the limit of budget and resources, we sampled 300 sentences from the Quora test set and only compared the unsupervised methods (which is the main focus of our work). Selecting a subset of models and data samples is a common practice for human evaluation in previous work (Wang et al., 2019).

We asked three human annotators to evaluate the generated paraphrases in terms of relevance and fluency; each aspect was scored from 1 to 5. We report the average human scores and the Cohen's kappa score (Cohen, 1960). It should be emphasized that our human evaluation was conducted in a blind fashion. Table 3 shows that UPSA achieves the highest human satisfaction scores in terms of both relevance and fluency, and the kappa scores indicate moderate inter-annotator agreement (Landis and Koch, 1977). The results are also consistent with the automatic metrics in Tables 1 and 2. We further conducted two-sided Wilcoxon signed rank tests. The improvement of UPSA is statistically significant with $p < 0.01$ in both aspects, compared with both competing methods.

4.5 Model Analysis

We analyze UPSA in more detail on the most widely-used Quora dataset, with a test subset of 2000 samples.

Ablation Study. We first evaluate the searching objective function (2) in Lines 1–4 of Table 4. The results show that each component of our objective (namely, keyword similarity, sentence similarity, and expression diversity) does play its role in paraphrase generation.

Line 5 of Table 4 shows the effect of our copy mechanism, which is used in word replacement and insertion. It yields roughly one iBLEU score improvement if we keep sampling those words in the original sentence.

Finally, we test the effect of the temperature decay in SA. Line 6 shows the performance if we fix the initial temperature during the whole searching process, which is similar to Metropolis–Hastings sampling.[3] The result shows the importance of the annealing schedule. It also verifies our intuition that sentence generation (in particular, paraphrasing in this paper) should be better modeled as a searching problem than a sampling problem.

Line #	UPSA Variant	iBLEU	BLEU	Rouge1	Rouge2
1	UPSA	**12.41**	18.48	57.06	31.39
2	w/o $f_{\text{sim,key}}$	10.28	15.34	50.85	26.42
3	w/o $f_{\text{sim,sen}}$	11.78	17.95	57.04	30.80
4	w/o f_{exp}	11.93	21.17	59.75	34.91
5	w/o copy	11.42	17.25	56.09	29.73
6	w/o annealing	10.56	16.52	56.02	29.25

Table 4: Ablation study.

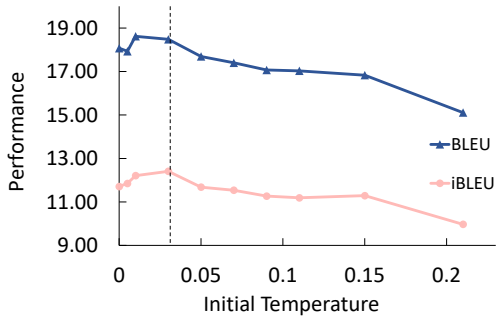

Figure 2: Analysis of the initial temperature T_{init}. The dashed line illustrates the selected hyperparameter in validation.

Analysis of the Initial Temperature. We fixed the decreasing rate to $C = 1 \times 10^{-4}$ and chose the initial temperature T_{init} from $\{0, 0.5, 1, 3, 5, 7, 9, 11, 15, 21\} \times 10^{-2}$. In particular, $T_{\text{init}} = 0$ is equivalent to hill climbing (greedy search). The trend is plotted in Figure 2.

It is seen that a high temperature yields worse performance (with other hyperparameters fixed), because in this case UPSA accepts more worse sentences and is less likely to settle down. On the other hand, a low temperature makes UPSA greedier, also resulting in worse performance. Especially, our simulated annealing largely outperforms greedy search, whose temperature is 0.

We further observe that BLEU and iBLEU peak at different values of the initial temperature. This is because a lower temperature indicates a greedier strategy with less editing, and if the input sentence is not changed much, we may indeed have a higher BLEU score. But our major metric iBLEU penalizes the similarity to the input and thus prefers

[3]The Metropolis–Hastings sampler computes its acceptance rate in a slightly different way from Eqn. (1).

Input	VAE	Lag VAE	CGMH	UPSA
where are best places for spring snowboarding in the us?	where are best places for running in the world? (3.33)	where are best places for honeymoon year near the us? (2.33)	where is best store for the snowboarding in the US? (3.67)	where can I find the best places in the US for snowboarding? (4.67)
how can i become good in studies?	how can i have a good android phone? (2.33)	how can i become good students? (4.33)	how can i become very rich in studies? (4.00)	how should i do to get better grades in my studies? (4.33)
what are the pluses and minuses about life as a foreigner in singapore?	what are the UNK and most interesting life as a foreigner in medieval greece? (2.33)	what are the UNK and interesting things about life as a foreigner? (2.33)	what are the misconception about UNK with life as a foreigner in western? (2.33)	what are the mistakes and pluses life as a foreigner in singapore? (2.67)

Table 5: Example paraphrases generated by different methods on the Quora dataset. The averaged score evaluated by three annotators is shown at the end of each generated sentence.

a higher temperature. We chose $T_{\text{init}} = 0.03$ by validating on iBLEU.

Case Study. We showcase several generated paraphrases in Table 5. We see qualitatively that UPSA can produce more reasonable paraphrases than the other methods in terms of both closeness in meaning and difference in expressions, and can make non-local transformations. For example, "*places for spring snowboarding in the US*" is paraphrased as "*places in the US for snowboarding*." Admittedly, such samples are relatively rare, and our current UPSA mainly synthesizes paraphrases by editing words in the sentence, whereas the syntax is mostly preserved. This is partially due to the difficulty of exploring the entire (discrete) sentence space even by simulated annealing, and partially due to the insensitivity of the similarity objective given two very different sentences.

5 Conclusion and Future Work

In this paper, we proposed a novel unsupervised approach UPSA that generates paraphrases by simulated annealing. Experiments on four datasets show that UPSA outperforms previous state-of-the-art unsupervised methods to a large extent.

In the future, we plan to apply the SA framework on syntactic parse trees in hopes of generating more syntactically different sentences (motivated by our case study).

Acknowledgments

We thank the anonymous reviewers for their insightful suggestions. This work was supported in part by the Beijing Innovation Center for Future Chip. Lili Mou is supported by AltaML, the Amii Fellow Program, and the Canadian CIFAR AI Chair Program; he also acknowledges the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), RGPIN-2020-04465. Sen Song is the corresponding author of this paper.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2017. Guided open vocabulary image captioning with constrained beam search. In *EMNLP*, pages 936–945.

Yu Bao, Hao Zhou, Shujian Huang, Lei Li, Lili Mou, Olga Vechtomova, Xin-yu Dai, and Jiajun Chen. 2019. Generating sentences from disentangled syntactic and semantic spaces. In *ACL*, pages 6008–6019.

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *ACL*, pages 16–23.

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *CoNLL*, pages 10–21.

Xi Chen, Diederik P Kingma, Tim Salimans, Yan Duan, Prafulla Dhariwal, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2017. Variational lossy autoencoder. *ICLR*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING*, pages 350–356.

Tobias Domhan and Felix Hieber. 2017. Using target-side monolingual data for neural machine translation through multi-task learning. In *EMNLP*, pages 1500–1505.

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *ACL*, pages 3393–3402.

Stefan Edelkamp and Stefan Schroedl. 2011. *Heuristic Search: Theory and Applications*. Elsevier.

Michael Ellsworth and Adam Janin. 2007. Mutaphrase: Paraphrasing with framenet. In *Proc. ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 143–150.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *ACL*, pages 1608–1618.

Vincent Granville, Mirko Krivanek, and Jeanpaul Rasson. 1994. Simulated annealing: a proof of convergence. *TPAMI*, 16(6):652–656.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL*, pages 1631–1640.

Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loic Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2015. On using monolingual corpora in neural machine translation. *arXiv preprint arXiv:1503.03535*.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A deep generative framework for paraphrase generation. In *AAAI*, pages 5149–5156.

Kelvin Guu, Tatsunori B Hashimoto, Yonatan Oren, and Percy Liang. 2018. Generating sentences by editing prototypes. *TACL*, 6:437–450.

Junxian He, Daniel Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Lagging inference networks and posterior collapse in variational autoencoders. In *ICLR*.

Chii-Ruey Hwang. 1988. Simulated annealing: theory and applications. *Acta Applicandae Mathematicae*, 12(1):108–111.

S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science*, 220(4598):671–680.

Kevin Knight and Daniel Marcu. 2000. Statistics-based summarization step one: Sentence compression. In *AAAI*, pages 703–710.

Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova. 2020. Iterative edit-based unsupervised sentence simplification. In *ACL*.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *EMNLP*, pages 1224–1234.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Zichao Li, Xin Jiang, Lifeng Shang, and Hang Li. 2018. Paraphrase generation with deep reinforcement learning. In *EMNLP*, pages 3865–3878.

Zichao Li, Xin Jiang, Lifeng Shang, and Qun Liu. 2019. Decomposable neural paraphrase generation. In *ACL*, pages 3403–3414.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proc. Workshop on Text Summarization Branches Out*, pages 74–81.

Tsungyi Lin, Michael Maire, Serge J Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755.

Kathleen R Mckeown. 1983. Paraphrasing questions using given and new information. *Computational Linguistics*, 9(1):1–10.

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. 1953. Equation of state calculations by fast computing machines. *J. Chemical Physics*, 21(6):1087–1092.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. Constrained sentence generation by Metropolis–Hastings sampling. In *AAAI*, pages 6834–6842.

Shashi Narayan, Siva Reddy, and Shay B Cohen. 2016. Paraphrase generation from latent-variable PCFGs for semantic parsing. In *INLG*, pages 153–162.

Julia Neidert, Sebastian Schuster, Spence Green, Kenneth Heafield, and Christopher Manning. 2014. Stanford University's submissions to the WMT 2014 translation task. In *Proc. 9th Workshop on Statistical Machine Translation*, pages 150–156.

Aaron Van den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. In *NIPS*, pages 6306–6315.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2017. Unsupervised learning of sentence embeddings using compositional n-gram features. In *NAACL*, pages 528–540.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: global vectors for word representation. In *EMNLP*, pages 1532–1543.

Aaditya Prakash, Sadid A Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual LSTM networks. In *COLING*, pages 2923–2934.

Chris Quirk, Chris Brockett, and William Dolan. 2004. Monolingual machine translation for paraphrase generation. In *EMNLP*, pages 142–149.

Stuart Rose, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. *Text Mining: Applications and Theory*, 1:1–20.

Aurko Roy and David Grangier. 2019. Unsupervised paraphrasing without translation. In *ACL*, pages 6033–6039.

Raphael Schumann, Lili Mou, Yao Lu, Olga Vechtomova, and Katja Markert. 2020. Discrete optimization for unsupervised sentence summarization with word level extraction. In *ACL*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*, pages 1073–1083.

Pararth Shah, Dilek Hakkani-Tür, Bing Liu, and Gokhan Tür. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *NAACL*, pages 41–51.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In *ACL*, pages 38–42.

Christoph Tillmann, Stephan Vogel, Hermann Ney, A. Zubiaga, and Hassan Sawaf. 1997. Accelerated DP based search for statistical translation. In *EUROSPEECH*, pages 2667–2670.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Su Wang, Rahul Gupta, Nancy Chang, and Jason Baldridge. 2019. A task in a suit and a tie: Paraphrase generation with semantic augmentation. In *AAAI*, pages 7176–7183.

John Wieting and Kevin Gimpel. 2017. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *ACL*, pages 451–462.

Xinyuan Zhang, Yi Yang, Siyang Yuan, Dinghan Shen, and Lawrence Carin. 2019. Syntax-infused variational autoencoder for text generation. In *ACL*, pages 2069–2078.

Jiawei Zhou and Alexander Rush. 2019. Simple unsupervised summarization by contextual matching. In *ACL*, pages 5101–5106.