Automatically Paraphrasing via Sentence Reconstruction and Back-translation

Cover Letter

Dear reviewers:

We are submitting the enclosed paper entitled “Automatically Paraphrasing via Sentence Reconstruction and Back-translation” to AAAI-2021 (EMNLP Fast Track).

This article was submitted to EMNLP-2020 under the name “Unsupervised Paraphrasing via Sentence Reconstruction and Back-translation”. However, but it was not accepted in that two of the reviewers believed that our method was not unsupervised.

Although by definition, we believe that our approach is not wrong to be identified as “unsupervised”, we have revised our article for more rigorous expression. The main changes in our article are as follows:

- We redefined our framework as “distance-supervised without using parallel training data” instead of “unsupervised”.
- We conducted a more detailed human evaluation.
- We drew a more accurate picture to show the structure of our framework.
- We have our code and model ready, and they can be open source immediately after the double-blind period.

PS: In August, an article about unsupervised paraphrasing called “Unsupervised Paraphrasing via Deep Reinforcement Learning” was published on KDD. We tried to reproduce the method in this paper as a baseline. However, in the process of communicating with their authors, we found that the test set they use on Quora is incorrect, and their results are not convincing. We elaborate on this problem in the related work section, and provide the test set they used in the supplementary materials (in dir PUP(KDD-2020)).

Very sincerely yours,
Author Team

Appendix

Here is our response to the reviews from EMNLP.

reviewer 1

Q: Can our approach be regarded as unsupervised?
A: Many reviewers have such questions, so I gave a unified answer in the general response.

reviewer 2

Q: Why zh-en pair for the back-translation model?
A: We don't have strict requirements for the language used in back-translation, so we just use a commonly used translation pair.

Q: Lack of novelty
A: Many reviewers have such questions, so I gave a unified answer in the general response.

Q: Can our approach be regarded as unsupervised?
A: Many reviewers have such questions, so I gave a unified answer in the general response.

reviewer 3

Q: Results for simply copying the source sentence?
A: For Quora WikiAnswers and MSCOCO, copying the source sentence can bring a higher BLEU but a lower IBLEU. For Quora and WikiAnswers, copying the source sentence even performs better than our baselines.
Dataset / BLEU / IBLEU / ROUGE-1 / ROUGE-2
Quora / 25.63 / 13.07 / 63.93 / 38.11 ;
WikiAnswers / 40.52 / 26.47 / 57.29 / 25.60 ;
MSCOCO / 17.28 / 5.55 / 39.85 / 12.42 ;
Twitter / 38.07 / 24.26 / 70.94 / 54.09 ;

Q: Are the same evaluation scripts used for all models?
A: We use the same evaluation scripts for all datasets and all models.

Q: Statistical significance of human evaluation.
A: I calculate the kappa aggrement for the results from different methods:
Method / Aggrement For Accuracy / Aggrement For Fluency
VAE / 0.45 / 0.53 ;
CGMH / 0.55 / 0.50 ;
UPSA / 0.54 / 0.55 ;
BT / 0.59 / 0.58 ;
set2seq+BT / 0.57 / 0.55 ;

Q: PPDB has been used for MT data augmentation?
A: We didn't compared with other methods on MT data augmentation, this section is there to demonstrate one possible downstream task and our system's capability to deal with long sentences (translation data contains longer sentences). This can be a new research topic left for future works.

Q: BT-only model?
A: ParaNMT-50M is a dataset obtained by back-translation, so I use it to represent "BT-only" in the paper.

Q: Lack of novelty
A: Many reviewers have such questions, so I gave a unified answer in the general response.

Q: Use of synonym replacement?

A: I think this is already shown in the ablation study. If I remove the replacement, BLEU will increase to 23.92, but IBLUE will decrease to 13.78, the generated paraphrases will be very similar to the original sentence.

Q: Word Order?
A: We used back-translation to solve that problem. Our method is a whole. Looking at one part alone may be flawed, but the overall performance is good.

General

First of all, allow me to express my gratitude to everyone for your priceless suggestions!

1. Notice that many reviews think that our method cannot be counted as unsupervised, let me explain it here.

Firstly, sorry for our vague definition of unsupervised. In our paper, unsupervised refers to not using parallel paraphrase data.

We think that unsupervised means that there is no direct ground truth guidance in the learning process, so we need to train the model through some indirect methods, and back-translation is one of these indirect methods.

The reason why we don't use supervised methods is not because of the defects of the supervised learning itself, but because of the limitation of the datasets. There is no high-quality, cross-domain dataset for paraphrase generation. However, there are such datasets in the Machine Translation area between English and almost any language. We have not restricted our baseline methods to use parallel MT data (such as ParaNMT), but they cannot achieve the same results as ours.

Please imagine a scenario where we need to generate paraphrases for some sentences in a particular domain. We cannot use supervised methods since their datasets are all domain-specific, but our method can be used. Unsupervised methods are designed for scenarios like this. We think that a method is valuable as long as it can play a role in scenarios similar to this one.

2. Some reviewers pointed out that lack of novelty is another problem for our paper, let me will explain it here.

In my opinion, as a very simple method, set2seq haven't been used in paraphrase generation because it is difficult to balance accuracy and diversity. If too much information is reserved in wordset, the generated sentence will be very close to the original sentence. If the reserved information is insufficient, it will not be able to generate paraphrases with similar meaning.

Wordset in our method has lost some information, which is not enough to retain the meaning of the original sentence completely. Also, in the process of back-translation, some information will be lost. Neither method is good enough to get SOTA performance.

Our main contribution is to combine these two methods, the information lost in one of them can be obtained in another. This refers to boosting's idea that combines multiple weak models into a strong model.