Improving Large-scale Paraphrase Acquisition and Generation

Yao Dou, Chao Jiang, Wei Xu
School of Interactive Computing
Georgia Institute of Technology
{douy, chaojiang}@gatech.edu; wei.xu@cc.gatech.edu
http://twitter-paraphrase.com/

Abstract

This paper addresses the quality issues in existing Twitter-based paraphrase datasets, and discusses the necessity of using two separate definitions of paraphrase for identification and generation tasks. We present a new Multi-Topic Paraphrase in Twitter (MULTIPIIT) corpus that consists of a total of 130k sentence pairs with crowdsourcing (MULTIPIIT\_CROWD) and expert (MULTIPIIT\_EXPERT) annotations using two different paraphrase definitions for paraphrase identification, in addition to a multi-reference test set (MULTIPIIT\_MRK) and a large automatically constructed training set (MULTIPIIT\_AUTO) for paraphrase generation. With improved data annotation quality and task-specific paraphrase definition, the best pre-trained language model fine-tuned on our dataset achieves the state-of-the-art performance of 84.2 \( F_1 \) for automatic paraphrase identification. Furthermore, our empirical results also demonstrate that the paraphrase generation models trained on MULTIPIIT\_AUTO generate more diverse and high-quality paraphrases compared to their counterparts fine-tuned on other corpora such as Quora, MSCOCO, and ParaNMT.

1 Introduction

Paraphrases are alternative expressions that convey a similar meaning (Bhagat and Hovy, 2013). Studying paraphrase facilitates research in both natural language understanding and generation. For instance, identifying paraphrases on social media is important for tracking the spread of misinformation (Bakshy et al., 2011) and capturing emerging events (Vosoughi and Roy, 2016). On the other hand, paraphrase generation improves the linguistic diversity in conventional agents (Li et al., 2016) and machine translation (Thompson and Post, 2020). It has also been successfully applied in data augmentation to improve information extraction (Zhang et al., 2015; Ferguson et al., 2018) and question answering systems (Gan and Ng, 2019).

Many researchers have been leveraging Twitter data to study paraphrase given its lexical and style diversity as well as coverage of up-to-date events. However, existing Twitter-based paraphrase datasets, namely PIT-2015 (Xu et al., 2015) and Twitter-URL (Lan et al., 2017), suffer from quality issues such as topic unbalance and annotation noise,\(^{1}\) which limit the performance of the models trained using them. Moreover, past efforts on creating paraphrase corpora only consider one paraphrase criteria without taking into account the fact that the desired “strictness” of semantic equivalence in paraphrases varies from task to task (Bhagat and Hovy, 2013; Liu and Soh, 2022). For example, for the purpose of tracking unfolding events, “A tsunami hit Haiti.” and “303 people died because of the tsunami in Haiti” are sufficiently close to be considered as paraphrases; whereas for paraphrase generation, the extra information “303 people dead” in the latter sentence may lead models to learn to

\(^{1}\)63% of sentences in Twitter-URL are related to the 2016 US presidential election, and 58% of sentences in PIT-2015 are about NFL draft (more detailed analysis in §2.4).
hallucinate and generate more unfaithful content.

In this paper, we present an effective data collection and annotation method to address these issues. We curate the Multi-Topic Paraphrase in Twitter (MULTIPIT) corpus, which includes MULTIPIT, a large crowdsourced set of 125K sentence pairs that is useful for tracking information on Twitter, and MULTIPITEXPERT, an expert annotated set of 5.5K sentence pairs using a stricter definition that is more suitable for acquiring paraphrases for generation purpose. Compared to PIT-2015 and Twitter-URL, our corpus contains more than twice as much data with more balanced topic distribution and better annotation quality. Two sets of examples from MULTIPIT are shown in Figure 1.

We extensively evaluate several state-of-the-art neural language models on our datasets to demonstrate the importance of having task-specific paraphrase definition. Our best model achieves 84.2 F1 for automatic paraphrase identification. In addition, we construct a continually growing paraphrase dataset, MULTIPITAUTO, by applying the automatic identification model to unlabelled Twitter data. Empirical results and analysis show that generation models fine-tuned on MULTIPITAUTO generate more diverse and high-quality paraphrases compared to models trained on other corpora, such as MSCOCO (Lin et al., 2014), ParaNMT (Wieting and Gimpel, 2018), and Quora.2 We hope our MULTIPIT corpus will facilitate future innovation in paraphrase research.

2 Multi-Topic PIT Corpus

In this section, we present our data collection and annotation methodology for creating MULTIPIT and MULTIPITEXPERT datasets. The data statistics is detailed in Table 1.

2.1 Collection of Tweets

To gather paraphrases about a diverse set of topics as illustrated in Figure 1, we first group tweets that contain the same trending topic3 (year 2014–2015) or the same URL (year 2017–2019) retrieved through Twitter public APIs4 over a long time period. Specifically, for the URL-based method, we extract the URLs embedded in the tweets that are posted by 15 news agency accounts (e.g., NYTScience, CNNPolitics, and ForbesTech). To get cleaner paraphrases, we split the tweets into sentences, eliminating the extra noises caused by multi-sentence tweets. More details of the improvements we made to address the data preprocessing issues in prior work are described in Appendix B.

2.2 Topic Classification and Balancing

To avoid a single type of topics dominating the entire dataset as in prior work (Xu et al., 2015; Lan et al., 2017), we manually categorize the topics for each group of tweets and balance their distribution. For trending topics, we ask three in-house annotators to classify them into 4 different categories: sports, entertainment, event, and others. All three

---

2https://www.kaggle.com/c/quora-question-pairs
3https://www.twitter.com/explore/tabs/trending
4https://developer.twitter.com/en/docs/twitter-api
annotators are college students with varied linguistic annotation experience, and each received an hour-long training session. For URLs, most of them are linked to news articles and have already been categorized by the news agency.\(^5\) We include the tweets grouped by URLs that belong to the science/tech, health, politics, and finance categories.

### 2.3 Candidate Selection

The PIT-2015 (Xu et al., 2015) and Twitter-URL (Lan et al., 2017) corpora contain only 23\% and 31\% sentence pairs that are paraphrases, respectively. To increase the portion of paraphrases and improve the annotation efficiency, we introduce an additional step to filter out the tweet groups that contain either too much noise or too few paraphrases, and adaptively select sentence pairs for annotation (§2.4). For each of the trend-based groups, we first select the top 2 sentences using a simple ranking algorithm (Xu et al., 2015) based on the averaged probability of words. We pair each of these two sentences with 10 other sentences that are randomly sampled from the top 20 in each group. Among these 20 sentence pairs, if the annotators found \(n \in [4, 6]\) or \([7, 9]\) or \([10, 12]\) or \([13, 20]\) pairs as paraphrases, then we further deploy 20, 30, 40, or 50 sentence pairs for annotation, respectively. We pair one of the top 5 ranked sentences with 10 sentences randomly selected from those ranked between top 6 and top 50. Since the URL-based groups generally contain fewer sentences, we select the top 11 sentences and ask annotators to choose one as the seed sentence that can be paired with the rest 10 sentences to produce at least 3 paraphrase pairs. If such a seed sentence exists, we pair it with the rest 10 sentences and deploy them for annotation. Otherwise, we skip the entire group.

### 2.4 Crowd Annotation for Paraphrase Identification

We then annotate the selected sentence pairs using the crowdsourcing platform Figure-Eight\(^6\) to construct \(\text{MULTI}^\text{CROWD}\).

#### Annotation Process.

We design a 1-vs-1 annotation schema, where we present one sentence pair to workers at a time and ask them to annotate whether it is a paraphrase pair or not. A screenshot of the annotation interface is provided in Appendix A.1. We collect 6 judgments for every sentence pair and pay \$0.2 per annotation (>\$7 per hour). For creating \(\text{MULTI}^\text{CROWD}\), with the purpose of identifying similar sentences and tracking information spreading on Twitter in mind, we consider two sentences as paraphrases even if one contains some new information that does not appear in the other sentence (see Figure 3 for examples). As a side note, because these sentences are grouped under the same trend or URL, the new information is always relevant and based on the context, otherwise, we will consider them non-paraphrases.

#### Quality Control.

In every five sentence pairs, we embed one hidden test sentence pair that are pre-labeled by one of the authors, and constantly monitor the workers’ performance. Whenever annotators make a mistake on the test pair, they will be alerted and provided with an explanation. Workers can continue in the task if they achieve >85\% accuracy on the test pairs and >0.2 Cohen’s (Cohen, 1960) kappa when compared with the major vote of other workers. All workers are in the U.S.

#### Inter-Annotator Agreement.

The average Cohen’s kappa is 0.75 for URL-sourced sentence pairs,
Sentence2 is a paraphrase of Sentence1 if:

For Tracking Info On Twitter (MULTI-PIT\textsubscript{CROWD})

For Generation (MULTI-PIT\textsubscript{EXPERT})

\[ S_2 \leftarrow S_1 \rightarrow \text{Simplification} \]
\[ S_2 \leftrightarrow S_1 \rightarrow \text{Add Commonsense} \]
\[ S_2 \rightarrow S_1 \rightarrow \text{Add World Knowledge} \]
\[ plus \]
\[ S_2 \rightarrow S_1 \rightarrow \text{Add Info Requires Fact Checking} \]

Examples:

- **Simplification**
  - A1: Sweden’s first female PM Magdalena Andersson, resigns on DAY ONE!
  - A2: Swedish PM Magdalena Andersson resigns hours after taking job.

- **Add World Knowledge**
  - B1: Facebook announces it will be changing its name to Meta.
  - B2: Facebook relaunches itself as ‘Meta’ in a clear bid to dominate the metaverse.

- **Add Info Requires Fact Checking**
  - C1: 100% of the 140,000 U.S. jobs lost in December were held by women.
  - C2: In fact women lost 111% of the jobs in December because men gained 16,000 jobs.

Figure 3: Two different paraphrase definitions used for creating MULTI-PIT\textsubscript{CROWD} and MULTI-PIT\textsubscript{EXPERT}, with examples. The difference between the two criteria is whether considering Sentence2 that contains new information that requires fact-checking as a paraphrase of Sentence1.\(^3\)

0.69 for Trends-sourced ones, and 0.70 for all. We also sample 400 sampled sentence pairs and hire two experienced in-house annotators to label them. Assuming the in-house annotation is gold, the \(F_1\) of crowdworkers’ majority vote is 89.1.

**Accessing Topic Diversity.** We manually examine 100 sentence pairs randomly sampled from MULTI-PIT\textsubscript{CROWD}, PIT-2015 (Xu et al., 2015) and Twitter-URL (Lan et al., 2017). Figure 2 shows the results of the manual inspection. MULTI-PIT\textsubscript{CROWD} has a much more balanced topic distribution, compared to prior work where 58% of sentences in PIT-2015 are about sports and 63% of sentences in Twitter-URL are politics-related. This improvement can be attributed to the long time period (§2.1) and topic classification step (§2.2) in our data collection process. In contrast, PIT-2015 was collected within only 10 days (04/24/2013 – 05/03/2013) that was overwhelmed by a popular sports event – the 2013 NFL draft (04/25 - 04/27), and Twitter-URL was collected during the 3 months of the 2016 US presidential election.

2.5 Expert Annotation for Paraphrase Generation

Text generation models are prone to memorize training data and generate unfaithful hallucinations (Maynez et al., 2020; Carlini et al., 2021). Including paraphrase pairs that contain extra information other than world or commonsense knowledge in the training data only worsens the problem, as shown in Table 15 in Appendix F. For the purpose of paraphrase generation, we further create MULTI-PIT\textsubscript{EXPERT} with expert annotations, using a stricter paraphrase definition than the one used in MULTI-PIT\textsubscript{CROWD}. The different paraphrase criteria used for creating these two datasets and their corresponding examples are illustrated in Figure 3.

**Data Selection.** To create a high-quality corpus that focuses on differentiating strict paraphrases from the more loosely defined ones, we first use our best paraphrase identifier (§3) fine-tuned on MULTI-PIT\textsubscript{CROWD} to filter the sentence pairs and then have experienced in-house annotators to further annotate them. Specifically, we gather sentence pairs that are identified as paraphrases by the automatic classifier from 9,762 trending topic groups (from Oct-Dec 2021) and 181,254 URL groups (from Jan 2020-Jun 2021). To improve the diversity of our dataset, instead of presenting these pairs directly to the experts for annotation, we cluster the sentences by considering the paraphrase relationship transitive, i.e., if sentence pairs \((s_1, s_2)\) and \((s_2, s_3)\) are both identified as paraphrases, then \((s_1, s_2, s_3)\) is a cluster. For each trend or URL, we show two seed sentences paired with up to 30 sentences in the largest cluster for the experts to annotate. In total, we have 5,570 sentence pairs annotated for MULTI-PIT\textsubscript{EXPERT}, in which 100 sentences sourced by trend and 100 ones sourced by URL have at least 8 corresponding paraphrases. We use these 200 sets to form MULTI-PIT\textsubscript{NMR}, the first multi-reference test set for paraphrase generation evaluation (§4).

**Expert Annotation.** We ask two experienced annotators with linguistic backgrounds and rich annotation experience to annotate each sentence pair as paraphrases or not. Annotators thoroughly discuss
pairs that have inconsistent judgments until reaching an agreement. A screenshot of the updated annotation instruction is provided in Appendix A.2.

3 Paraphrase Identification

Paraphrase identification is a task that determines whether two given sentences are paraphrases or not. The two paraphrase definitions used in MULTI\textsc{Pit} and MULTI\textsc{Pit} suit different downstream applications: tracking information on Twitter and acquiring high-quality paraphrase pairs for training generation models. Paraphrase identification models trained on our datasets achieve over 84 $F_1$ for each case.

Experimental Setup. As each sentence pair in MULTI\textsc{Pit} has six judgments, we use 3 as the threshold, where pairs with $>3$ paraphrase judgments are labeled as paraphrase, and the ones with $<3$ paraphrase judgments are labeled as non-paraphrase. We split MULTI\textsc{Pit} and MULTI\textsc{Pit} into 80/10/10\% for train/dev/test partitions by time such that the oldest data are used for training. More details on the implementation and hyperparameter tuning are in Appendix C.

3.1 Models

We consider an encoder-decoder language model, T5 (Raffel et al., 2020), five masked language models, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2019), BERTweet (Nguyen et al., 2020), and DeBERTaV3 (He et al., 2021). We also include two competitive BiLSTM-based models, Infersent (Conneau et al., 2017) and ESIM (Chen et al., 2017), to establish comparison with pre-BERT era work.

Table 2: Results on the test sets of MULTI\textsc{Pit} \textsc{crowd} and MULTI\textsc{Pit} \textsc{expert}. Models are fine-tuned on the corresponding training set. DeBERTaV3\textsubscript{large} performs the best on both datasets. LR: learning rate.

| Method          | Data                    | P. | R. | $F_1$ | Acc. |
|-----------------|-------------------------|----|----|-------|------|
| Fine-tuning     | M\textsubscript{c}      |    |    | 61.81 | 88.58|
| Fine-tuning     | M\textsubscript{c} + M\textsubscript{p} |72.82 | 69.84 |
| + Filtering     | M\textsubscript{c} + M\textsubscript{p} | 82.56 | 83.86 |
| + Flipping      | M\textsubscript{c} + M\textsubscript{p} | 83.40 | 85.04 |
|                 |                         |    |    | 84.21 | 85.46|

Table 3: Results of different methods on the test set of MULTI\textsc{Pit} \textsc{expert}. M\textsubscript{c}: MULTI\textsc{Pit} \textsc{crowd}; M\textsubscript{p}: MULTI\textsc{Pit} \textsc{expert}. We use DeBERTaV3\textsubscript{large} in the experiments.

3.2 Results

Table 2 presents results for the models fine-tuned on each dataset. DeBERTaV3\textsubscript{large} achieves the best results with 92 $F_1$ on MULTI\textsc{Pit} \textsc{crowd} and 83.2 $F_1$ on MULTI\textsc{Pit} \textsc{expert}. Transformer-based models consistently outperform BiLSTM-based models, especially on MULTI\textsc{Pit} \textsc{expert}.

Beyond Fine-tuning. As MULTI\textsc{Pit} \textsc{crowd} is a large-scale dataset annotated with a loose paraphrase definition, we test whether leveraging these “noisy” data improves model performance on MULTI\textsc{Pit} \textsc{expert}. To reduce the noise that comes from the difference in definitions, we first adjust the labeling threshold for MULTI\textsc{Pit} \textsc{crowd} from 3 to 4. Then we consider two noisy training techniques adopted in prior work (Xie et al., 2020; Zhang and Sabuncu, 2018), namely filtering and flipping. Specifically, we fine-tune a teacher model on MULTI\textsc{Pit} \textsc{expert} and use it to go through MULTI\textsc{Pit} \textsc{crowd} as follows: for each sentence pair $p$, if its label is $i$ ($0$ for non-paraphrase, $1$ for paraphrase) and $P_{\text{teacher}}(y = i | p) \leq \lambda$, we filter out $p$ or flip its label to $1-i$ (i.e. $0 \rightarrow 1$).

\footnote{We perform a small grid search on $\lambda$ over $\{0.05, 0.15, 0.25, 0.35, 0.45\}$, and find 0.35 works well for the filtering method and 0.25 for the flipping method.}
model on the combination of \textsc{MultiPIT}_\textsc{expert} and the re-labeled \textsc{MultiPIT}_\textsc{crowd}. The experimental results are shown in Table 3. Compared to fine-tuning on \textsc{MultiPIT}_\textsc{expert}, adding the original \textsc{MultiPIT}_\textsc{crowd} to the training data results in a 9.8 and 19.5 points drop in \(F_1\) and precision, respectively, demonstrating the necessity of task-specific paraphrase definition. Among all methods, the flipping approach achieves the best \(F_1\) of 84.2. We thus use it to create \textsc{MultiPIT}_\textsc{auto} (§4).

3.3 Impact of Data Size

Figure 4 shows test set performance of DeBERTaV3\textsubscript{large} fine-tuned on different amounts of data in \textsc{MultiPIT}_\textsc{expert}. As there are 156 trend/URL groups in the train set, we truncate the data by group. With more training data, the model achieves better \(F_1\) and accuracy but in a slower fashion compared to the early stage. This finding suggests that annotating more data can further improve the model’s performance.

4 Paraphrase Generation

Paraphrase generation is a task that rewrites the input sentence while preserving its semantic meaning. Since new data is generated on Twitter every day, we introduce \textsc{MultiPIT}_\textsc{auto}, an automated continual growing dataset for paraphrase generation. We show that the model fine-tuned on \textsc{MultiPIT}_\textsc{auto} generates more diverse and high-quality paraphrases than other paraphrase datasets.

4.1 Comparison with Existing Datasets

\textsc{MSCOCO} (Lin et al., 2014), and \textsc{ParaNMT} (Wieting and Gimpel, 2018), and \textsc{Quora}\footnote{https://www.kaggle.com/c/quora-question-pairs} are three widely used datasets in paraphrase generation research (Zhou and Bhat, 2021). The \textsc{Quora} dataset contains over 400K question pairs, including 144K pairs labeled as duplicated (i.e., paraphrase), which are split into 134K/5K/5K as train/dev/test sets, respectively. \textsc{MSCOCO} consists of over 120K images, each of which has five captions. Following Chen et al. (2020), for each image, we randomly pick a caption and pair it with each of the other four captions, resulting in about 490K paraphrase pairs. We split them into train/dev/test sets with 330K/80K/80K pairs, respectively. \textsc{ParaNMT} is a dataset with more than 50 million paraphrase pairs that are automatically generated through back-translation. Since back-translation may introduce noise, we use the manually labeled dev and test sets from Chen et al. (2019), which contain 499 and 871 instances, respectively.

\textsc{MultiPIT}_\textsc{auto}. We use the best performing model in Section 3 to extract paraphrase pairs from recent Twitter data (trending topics in Oct-Dec 2021 and URLS in Jan 2020-Jun 2021). We call these automated identified paraphrase pairs \textsc{MultiPIT}_\textsc{auto},\footnote{Future identified paraphrase pairs will be released every month} which contains 302,307 pairs. One of the authors manually annotates 215 paraphrase pairs and uses them as the dev set. We use the multi-reference \textsc{MultiPIT}_\textsc{nrm} test set (§2.5) for evaluation. As the test set and \textsc{MultiPIT}_\textsc{auto} come from the same time period, we filter out sentence pairs in \textsc{MultiPIT}_\textsc{auto} that share similar trends or URLs with the pairs from the test set. This leaves us with 290,395 pairs as the training set.

Following Chen et al. (2019), we remove paraphrase pairs with high BLEU scores in each training set to ensure there is enough variation between paraphrases, leaving about 137K pairs for \textsc{MultiPIT}_\textsc{auto}, 47K for \textsc{Quora}, 275K for \textsc{MSCOCO}, and 443K for \textsc{ParaNMT}. Table 14 in Appendix F shows BLEU filtering improves model performance for all datasets. Detailed dataset statistics are provided in Appendix E.

4.2 Evaluation Metrics

We consider four automated metrics that are commonly used in previous work (Li et al., 2019; Niu et al., 2021) for paraphrase generation: \textsc{BLEU} (Papineni et al., 2002), \textsc{Self-BLEU} (Liu et al., 2021), \textsc{BERT-Score} (Zhang et al., 2020), and \textsc{BERT-IBLEU} (Niu et al., 2021). \textsc{Self-BLEU} is BLEU
Table 4: Test set results of different transformer models fine-tuned on MultiPIT\textsubscript{AUTO}, except GPT-3, where in-context learning is used. BL: BLEU, S-B: BERT-Score, B-iB: BERT-iBLEU. LR: learning rate. \textbf{Bold}: the best. The Self-BLEU of human reference is calculated by taking the min/avg/max score of the 8 references for each input sentence first, and then averaging across all scores.

| Model               | #Para. | LR  | BL  | S-B ↓ | B-S  | B-iB  |
|---------------------|--------|-----|-----|-------|------|-------|
| GPT-2\textsubscript{small} | 117M   | 3e-5| 41.15| 51.38 | 88.18| 65.23 |
| GPT-2\textsubscript{large} | 774M   | 3e-5| 42.89| 39.61 | 86.16| 74.01 |
| BART\textsubscript{base} | 139M   | 1e-5| 46.91| 46.38 | 87.65| 71.40 |
| BART\textsubscript{large} | 406M   | 1e-5| 47.22| 38.26 | 86.40| 75.17 |
| T5\textsubscript{small} | 60M    | 3e-4| 38.27| 52.16 | 88.32| 68.37 |
| T5\textsubscript{base} | 220M   | 1e-4| 42.10| 46.43 | 87.75| 72.29 |
| T5\textsubscript{large} | 770M   | 1e-4| 41.14| 33.34 | 85.86| 77.79 |

| Diversity (S-B ↓) | Min. | Avg. | Max. |
|-------------------|------|------|------|
| Human Reference   | 6.52 | 31.68| 86.66| 80.16|

Figure 5: Test set performance of model fine-tuned on varying amount of data in MultiPIT\textsubscript{AUTO}, in terms of Self-BLEU (lower is better) and BERT-iBLEU.

much more diverse with a decrease of 24.5 in Self-BLEU under the best case and 13.5 under the average case, indicating that there is still a big gap between large language models and humans. For supervised small-scale models, T5\textsubscript{large} outperforms others with the best Self-BLEU and BERT-iBLEU scores. Although BART\textsubscript{large} gets the highest BLEU score, our experiments in Appendix F show BERT-iBLEU has the best correlation with human evaluation. We thus use T5\textsubscript{large} in all the rest experiments. For all models except GPT-3, we use beam search with beam size = 4. Please refer to Appendix C for details on the training setup and hyperparameter tuning. GPT-3 prompting and hyperparameter setup are provided in Appendix D. Generation examples are displayed in Figure 16 in Appendix G.

Impact of Data Size. Figure 5 shows test set performance of T5\textsubscript{large} fine-tuned on different amount of data in MultiPIT\textsubscript{AUTO} from 1K to 137K. With more training data, the model generates more diverse and high-quality paraphrases as Self-BLEU decreases (improves) and BERT-iBLEU increases. This suggests that the paraphrase generation models will benefit from the continually growing size of our MultiPIT\textsubscript{AUTO} corpus.

4.3 Generation Models

We consider two autoregressive language models, GPT-2 (Radford et al., 2019) and GPT-3\textsuperscript{12} (Brown et al., 2020), and two encoder-decoder language models, BART (Lewis et al., 2020) and T5 (Raffel et al., 2020). For GPT-3, we try both zero-shot and few-shot (4 examples) setups using in-context learning without any fine-tuning. For other models, we fine-tune seven configurations of them on MultiPIT\textsubscript{AUTO}. Table 4 shows the test set results of each model and the diversity of human references measured by Self-BLEU. Among all models, the few-shot setting of GPT-3 achieves the highest BERT-iBLEU score, and the zero-shot setting achieves the second-best number with only 1 point behind, which is not surprising given its size. Compared to GPT-3 generations, human references are

\textsuperscript{11}https://www.github.com/Tiiiger/bert_score

\textsuperscript{12}We use text-davinci-002, which is the most capable GPT-3 model.
Table 5: Automatic evaluation of models fine-tuned on four datasets. Here, BL: BLEU, S-B: Self-BLEU, B-S: BERT-Score, B-iB: BERT-iBLEU. Bold: the best, Underline: the second best.

| Training set | MULTIPI T\textsubscript{AUTO} | Quora | MSCOCO | ParaNMT |
|--------------|-----------------|--------|--------|---------|
|               | BL  | S-B | ↓  | B-S | B-iB | BL  | S-B | ↓  | B-S | B-iB | BL  | S-B | ↓  | B-S | B-iB |
| MULTIPI T\textsubscript{AUTO} | 41.14 | 33.43 | 85.86 | 77.79 | 26.28 | 46.98 | 91.73 | 67.31 | 19.69 | 56.59 | 92.86 | 66.44 | 14.32 | 42.69 | 86.10 | 70.56 |
| Quora       | 32.13 | 32.48 | 83.24 | 76.07 | 28.72 | 34.23 | 87.97 | 73.54 | 15.37 | 51.15 | 88.28 | 61.65 | 8.70 | 28.73 | 79.79 | 67.67 |
| MSCOCO      | 8.37  | 4.83  | 59.25 | 63.47 | 0.97  | 1.26  | 56.52 | 61.55 | 26.14 | 15.46 | 81.00 | 80.30 | 0.70 | 0.59  | 55.52 | 60.56 |
| ParaNMT     | 38.69 | 47.74 | 90.98 | 75.66 | 28.20 | 52.77 | 93.13 | 64.66 | 19.75 | 49.36 | 92.59 | 73.70 | 20.36 | 33.35 | 86.90 | 77.51 |

Figure 6: Human evaluation distributions on generations by model fine-tuned on MULTIPI T\textsubscript{AUTO} or ParaNMT.

Table 6: Human evaluation results on generations by model fine-tuned on MULTIPI T\textsubscript{AUTO} or ParaNMT.

| Model      | Fluency | Semantic Similarity | Diversity |
|------------|---------|---------------------|-----------|
| MULTIPI T\textsubscript{AUTO} | 4.98    | 4.67                | 3.59      |
| ParaNMT    | 4.95    | 4.64                | 3.40      |

5 Other Related Work

Besides the several frequently used paraphrase datasets we mentioned above, here are a few other paraphrase corpora. The MSR Paraphrase corpus (Dolan and Brockett, 2005) contains 5,801 sentence pairs from news articles, but it has a deficiency that skewing toward over-identification (Das and Smith, 2009) and having high lexical overlap (Rus et al., 2014). PPDB (Ganitkevitch et al., 2013) contains over 220 million phrase and lexical paraphrases without any sentence paraphrases. WikiAnswer (Fader et al., 2013) consists of 18 million word-aligned question pairs. However, same as Quora, WikiAnswer is restricted to only questions. In addition, the Semantic Textual Similarity (STS) shared task Cer et al. (2017) measures the degree using ordinal Krippendorff’s alpha (Krippendorff, 2011), which yields 0.31 for fluency,\(^{14}\) 0.56 for semantic similarity, and 0.81 for diversity. All values are considered fair to good (Krippendorff, 2004). Additionally, we perform a manual inspection and observe that model fine-tuned on MULTIPI T\textsubscript{AUTO} generates more diverse kinds of good paraphrases and much fewer poor paraphrases than the one trained on ParaNMT. We define five good paraphrase types and six poor paraphrase types. The definitions and results are shown in Table 7.

---

\(^{13}\)The input is 4 × 50 sentences from each test set.

\(^{14}\)Since over 95% ratings of fluency fall into the same point (see Figure 7 in the Appendix), Krippendorff’s alpha will stay low no matter how the raters agree.
### Table 7: Paraphrase types with examples and statistics observed in the generations by models fine-tuned on MULTIPIT\textsubscript{AUTO} (M\textsubscript{AUTO}) or ParaNMT. Statistics are based on manual inspection of generations by each model on 200 sampled sentences. The shown generation example for each type is by model with the higher value (bold).

| Type               | Definition                                                                 | Generation Example                                                                 | $M_{\text{AUTO}}$ | ParaNMT |
|--------------------|-----------------------------------------------------------------------------|----------------------------------------------------------------------------------|-------------------|---------|
| **Good Paraphrase Type** |                                                                             | Number of occurrences per generation: |                  |         |
| Add New            | Add new phrases while keeping the meaning of given sentence.                | Sent: relax, take it easy. Gen: Relax, take a **deep breath**, and enjoy the moment. | 0.18              | 0.03    |
| Word Syn           | Substitute a word with its synonym (another word).                          | Sent: *Which is the best GRE coaching* centre in Bangalore? Gen: *what is the best gre training* centre... | 0.39              | 0.54    |
| Phrase Syn         | Replace a phrase with synonym or expand a word to a phrase.                 | Sent: it **looks goddamned foolish** to put an oyster on the clam. Gen: Putting an oyster on a clam is **a fucking joke**. | 0.28              | 0.16    |
| Structure          | Use different sentence structures to express the same thing.                | Sent: Two big plates **filled with some tasty looking** food. Gen: Two big plates of food, and the food looks good. | 0.28              | 0.23    |
| Simplification     | Delete minor details or shorten phrases while maintaining the meaning of given sentence. | Sent: Daniel Farke sacked by Norwich after first win of Premier League season **over Brentford.** Gen: Norwich sack Daniel Farke after first win of Premier League season. | 0.39              | 0.26    |
| **Poor Paraphrase Type** |                                                                             | Percentage in 200 generations: |                  |         |
| Copy               | Copy the given sentence.                                                   | Sent: Did you have a good day today? Gen: Did you have a good day today? | 14.5%             | 44%     |
| Small Change       | Only have small changes such as changing article, tense, or prep.           | Sent: FDA approves new test that can detect coronavirus in 45 minutes. Gen: the **fda has approved** a new test... | 5.5%             | 18%     |
| Hallucination      | Add new info that is not common sense or world knowledge.                   | Sent: A dog at a table wearing a birthday hat. Gen: A dog wearing a birthday hat at a **dinner party.** | 2.5%             | 0%      |
| Miss Info          | Miss important info in the given sentence.                                 | Sent: Very sad though that the amazing AJ and Kai will be missing the final. Gen: AJ and Kai will not be in the final. | 1.5%             | 1%      |
| Misinterpret       | Misinterpret or contradict meaning of the given sentence.                  | Sent: Why are most first basemen left handed? Gen: why do most of the first basemen have **left hands?** | 2%               | 14%     |
| Bad Grammar        | Contain grammar error.                                                     | Gen: what is the best headphones for rs 3000? | 0%               | 1%      |

Limitations

While our study shows MULTIPIT\textsubscript{AUTO} improves paraphrase generation quality and diversity, we observe model sometimes generates Twitter-specific artifacts (i.e. “@JoeBiden”). Future work could investigate techniques to mine paraphrases from other social media platforms such as Reddit. Another limitation is that our dataset is only in English, future work could extend this to multilingual as Twitter is used by users from different countries that speak different languages.

Acknowledgments

We thank Yang Chen as well as three anonymous reviewers for their helpful feedback on this work. We also thank Andrew Duffy, Elizabeth Liu, Ian Ligon, Rachel Choi, Jonathan Zhou, Chase Perry, Panya Bhinder for their help on annotations and human evaluation. This research is supported in part by the NSF awards IIS-2144493 and IIS-2112633, ODNI and IARPA via the BETTER program (contract 19051600004), and Figure Eight AI for Everyone Award. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF.
References

Eytan Bakshy, Jake M Hofman, Winter A Mason, and Duncan J Watts. 2011. Everyone’s an influencer: quantifying influence on twitter. In Proceedings of the fourth ACM international conference on Web search and data mining.

Rahul Bhagat and Eduard Hovy. 2013. What is a paraphrase? Computational Linguistics, pages 463–472.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Xiaodong Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In USENIX Security Symposium.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017). Association for Computational Linguistics.

Tuinh Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable paraphrase generation with a syntactic exemplar. In Proceedings of the Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In Proceedings of the Association for Computational Linguistics.

Wenqing Chen, Jidong Tian, Liqiang Xiao, Hao He, and Yaohui Jin. 2020. A semantically consistent and syntactically variational encoder-decoder framework for paraphrase generation. In Proceedings of International Conference on Computational Linguistics.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In Proceedings of Empirical Methods in Natural Language Processing.

Dipanjan Das and Noah A. Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In Proceedings of the Third International Workshop on Paraphrasing (IWP2005).

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. Parasci: A large scientific paraphrase dataset for longer paraphrase generation. In Proceedings of the European Chapter of the Association for Computational Linguistics.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In Proceedings of the Association for Computational Linguistics.

James Ferguson, Colin Lockard, Daniel Weld, and Hannaneh Hajishirzi. 2018. Semi-supervised event extraction with paraphrase clusters. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In Proceedings of the Association for Computational Linguistics.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. ArXiv.
Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations.

Qizhe Xie, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Self-training with noisy student improves imagenet classification. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Wei Xu, Chris Callison-Burch, and William B. Dolan. 2015. SemEval-2015 Task 1: Paraphrase and semantic similarity in Twitter (PIT). In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval).

Congle Zhang, Stephen Soderland, and Daniel S. Weld. 2015. Exploiting parallel news streams for unsupervised event extraction. Transactions of the Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. Proceedings of International Conference on Learning Representation.

Zhilu Zhang and Mert Rory Sabuncu. 2018. Generalized cross entropy loss for training deep neural networks with noisy labels. In Proceedings of Advances in Neural Information Processing Systems.

Jianing Zhou and Suma Bhat. 2021. Paraphrase generation: A survey of the state of the art. In Proceedings of Empirical Methods in Natural Language Processing.
A Annotation Interface

A.1 Crowdsourcing
Figure 9 and Figure 10 display screenshots of the instruction and an example question of our crowdsourcing annotation for MULTIPIT\textsubscript{CROWD}.

A.2 Expert
Figure 11 displays a screenshot of the instruction of our expert annotation for MULTIPIT\textsubscript{EXPERT}.

B Data Pre-processing
Both PIT-2015 (Xu et al., 2015) and Twitter URL (Lan et al., 2017) datasets share similar pre-processing steps that introduced tokenization and sentence splitting errors. Moreover, PIT-2015 contains some spam patterns, such as “Follow Me PLEASE”. We improved the quality of our dataset by fixing the pre-processing methods and removing spam patterns. More importantly, we split tweets into sentences to get cleaner paraphrases (see Table 8 for an example), without added noises from extra sentences in the tweet. We improve the sentence splitting script by Xu et al. (2015) and tokenization script by O’Connor et al. (2010) used in prior work with a number of errors fixed: (1) Emojis and most symbols are cleaned while punctuation are kept; (2) Extremely short sentences (<5 tokens) are filtered out while remaining sentences are deduplicated by comparing lowercased strings w/o any punctuation.

C Implementation Details
We use HuggingFace Transformers (Wolf et al., 2020) version of all pre-trained models. We use Python 3.8, PyTorch 1.9.0, and Transformers 4.12.0. For all experiments, we use 4 × 48GB NVIDIA A40 GPUs.

Paraphrase Identification. Hyperparameters for fine-tuning models in paraphrase identification experiments are given in Table 9.

| Hyperparameter              | Assignment |
|-----------------------------|------------|
| Max epochs                  | 5          |
| Eval steps                  | 500        |
| Effective batch size        | 32         |
| Learning rate optimizer     | AdamW      |
| Adam epsilon                | 1e-8       |
| Weight decay                | 0.01       |
| Learning rate               | {1e-5, 2e-5, 3e-5, 5e-5} |
| Learning rate decay         | Linear     |
| Warmup ratio                | 0.06       |

Table 9: Hyperparameters for paraphrase identification. We choose learning rate range based on Liu et al. (2019) consider learning rates ∈ {1e-5, 3e-5, 5e-6, 8e-6} following He et al. (2021). We fine-tune for 5 epochs and eval every 500 steps (every epoch if total training steps is less than 1500) on the dev set. The only hyperparameter we tune is the learning rate and use $F_1$ on the dev set for model selection.

For InferSent and ESIM models, we use their original implementation initialized with GloVe embedding (Pennington et al., 2014), and also only tune the learning rate based on the dev set.

Paraphrase Generation. Hyperparameters for fine-tuning models in paraphrase generation experiments are given in Table 10.

| Hyperparameter              | Assignment |
|-----------------------------|------------|
| Max epochs                  | 5          |
| Eval steps                  | 128        |
| Effective batch size        | 128        |
| Learning rate optimizer     | AdamW      |
| Adam epsilon                | 1e-8       |
| Weight decay                | 0.01       |
| Learning rate               | {1e-4, 3e-4, 1e-5, 3e-5} |
| Learning rate decay         | Linear     |
| Warmup ratio                | 0.06       |

Table 10: Hyperparameters for paraphrase generation. We use perplexity on the dev set for model selection.

As ParaNMT contains only lowercase letters, we lowercase the input and references for generation and evaluation of the model fine-tuned on ParaNMT and lowercase the other models’ generations while evaluating on ParaNMT.
D GPT-3 Setup

D.1 Hyperparameters
We use the text-davinci-002 GPT-3 model for paraphrase generation. To generate paraphrase, we use the following hyperparameters: temperature=1, max tokens=100, top-p=0.9, best of=1, frequency penalty=0.5, presence penalty=0.5, based on Chakrabarty et al. (2021).

D.2 Prompts

Zero-shot setting: Your task is to generate a diverse paraphrase for a given sentence.
Sentence: {sentence}
Paraphrase: {paraphrase}

Few-shot setting: You will be presented with examples of some input sentences and their paraphrases. Your task is to generate a diverse paraphrase for a given sentence.
Sentence: Mike Bloomberg is sending $18 million from his defunct presidential campaign to the DNC.
Paraphrase: Mike Bloomberg is transferring $18M from his campaign to DNC, stretching campaign finance law.

Sentence: Google Assistant on Android can read web pages to you
Paraphrase: Google Assist lets your Android devices read entire web pages aloud

Sentence: Charlie Patino scored a goal on his debut!
Paraphrase: Charlie Patino’s debut and he capped it off with a goal.

Sentence: khem birch is the difference maker for the raptors this game
Paraphrase: Khem Birch may be the MVP tonight for the Raptors.

Sentence: {sentence}
Paraphrase: {paraphrase}

E Generation Dataset Statistics

Table 11 presents the detailed statistics of MULTI-PIT AUTO, Quora, MSCOCO and ParaNMT.

| Genre       | M AUTO | Quora   | MSCOCO | ParaNMT |
|-------------|--------|---------|--------|---------|
| Sentence Length | 11.34  | 9.66    | 10.49  | 11.33   |
| Sentence BLEU  | 24.48  | 26.37   | 9.30   | 24.85   |

Table 11: Statistics of datasets for paraphrase generation. We calculate sentence length based on the number of tokens per unique sentence. As ParaNMT is too large, we sample 500K for the calculation of sentence length and BLEU. W/o BF denotes without BLEU filtering.

F Further Paraphrase Generation Experiments

Table 12: Spearman correlations with human evaluation on 100 generations on MULTI-PIT NMR (50 by model trained on MULTI-PIT AUTO and 50 by model trained on ParaNMT). Here, ***: p < 0.0001, **: p < 0.001, *: p < 0.01. Overall is the summation score of all three aspects.

| Metric             | Fluency | Semantic | Diversity | Overall |
|--------------------|---------|----------|-----------|---------|
| BLEU               | 0.212   | 0.209    | -0.233    | -0.091  |
| Self-BLEU ↓        | 0.068   | 0.412*** | -0.655*** | -0.452***|
| BERT-Score         | 0.062   | 0.523*** | -0.722*** | -0.507***|
| BERT-iBLEU         | -0.166  | -0.089   | 0.370***  | 0.381*** |

Table 13: Spearman correlations with human evaluation on all 400 generations. Here, ***: p < 0.0001, **: p < 0.001, *: p < 0.01. Overall is the summation score of all three aspects.

| Metric             | Fluency | Semantic | Diversity | Overall |
|--------------------|---------|----------|-----------|---------|
| Self-BLEU ↓        | 0.043   | 0.319*** | -0.638*** | -0.491***|
| BERT-Score         | 0.070   | 0.436*** | -0.744*** | -0.561***|
| BERT-iBLEU         | -0.036  | -0.096   | 0.346***  | 0.339*** |

Correlation Analysis. With human evaluation, we calculate Spearman correlation to evaluate automatic metric quality. Since the four test sets have different numbers of references and MULTI-PIT NMR has the most number of references, to evaluate BLEU, we examine 100 generations on MULTI-PIT NMR (50 by T5 large fine-tuned on MULTI-PIT AUTO and 50 by T5 large fine-tuned on ParaNMT). Results are shown in Table 12. BLEU gets a weak correlation around |0.2| with all as-
Diversity

| 33.8% | 22.9% | 14.4% | 16.8% | 12.1% |

Semantic Similarity

| 78.7% | 13.9% | 4.8% | 2.2% | 0.4% |

Fluency

| 97.7% | 0.3% | 0.2% | 0.1% | 1.7% |

Figure 7: Label distribution of 1200 ratings on 400 generations by models fine-tuned on MultiPIT_{AUTO} and ParaNMT.

BLEU Filtering. We evaluate different BLEU thresholds on the dev set of MultiPIT_{AUTO} as shown in Figure 8. The model achieves the best performance at the threshold of 14, which is used across our experiments.

Next, we compare model performance on all four datasets with and without BLEU filtering. Results are presented in Table 14. Applying BLEU filtering improves model performance with higher BERT-iBLEU on all datasets.

Impact of Definition. We investigate how different paraphrase definitions affect generation performance. As shown in Table 15, model fine-tuned on MultiPIT_{AUTO} outperforms fine-tuning on the loosely defined data such as MultiPIT_{CROWD}.

Table 14: In-domain test set results of fine-tuning model on data with or without BLEU filtering. w/o BF denotes without BLEU filtering.

| Data | Size | BL | S-B ↓ | B-S | B-iB |
|------|------|----|-------|-----|-----|
| MultiPIT_{CROWD} | 26,091 | 36.15 | 32.09 | 85.53 | 74.19 |
| MultiPIT_{CROWD} | 326,517 | 45.55 | 37.90 | 85.80 | 74.12 |
| MultiPIT_{AUTO} | 136,645 | 41.14 | 33.34 | 85.86 | 77.79 |

Table 15: Test set results of models fine-tuned on data constructed with different paraphrase definitions. MultiPIT_{CROWD} contains its paraphrase pairs. MultiPIT_{AUTO-}\_CROWD is the automatically identified paraphrase pairs by the identifier fine-tuned on MultiPIT_{CROWD}.

G Examples

Generation Examples. Table 16 presents generation examples by GPT-3 and fine-tuned T5_{large} on MultiPIT_{NMR}.

Table 17 presents generation examples by T5_{large} fine-tuned on MultiPIT_{AUTO}, Quora, MSCOCO, and ParaNMT.

Multi-Reference Examples. Table 18 displays three examples from the MultiPIT_{NMR} test set.

H Human Evaluation Details

We display our human evaluation instruction for each aspect (fluency, semantic similarity, diversity) in Figure 12,13,14.
Table 16: Paraphrase generation examples by GPT-3 and fine-tuned T5\textsubscript{large} on MULTI PIT NMR.
| Source origin: | MULTI PIT NMR |
| Source sentence: | In honor of Veterans Day, we salute all of the men and women who have served in America’s armed forces. |
| Generated sentences: | |
| $T_5^{MULTI PIT AUTO}$ | Today is Veterans Day, a day to remember those who have served. |
| $T_5^{Quora}$ | On this Veterans Day, we salute all the men and women who have served our country. |
| $T_5^{MSCOCO}$ | A group of men and women standing next to each other. |
| $T_5^{ParaNMT}$ | in honor of veterans’ day, we salute all the men and women who served in america’s armed forces. |

| Source origin: | Quora |
| Source sentence: | What is the scope for a food technologist in India? |
| Generated sentences: | |
| $T_5^{MULTI PIT AUTO}$ | What is the job outlook for a food technologist in India? |
| $T_5^{Quora}$ | What is the scope of food technology in India? |
| $T_5^{MSCOCO}$ | a man in a white shirt is working in a kitchen |
| $T_5^{ParaNMT}$ | what is the role of a food technologist in india? |

| Source origin: | MSCOCO |
| Source sentence: | A bear is taking a walk through the forest. |
| Generated sentences: | |
| $T_5^{MULTI PIT AUTO}$ | A bear is taking a walk in the woods. |
| $T_5^{Quora}$ | What is a bear doing in the forest? |
| $T_5^{MSCOCO}$ | A brown bear walking through a wooded area. |
| $T_5^{ParaNMT}$ | the bear is walking through the forest. |

| Source origin: | ParaNMT |
| Source sentence: | someone speaks in english over there. |
| Generated sentences: | |
| $T_5^{MULTI PIT AUTO}$ | Someone has to speak english in this room. |
| $T_5^{Quora}$ | What is the best way to learn English from someone who speaks fluently? |
| $T_5^{MSCOCO}$ | A man standing in front of a wall with a sign on it. |
| $T_5^{ParaNMT}$ | someone’s speaking english. |

Table 17: Paraphrase generation examples by $T_5^{large}$ fine-tuned on MULTI PIT AUTO, Quora, MSCOCO, and ParaNMT on each test set.
Source sentence: @GovStitt Please grant clemency for Julius Jones, an innocent man scheduled for execution in your state.

References:
1. @GovStitt Almost like murder if you execute the innocent Julius Jones tomorrow Governor.
2. @GovStitt Please commute the sentence of Julius Jones.
3. @GovStitt I join the many, many voices urging you to do the right thing and grant clemency to Julius Jones.
4. @GovStitt Please save the life of Julius Jones.
5. @GovStitt please do the right thing and don’t execute julius Jones.
6. @OKFirstLady Please urge your husband @GovStitt to grant Julius Jones clemency.
7. @GovStitt Respectfully I urge you to exercise all powers vested in your office to grant clemency to Mr. Julius Jones.
8. @GovStitt Please stop the needless execution of Julius Jones.

Source sentence: Austria imposes COVID-19 lockdown that Applies only to the unvaccinated

References:
1. Austria decided to have a lockdown of the unvaccinated.
2. Unvaccinated people forced into lockdown in Austria.
3. Austria enters hard-to-enforce Covid-19 lockdown for the unvaccinated.
4. Austria orders non-vaccinated people into COVID-19 lockdown.
5. Lockdown takes effect for unvaccinated people in Austria.
6. Unvaccinated People in Austria Are Now Being Put in Lockdown.
7. Austria orders lockdown for residents who have not received COVID-19 vaccine.
8. Austria brings back COVID-19 lockdown, this time for the unvaccinated.

Source sentence: Turn off Bluetooth when you are not using it.

References:
1. Reminder to turn off your blue tooth when not in use.
2. Turn your Bluetooth off while you’re not using it.
3. Best to turn off Bluetooth when you can.
4. Always turn off your Bluetooth when you’re not using it.
5. Whenever you don’t absolutely need it, you should go ahead and turn off your Bluetooth.
6. Keep Bluetooth off when you are not using it.
7. Whenever you don’t need BlueTooth, you should turn it off.
8. If you don’t need your Bluetooth enabled, then turn it off!

Table 18: Three examples from MULTI PIT NMR.
Definition

- A and B are a paraphrase pair:
  - Case 1: A and B are completely equivalent (mean the same thing, though differ in expression):
    A: Chad from World of Jenks is so adorable.
    B: Chad from world of Jenks is the absolute cutest!
    Two sentences convey the same meaning (liking Chad), while their expression are different.

  - Case 2: A and B are mostly equivalent, but some unimportant details differ:
    Although some unimportant details differ (job title, data, employment environment, etc.).
    A: 13 May 2013 Roberto Mancini sacked as Man City Manager.
    B: Roberto Mancini just got sacked after that shit season.
    Two sentences are mostly equivalent (Roberto Mancini was sacked).

- A and B are a non-paraphrase pair:
  - Case 1: A and B are not equivalent, though share some details:
    Though sharing some details (new Macbook Air).
    A: Apple unveils new Macbook Air and a Mac Pro.
    B: I was pumped for the new macbook air.
    Two sentences are talking about different things (Apple unveil v.s. I was pumped).

  - Case 2: A and B are not equivalent, though on a same or similar topic:
    Two sentences are talking about different things, though on the same or similar topic (8 Mile / movies).
    A: Ok good, the end of 8 Mile is on.
    B: I always get the movies 8 mile and green mile mixed up.

  - Case 3: A and B are not equivalent - they are on different topics:
    The two sentences are on different topics.
    A: The name Lydia sounds Spanish.
    B: Lydia and I send the creepiest snap chat videos to each other.

Figure 9: Instruction of our crowdsourcing annotation on the Figure Eight platform for creating MULTI PIT CROWD.
Figure 10: An example question of our crowdsourcing annotation on the Figure Eight platform for creating MULTIPIT CROWD.

| Sentence A | Sentence B |
|------------|------------|
| A Green-Haired Turtle That Can Breathe Through Its Genitals #NYT | This is a face that the world needs. |

What’s the relationship between Sentence A and Sentence B?

- **A and B is a paraphrase pair**
  - A and B are equivalent (convey the same meaning, though differ in expressing or some unimportant details).

- **A and B is a non-paraphrase pair**
  - A and B are talking about different things, though sharing some details or on the same/similar topic.

Comments (Optional)

*If you have any comment about this HIT, please type it here*
Instruction

A and B is a **paraphrase** pair if:

- **Case 1**: A and B are completely equivalent (mean the same thing, though differ in expression):

  A: Chad from World of Jenks is so adorable.
  B: Chad from World of Jenks is the absolute cutest!
  **Explanation**: Two sentences convey the same meaning (liking Chad) using different expressions.

- **Case 2**: B keeps the main meaning of A, but deletes some minor details from A:

  A: Sweden's first female PM Magdalena Andersson, resigns on day one!
  B: Swedish PM Magdalena Andersson resigns hours after taking job.
  **Explanation**: The main content of A is about Magdalena Andersson resigning on day one, so deleting "first female" is fine and considered as simplification.

- **Case 3**: B keeps the main meaning of A, and add new information based on commonsense or world knowledge:

  A: Facebook announces it will be changing its name to Meta.
  B: Facebook relaunches itself as 'Meta' in a clear bid to dominate the metaverse.
  **Explanation**: The new added "to dominate the metaverse" is world knowledge as many people know it. We consider B as a paraphrase of A.

A and B is a **non-paraphrase** pair if:

- **Case 1**: B adds new information that requires fact-checking:

  A: 100% of the 140,000 U.S. jobs lost in December were held by women.
  B: In fact women lost 111% of the jobs in December because men gained 16,000 jobs.
  **Explanation**: Even though both sentences are talking about the same thing, but B introduces new information that is not commonsense or world knowledge.

- **Case 2**: A and B share some details but focus on different things:

  A: Apple unveils new Macbook Airs and a Mac Pro.
  B: I was pumped for the new macbook air.
  **Explanation**: Two sentences are talking about different things: "Apple unveils" vs "I was pumped".

- **Case 3**: A and B are on different topics:

  A: Rhode Island Senate approves marriage equality by vote of 26-12
  B: So glad to hear that the Kings are staying in Sac.
  **Explanation**: Both sentences are completely irrelevant.

Figure 11: Instruction of our expert annotation for creating MULTI-PIT-EXPERT.
Fluency

To rate Fluency, you just answer the following question: Is sentence 2 natural and fluent? Does it have grammatical error?

Here is each score (1 to 5) represents:

5 - Without any grammatical error
4 - Fluent and has one minor grammatical error that does not affect understanding, e.g. Practising is the best way to learn programming., Is apples good?
3 - Basically fluent and has two or more minor grammatical errors or one serious grammatical error that does not have strong impact on understanding, e.g. Here are some good book for read.
2 - Can not understand what it means but it is still in the form of human language, e.g. what is the best movie of movie
1 - Non-sense composition of words and not in the form of human language, e.g. how world war iii world war?

Note 1: hashtag # or @ doesn’t count as grammatical error (e.g. @AskTarget Why did you pull #JohnnyTheWalrus? is a 5)

Note 2: We ignore lettercase and punctuation issue.

Figure 12: Instruction for rating fluency aspect in our human evaluation.

Semantic Similarity

To rate Semantic Similarity, you just answer the following question: Is sentence 2 semantically close to sentence 1?

Here is each score (1 to 5) represents:

5 - Keeps the main meaning of sentence 1. Correct interpretation, new addition of info or implication based on commonsense or world knowledge, and simplification by deleting unimportant details are fine.
4 - Has a similar meaning of sentence 1 but contains further aftermath, or misses a small part of the main content in sentence 1.
3 - Misses half or more than half of the main content in sentence 1, or adds hallucination or new info that requires fact-checking.
2 - Misinterprets, misrepresents, contradicts or doesn’t reflect the meaning of sentence 1 correctly.
1 - Doesn’t make sense or the main content is different from sentence 1.

Note: We ignore lettercase and punctuation issue.

Figure 13: Instruction for rating semantic similarity aspect in our human evaluation.
Diversity

To rate Diversity, you just answer the following question: Is sentence 2 different from sentence 1?

Here is each score (1 to 5) represents:

5 - Uses more than 1 score 4 and 3 changes. Note: must contain at least 1 4 type changes.

4 - Uses one of the following types of change 1 time:
   - change of sentence structure
   - simplifying
   - adding new phrase or meaningful word
   - rearranging word order
   - using idiomatic expressions
   - change of part of speech
   - expanding a word in detail
   - synonym replacement phrase-wise (e.g. "10 years" <-> "a decade", "hotel employee" <-> "bell boy")

Or uses synonym replacement word-wise more than 2 times. Note: mark 5 if sentence 1 contains less than 6 words.

3 - Uses synonym replacement word-wise 1 or 2 times.

2 - Very simple grammatical changes such as:
   - determiners changes (remove or add "the", "the" <-> "a", "a" <-> "one", "that" <-> "it", "his" <-> "this", "some" <-> "any", ...)
   - contraction changes ("n't" <-> "not", "re" <-> "are", "will" <-> "ll", "let's" <-> "let us", ...)
   - singular and plural switching ("a" <-> "some", "are" <-> "is", add "es/s", ...)
   - tense changes ("is" <-> "was", "did" <-> "have done", "is doing" <-> "do", "will" <-> "would", ...)
   - number and text switching ("7" <-> "seven", "five" <-> "5", ...)
   - preposition changes (remove or add "on", "at" <-> "on", "upon" <-> "on", "of" <-> "for", ...)
   - adding or removing conjunction word or meaningless word ("... that ..." <-> .....", "And ...
   "", "just", ...)
   - other cases ("to" <-> "will")

Note: Multiple 2 changes is still a 2.

1 - Copies sentence 1 completely.

Note: we ignore lettercase and punctuation issue.

Figure 14: Instruction for rating diversity aspect in our human evaluation.