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 (MULTIPIT) corpus that consists of a total of 130k sentence pairs with crowdsoursing ($\text{MULTIPIT}_{\text{CROWD}}$) and expert ($\text{MULTIPIT}_{\text{EXPERT}}$) annotations using two different paraphrase definitions for paraphrase identification, in addition to a multi-reference test set ($\text{MULTIPIT}_{\text{NMR}}$) and a large automatically constructed training set ($\text{MULTIPIT}_{\text{AUTO}}$) 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 $\text{MULTIPIT}_{\text{AUTO}}$ 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 argumentation to improve information extraction (Zhang et al., 2015; Ferguson et al., 2018) and question answering systems (Gan and Ng, 2019).

Figure 1: Two sets of paraphrases in MULTIPIT, discussing a trending topic or a news article, respectively.

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).

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 9301 - 9323
December 7-11, 2022

	Topic Domains	#Train	#Dev	#Test	Sent/Tweet Len	%Paraphrase	#Trends/URLs	#Uniq Sent	%Multi-Ref
Our Multi-Topic Paraphrase in Twitter (MULTIPIT$_{\text{CROWD}}$) Dataset									
Trends	Sports	25,255	3,157	3,157	10.24 / 13.79	40.52%	1,201	34,786	17.89%
	Entertainment	11,547	1,443	1,444	10.44 / 13.80	62.33%	610	15,784	18.11%
	Event	8,624	1,078	1,079	10.86 / 15.32	82.83%	359	11,746	17.75%
	Others	17,751	2,219	2,219	10.41 / 14.56	67.16%	817	24,286	18.33%
URL	Science/Tech	7,384	923	923	10.94 / 19.17	46.13%	1,032	10,327	17.74%
	Health	9,123	1,140	1,141	11.29 / 21.68	46.78%	1,298	12,772	17.86%
	Politics	7,981	998	998	10.95 / 18.48	56.56%	1,063	10,999	17.68%
	Finance	4,552	569	569	11.19 / 23.08	18.96%	554	5,907	20.13%
	Total	**92,217**	**11,527**	**11,530**	**10.62 / 16.10**	**53.73%**	**6,934**	**124,438**	**18.65%**
Our MULTIPIT$_{\text{EXPERT}}$ Dataset		4,458	555	557	12.08 / 17.02	53.11%	200	5,743	100%
Existing Twitter Paraphrase Datasets									
PIT-2015 (Xu et al.)		13,063	4,727	972	11.9 / –	30.60%	420	19,297	24.67%
Twitter URL (Lan et al.)		42,200	–	9,324	– / 14.8	22.77%	5,187	48,906	23.91%

Table 1: Statistics of MULTIPIT$_{\text{CROWD}}$ and MULTIPIT$_{\text{EXPERT}}$ datasets. The sentence/tweet lengths are calculated based on the number of tokens per unique sentence/tweet. %Multi-Ref denotes the percentage of source sentences with more than one paraphrase. Compared with prior work, our MULTIPIT$_{\text{CROWD}}$ dataset has a significantly larger size, a higher portion of paraphrases, and a more balanced topic distribution.

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$_{\text{CROWD}}$, a large crowdsourced set of 125K sentence pairs that is useful for tracking information on Twitter, and MULTIPIT$_{\text{EXPERT}}$, 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 F_1 for automatic paraphrase identification. In addition, we construct a continually growing paraphrase dataset, MULTIPIT$_{\text{AUTO}}$, by applying the automatic identification model to unlabelled Twitter data. Empirical results and analysis show that generation models fine-tuned on MULTIPIT$_{\text{AUTO}}$ 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] https://www.kaggle.com/c/quora-question-pairs

2 Multi-Topic PIT Corpus

In this section, we present our data collection and annotation methodology for creating MULTIPIT$_{\text{CROWD}}$ and MULTIPIT$_{\text{EXPERT}}$ 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 topic[3] (year 2014–2015) or the same URL (year 2017–2019) retrieved through Twitter public APIs[4] 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

[3] https://www.twitter.com/explore/tabs/trending

[4] https://developer.twitter.com/en/docs/twitter-api

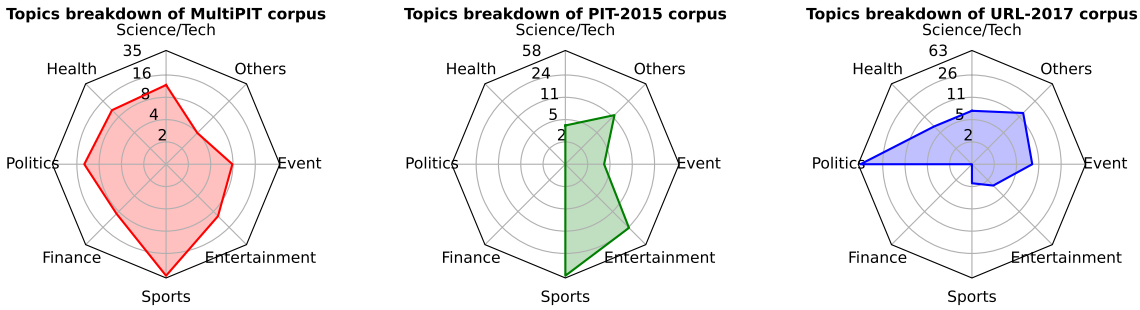

Figure 2: Topic breakdown on 100 randomly sampled sentence pairs from MULTIPIT$_{\text{CROWD}}$, PIT-2015 and Twitter-URL. Our MULTIPIT$_{\text{CROWD}}$ corpus has a more balanced topic distribution.

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 MULTIPIT$_{\text{CROWD}}$.

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 MULTIPIT$_{\text{CROWD}}$, 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,

[5]For example, URL https://www.nytimes.com/2019/08/09/science/komodo-dragon-genome.html belongs to science topic.

[6]https://www.appen.com/

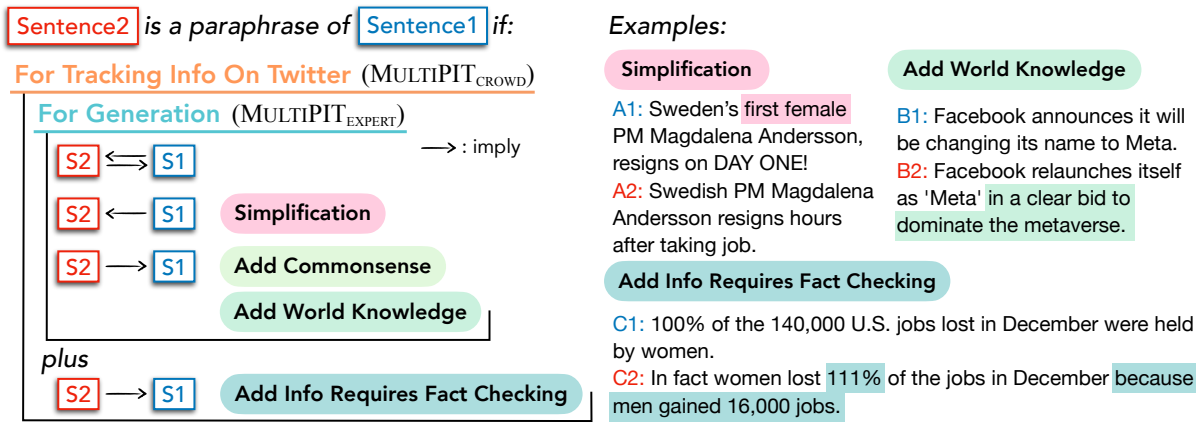

Figure 3: Two different paraphrase definitions used for creating MULTIPIT$_{\text{CROWD}}$ and MULTIPIT$_{\text{EXPERT}}$, with examples. The difference between the two criteria is whether considering *Sentence2* that contains new information that requires fact-checking as a paraphrase of *Sentence1*.[7]

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 MULTIPIT$_{\text{CROWD}}$, PIT-2015 (Xu et al., 2015) and Twitter-URL (Lan et al., 2017). Figure 2 shows the results of the manual inspection. MULTIPIT$_{\text{CROWD}}$ 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 periodd (§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 MULTIPIT$_{\text{EXPERT}}$ with expert annotations, using a stricter paraphrase definition than the one used in MULTIPIT$_{\text{CROWD}}$. 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 MULTIPIT$_{\text{CROWD}}$ 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 MULTIPIT$_{\text{EXPERT}}$, 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 MULTIPIT$_{\text{NMR}}$, 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

[7]The example C1 and C2 is on the more extreme side of the "loose" paraphrase criterion from the linguistic perspective, more average cases are shown in Figure 1.

Model	#Para.	MULTIPIT$_{\text{CROWD}}$					MULTIPIT$_{\text{EXPERT}}$				
		LR	Precision	Recall	F_1	Accuracy	LR	Precision	Recall	F_1	Accuracy
ESIM	17M	4e-4	89.55	70.15	78.67	82.15	4e-4	47.07	91.73	62.22	49.19
Infersent	47M	1e-3	87.03	87.57	87.29	86.47	1e-3	45.87	**98.43**	62.58	46.32
T5$_{\text{base}}$	220M	1e-4	89.21	93.76	91.43	90.67	1e-4	71.96	83.86	77.45	77.74
T5$_{\text{large}}$	770M	1e-5	90.36	93.58	91.94	91.29	1e-4	79.78	85.43	82.51	83.48
BERT$_{\text{base}}$	109M	3e-5	88.59	91.24	89.90	89.12	2e-5	71.66	86.61	78.43	78.28
BERT$_{\text{large}}$	335M	2e-5	88.73	93.17	90.90	90.10	2e-5	72.22	87.01	78.93	78.82
RoBERTa$_{\text{large}}$	355M	2e-5	**90.81**	92.70	91.74	91.14	2e-5	77.01	83.07	79.92	80.97
BERTweet$_{\text{large}}$	355M	2e-5	89.72	**93.95**	91.79	91.08	2e-5	82.47	81.50	81.98	83.66
ALBERTV2$_{\text{xxlarge}}$	235M	1e-5	90.36	92.96	91.64	91.00	2e-5	**82.68**	82.68	82.68	84.20
DeBERTaV3$_{\text{large}}$	400M	5e-6	90.46	93.59	**92.00**	**91.36**	5e-6	82.56	83.86	**83.20**	**84.56**

Table 2: Results on the test sets of MULTIPIT$_{\text{CROWD}}$ and MULTIPIT$_{\text{EXPERT}}$. Models are fine-tuned on the corresponding training set. DeBERTaV3$_{\text{large}}$ performs the best on both datasets. LR: learning rate.

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 MULTIPIT$_{\text{CROWD}}$ and MULTIPIT$_{\text{EXPERT}}$ 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 MULTIPIT$_{\text{CROWD}}$ 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 MULTIPIT$_{\text{CROWD}}$ and MULTIPIT$_{\text{EXPERT}}$ 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.

Method	Data	P.	R.	F_1	Acc.
Fine-tuning	M$_{\text{C}}$	61.81	**88.58**	72.82	69.84
Fine-tuning	M$_{\text{E}}$	82.56	83.86	83.20	84.56
Fine-tuning	M$_{\text{C}}$ + M$_{\text{E}}$	62.99	87.80	73.36	70.92
+ Filtering	M$_{\text{C}}$ + M$_{\text{E}}$	77.24	88.19	82.35	82.76
+ Flipping	M$_{\text{C}}$ + M$_{\text{E}}$	**83.40**	85.04	**84.21**	**85.46**

Table 3: Results of different methods on the test set of MULTIPIT$_{\text{EXPERT}}$. M$_{\text{C}}$: MULTIPIT$_{\text{CROWD}}$, M$_{\text{E}}$: MULTIPIT$_{\text{EXPERT}}$. We use DeBERTaV3$_{\text{large}}$ in the experiments.

3.2 Results

Table 2 presents results for the models fine-tuned on each dataset. DeBERTaV3$_{\text{large}}$ achieves the best results with 92 F_1 on MULTIPIT$_{\text{CROWD}}$ and 83.2 F_1 on MULTIPIT$_{\text{EXPERT}}$. Transformer-based models consistently outperform BiLSTM-based models, especially on MULTIPIT$_{\text{EXPERT}}$.

Beyond Fine-tuning. As MULTIPIT$_{\text{CROWD}}$ is a large-scale dataset annotated with a loose paraphrase definition, we test whether leveraging these "noisy" data improves model performance on MULTIPIT$_{\text{EXPERT}}$. To reduce the noise that comes from the difference in definitions, we first adjust the labeling threshold for MULTIPIT$_{\text{CROWD}}$ 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 MULTIPIT$_{\text{EXPERT}}$ and use it to go through MULTIPIT$_{\text{CROWD}}$ 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. $0 \rightarrow 1$).[8] Next, we fine-tune a new

[8]We perform a small grid search on λ 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.

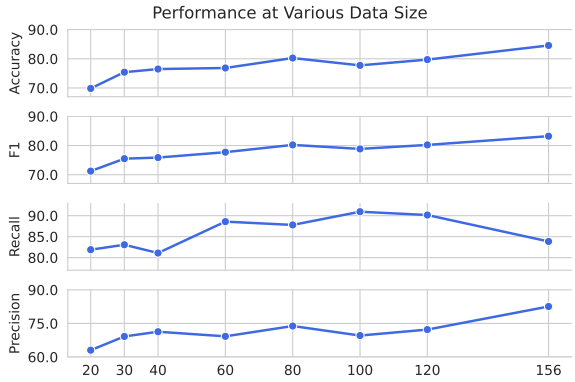

Figure 4: Test set performance of model fine-tuned on varying amounts of data in MULTIPIT$_{\text{EXPERT}}$.

model on the combination of MULTIPIT$_{\text{EXPERT}}$ and the re-labeled MULTIPIT$_{\text{CROWD}}$. The experimental results are shown in Table 3. Compared to fine-tuning on MULTIPIT$_{\text{EXPERT}}$, adding the original MULTIPIT$_{\text{CROWD}}$ 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 MULTIPIT$_{\text{AUTO}}$ (§4).

3.3 Impact of Data Size

Figure 4 shows test set performance of DeBERTaV3$_{\text{large}}$ fine-tuned on different amounts of data in MULTIPIT$_{\text{EXPERT}}$. 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 MULTIPIT$_{\text{AUTO}}$, an automated continual growing dataset for paraphrase generation. We show that the model fine-tuned on MULTIPIT$_{\text{AUTO}}$ generates more diverse and high-quality paraphrases than other paraphrase datasets.

4.1 Comparison with Existing Datasets

MSCOCO (Lin et al., 2014), and **ParaNMT** (Wieting and Gimpel, 2018), and **Quora**[9] are three widely used datasets in paraphrase generation research (Zhou and Bhat, 2021). The 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. 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. 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.

MULTIPIT$_{\text{AUTO}}$. 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 MULTIPIT$_{\text{AUTO}}$,[10] 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 MULTIPIT$_{\text{NMR}}$ test set (§2.5) for evaluation. As the test set and MULTIPIT$_{\text{AUTO}}$ come from the same time period, we filter out sentence pairs in MULTIPIT$_{\text{AUTO}}$ 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 MULTIPIT$_{\text{AUTO}}$, 47K for Quora, 275K for MSCOCO, and 443K for 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: **BLEU** (Papineni et al., 2002), **Self-BLEU** (Liu et al., 2021), **BERT-Score** (Zhang et al., 2020), and **BERT-iBLEU** (Niu et al., 2021). Self-BLEU is BLEU

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

[10] Future identified paraphrase pairs will be released every month.

Model	#Para.	LR	BL	S-B ↓	B-S	B-iB
$\text{GPT-2}_{\text{small}}$	117M	3e-5	41.15	51.38	88.18	65.23
$\text{GPT-2}_{\text{large}}$	774M	3e-5	42.89	39.61	86.16	74.01
$\text{BART}_{\text{base}}$	139M	1e-5	46.91	46.38	87.65	71.40
$\text{BART}_{\text{large}}$	406M	1e-5	**47.22**	38.26	86.40	75.17
T5_{small}	60M	3e-4	38.27	52.16	**88.32**	68.37
T5_{base}	220M	1e-4	42.10	46.43	87.75	72.29
T5_{large}	770M	1e-4	41.14	**33.34**	85.86	**77.79**
$\text{GPT-3}_{\text{zero-shot}}$	175B	-	28.05	31.68	86.66	80.16
$\text{GPT-3}_{\text{few-shot}}$	175B	-	30.17	30.93	86.84	81.13

Diversity (S-B ↓)	*Min.*	*Avg.*	*Max.*
Human Reference	6.52	$17.45_{\pm 9.14}$	34.06

Table 4: Test set results of different transformer models fine-tuned on MULTIPIT$_{\text{AUTO}}$, except GPT-3, where in-context learning is used. BL: BLEU, S-B: Self-BLEU, B-S: BERT-Score, B-iB: BERT-iBLEU. LR: learning rate. **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.

computed between the source sentence and the output, which measures surface-form diversity. BERT-Score is also calculated between the source sentence and the output, measuring semantic similarity. BERT-iBLEU is a harmonic mean of BERT-Score and 1−Self-BLEU, encouraging both semantic similarity and diversity. We use SacreBLEU (Post, 2018) to compute BLEU and *bert-score*[11] to compute BERT-Score.

4.3 Generation Models

We consider two autoregressive language models, **GPT-2** (Radford et al., 2019) and **GPT-3**[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$_{\text{AUTO}}$. 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 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_{large} outperforms others with the best Self-BLEU and BERT-iBLEU scores. Although $\text{BART}_{\text{large}}$ gets the highest BLEU score, our experiments in Appendix F show BERT-iBLEU has the best correlation with human evaluation. We thus use T5_{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.

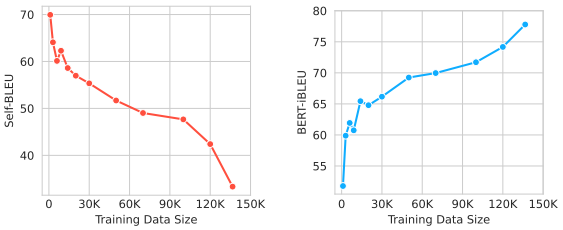

Figure 5: Test set performance of model fine-tuned on varying amount of data in MULTIPIT$_{\text{AUTO}}$, in terms of Self-BLEU (lower is better) and BERT-iBLEU.

Impact of Data Size. Figure 5 shows test set performance of T5_{large} fine-tuned on different amount of data in MULTIPIT$_{\text{AUTO}}$ 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$_{\text{AUTO}}$ corpus.

4.4 Cross-Dataset Generalization

Building a paraphrase generation model that generalizes to new data is always an ambitious goal. To better understand the generalizability of each dataset, we fine-tune T5_{large} on MULTIPIT$_{\text{AUTO}}$, Quora, MSCOCO, and ParaNMT separately and evaluate their performance across datasets. For fair comparisons, we use the same architecture, T5_{large}, in this experiment. Appendix G displays examples generated by these models on each dataset.

Table 5 presents **automatic** evaluation of test set performance across all four datasets. As MULTIPIT$_{\text{AUTO}}$ and ParaNMT consist of sentences in different styles, models trained on them have better generalizability, achieving the best cross-domain

[11] https://www.github.com/Tiiiger/bert_score We use DeBERTa-xlarge-mnli since it has the best correlation with human evaluation according to Zhang et al. (2020).

[12] We use text-davinci-002, which is the most capable GPT-3 model.

Training set \ Test set	MULTIPIT$_{NMR}$				Quora				MSCOCO				ParaNMT			
	BL	S-B ↓	B-S	B-iB	BL	S-B ↓	B-S	B-iB	BL	S-B ↓	B-S	B-iB	BL	S-B ↓	B-S	B-iB
MULTIPIT$_{AUTO}$	**41.14**	33.34	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**

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.

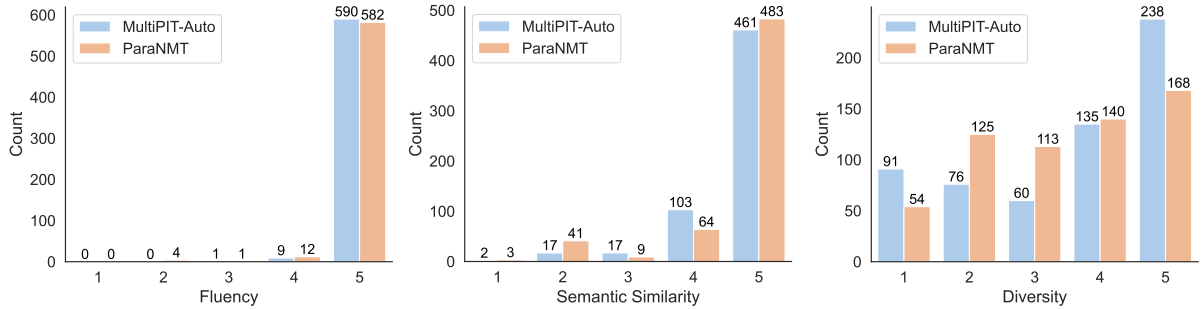

Figure 6: Human evaluation distributions on generations by model fine-tuned on MULTIPIT$_{AUTO}$ or ParaNMT.

Model	Fluency	Semantic Similarity	Diversity
MULTIPIT$_{AUTO}$	**4.98**	**4.67**	**3.59**
ParaNMT	4.95	4.64	3.40

Table 6: Human evaluation results on generations by model fine-tuned on MULTIPIT$_{AUTO}$ or ParaNMT.

performance. On the contrary, since Quora and MSCOCO contain only questions or captions, models fine-tuned on them always generate question- or description-style sentences. For example, given *"we should take shots."*, model fine-tuned on Quora generates *"Why do we take shots?"*.

We conduct a **human evaluation** to further compare MULTIPIT$_{AUTO}$ and ParaNMT datasets, by evaluating 200 randomly sampled generations from the model trained on each corpus.[13] As shown in Table 6, MULTIPIT$_{AUTO}$'s generations receive the highest scores in all three dimensions: *fluency*, *semantic similarity* and *diversity*. Each generation is rated by three annotators on a 5-point Likert-scale per aspect, with 5 being the best. We also show the distribution of human evaluation results on each dimension in Figure 6 for a deeper comparison. Specifically, MULTIPIT$_{AUTO}$ model generates fewer really poor paraphrases (semantic similarity <3) and much more diverse paraphrases (diversity >3). We include our evaluation template in Appendix H. We measure inter-annotator agreement 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 MULTIPIT$_{AUTO}$ 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.

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 sentences 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

[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 often the raters agree.

Type	Definition	Generation Example	M_{AUTO}	ParaNMT
	Good Paraphrase Type	Number of occurrences per generation:	**1.53**	1.22
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:	26%	**44%**
Copy	Copy the given sentence.	**Sent:** Did you have a good day today? **Gen:** Did you have a good day today?	**14.5%**	10%
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 commonsense 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 earphones for rs 3000?	0%	**1%**

Table 7: Paraphrase types with examples and statistics observed in the generations by models fine-tuned on MULTIPIT$_{AUTO}$ (M_{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**).

to which two sentences are semantically similar to each other. Since it doesn't make a binary judgment for paraphrase relationships, it is not frequently used in paraphrase research. Recently, Dong et al. (2021) presents ParaSci, a large paraphrase dataset in the scientific field, and Kim et al. (2021) proposes BiSECT, a large split and rephrase corpus constructed using machine translation. Our work focuses on creating a large paraphrase corpus that contains more diverse and natural human-authored texts and investigating different paraphrase criteria.

6 Conclusion

In this paper, we present the Multi-Topic Paraphrase in Twitter (MULTIPIT) corpus. Our work surpasses prior Twitter-based paraphrase corpora in topic diversity as well as the quality and quantity of annotation. Experimental results demonstrate the necessity of defining paraphrases based on downstream tasks. Our paraphrase generation evaluation shows that models trained on our corpus have better generation quality and generalizability compared to models fine-tuned on existing widely-used paraphrase datasets. We believe that MULTIPIT will facilitate further research in both paraphrase identification and paraphrase generation.

Limitations

While our study shows MULTIPIT$_{AUTO}$ 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,

ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

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.

Tuhin 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 pretraining with gradient-disentangled embedding sharing. *ArXiv*.

Joongwon Kim, Mounica Maddela, Reno Kriz, Wei Xu, and Chris Callison-Burch. 2021. BiSECT: Learning to split and rephrase sentences with bitexts. In *Proceedings of Empirical Methods in Natural Language Processing*.

Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Wuwei Lan, Siyu Qiu, Hua He, and Wei Xu. 2017. A continuously growing dataset of sentential paraphrases. In *Proceedings of Empirical Methods in Natural Language Processing*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of International Conference on Learning Representation*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Association for Computational Linguistics*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Zichao Li, Xin Jiang, Lifeng Shang, and Qun Liu. 2019. Decomposable neural paraphrase generation. In *Proceedings of the Association for Computational Linguistics*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755.

Timothy Liu and De Wen Soh. 2022. Towards better characterization of paraphrases. In *Proceedings of the Association for Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Proceedings of International Conference on Learning Representation*.

Yixin Liu, Graham Neubig, and John Wieting. 2021. On learning text style transfer with direct rewards. In *NAACL*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan T. McDonald. 2020. On faithfulness and factuality in abstractive summarization. *Proceedings of the Association for Computational Linguistics*.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Tong Niu, Semih Yavuz, Yingbo Zhou, Nitish Shirish Keskar, Huan Wang, and Caiming Xiong. 2021. Unsupervised paraphrasing with pretrained language models. In *Proceedings of Empirical Methods in Natural Language Processing*.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. Tweetmotif: Exploratory search and topic summarization for twitter. In *Fourth International AAAI Conference on Weblogs and Social Media*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the Association for Computational Linguistics*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of Empirical Methods in Natural Language Processing*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Vasile Rus, Rajendra Banjade, and Mihai C. Lintean. 2014. On paraphrase identification corpora. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of Empirical Methods in Natural Language Processing*.

Soroush Vosoughi and Deb Roy. 2016. A semi-automatic method for efficient detection of stories on social media. In *Tenth International AAAI Conference on Web and Social Media*.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the Association for Computational Linguistics*.

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$_{\text{CROWD}}$.

A.2 Expert

Figure 11 displays a screenshot of the instruction of our expert annotation for MULTIPIT$_{\text{EXPERT}}$.

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.

Raw Tweets w/o Sent. Splitting
◦ *Horrible Crash on the Aurora Bridge in Seattle.*
◦ *The crash on the Aurora Bridge in Seattle looks horrible. That was the bridge I took to work everyday. Yikes.*

Table 8: An example pair of raw tweets from our corpus. Annotating at tweet-level will include mismatched content and ambiguity. Cleaner paraphrase annotations can be acquired after sentence splitting.

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 $\times$ 48GB NVIDIA A40 GPUs.

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

For T5 model, we consider learning rates $\in$ {1e-4, 3e-4, 1e-5, 3e-5}. For DeBERTaV3 model, we consider learning rates $\in$ {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.

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)

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	500
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:

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:

E Generation Dataset Statistics

Table 11 presents the detailed statistics of MULTIPIT$_{\text{AUTO}}$, Quora, MSCOCO and ParaNMT.

	M$_{\text{AUTO}}$	Quora	MSCOCO	ParaNMT
Genre	Twitter	Question	Description	Novels, Laws
Sentence Length	11.34	9.66	10.49	11.33
Sentence BLEU	24.48	26.37	9.30	24.85
Train/dev/test split				
#Train w/o BF	290,395	134,378	331,330	50M
#Train	136,645	47,393	275,583	443,512
#Dev	215	5,255	20,186	499
#Test	200	5,255	20,187	781
#Test Refs	8	1.34	4	1

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

Metric	Fluency	Semantic Similarity	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 12: Spearman correlations with human evaluation on 100 generations on MULTIPIT$_{\text{NMR}}$ (50 by model trained on MULTIPIT$_{\text{AUTO}}$ 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 Similarity	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***

Table 13: Spearman correlations with human evaluation on all 400 generations. Here, ***: $p < 0.0001$, **: $p < 0.001$, *: $p < 0.01$.

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 MULTIPIT$_{\text{NMR}}$ has the most number of references, to evaluate BLEU, we examine 100 generations on MULTIPIT$_{\text{NMR}}$ (50 by T5$_{\text{large}}$ fine-tuned on MULTIPIT$_{\text{AUTO}}$ and 50 by T5$_{\text{large}}$ fine-tuned on ParaNMT). Results are shown in Table 12. BLEU gets a weak correlation around $|0.2|$ with all as-

Figure 7: Label distribution of 1200 ratings on 400 generations by models fine-tuned on MULTIPIT$_{\text{AUTO}}$ and ParaNMT.

pects and ∼0.1 with the overall score. Table 13 presents Spearman correlations for Self-BLEU, BERT-Score and BERT-iBLEU on all 400 generations. BERT-iBLEU outperforms the other two metrics. Because Self-BLEU measures diversity and BERT-Score measures semantic similarity, both metrics get the best correlation with human evaluation on the corresponding aspect but the worst correlation on the other one. Notably, Self-BLEU gets the highest correlation with the overall measurement, but the reason behind it is more differentiation in diversity ratings compared to semantic similarity, as shown in Figure 7. This makes diversity the biggest role in the overall score.

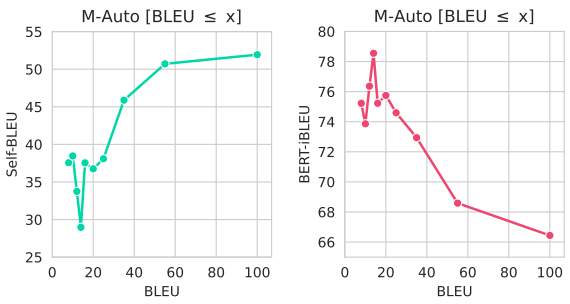

Figure 8: MULTIPIT$_{\text{AUTO}}$ dev set performance on various BLEU filtering thresholds.

BLEU Filtering. We evaluate different BLEU thresholds on the dev set of MULTIPIT$_{\text{AUTO}}$ 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.

Training Data	LR	BL	S-B ↓	B-S	B-iB
M$_{\text{AUTO}}$ w/o BF	1e-4	45.85	65.03	91.86	59.87
M$_{\text{AUTO}}$	1e-4	41.14	33.34	85.86	**77.79**
Quora w/o BF	3e-4	36.63	54.54	91.27	61.58
Quora	1e-4	28.72	34.23	87.97	**73.54**
MSCOCO w/o BF	3e-4	28.15	23.39	82.99	78.89
MSCOCO	1e-4	26.14	15.46	81.00	**80.30**
ParaNMT w/o BF	1e-5	19.43	37.59	88.07	76.28
ParaNMT	3e-5	20.36	33.35	86.90	**77.51**

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.

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

Data	Size	BL	S-B ↓	B-S	B-iB
MULTIPIT$_{\text{CROWD}}$	26,091	36.15	**32.09**	85.53	74.19
M$_{\text{AUTO-CROWD}}$	326,517	**45.55**	37.90	85.80	74.12
M$_{\text{AUTO}}$	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$_{\text{CROWD}}$ contains its paraphrase pairs. M$_{\text{AUTO-CROWD}}$ is the automatically identified paraphrase pairs by the identifier fine-tuned on MULTIPIT$_{\text{CROWD}}$.

G Examples

Generation Examples. Table 16 presents generation examples by GPT-3 and fine-tuned T5$_{\text{large}}$ on MULTIPIT$_{\text{NMR}}$.

Table 17 presents generation examples by T5$_{\text{large}}$ fine-tuned on MULTIPIT$_{\text{AUTO}}$, Quora, MSCOCO, and ParaNMT.

Multi-Reference Examples. Table 18 displays three examples from the MULTIPIT$_{\text{NMR}}$ test set.

H Human Evaluation Details

We display our human evaluation instruction for each aspect (fluency, semantic similarity, diversity) in Figure 12,13,14.

Source sentence: Today is the winter solstice, marking the beginning of winter in the Northern Hemisphere.
Generated sentences:
$T5_{large}$ → Today marks the beginning of winter in the Northern Hemisphere.
$GPT\text{-}3_{zero\text{-}shot}$ → Today is the first day of winter in the Northern Hemisphere.
$GPT\text{-}3_{few\text{-}shot}$ → Today is the winter solstice, which means winter has arrived in the Northern Hemisphere.

Source sentence: Elizabeth Warren and Cory Booker test positive for COVID in breakthrough cases
Generated sentences:
$T5_{large}$ → Cory Booker and Elizabeth Warren have tested positive for COVID-19.
$GPT\text{-}3_{zero\text{-}shot}$ → In a breakthrough, Elizabeth Warren and Cory Booker test positive for COVID.
$GPT\text{-}3_{few\text{-}shot}$ → Elizabeth Warren and Cory Booker have both contracted COVID-19 in what are being called breakthrough cases.

Source sentence: that malcolm subban save is f*cking incredible.
Generated sentences:
$T5_{large}$ → Malcolm Subban with the save of the year.
$GPT\text{-}3_{zero\text{-}shot}$ → That save by Malcolm Subban was insane!
$GPT\text{-}3_{few\text{-}shot}$ → That Malcolm Subban save was absolutely incredible.

Source sentence: Former Oklahoma quarterback Spencer Rattler is transferring to University of South Carolina.
Generated sentences:
$T5_{large}$ → Former Oklahoma QB Spencer Rattler is transferring to the University of South Carolina.
$GPT\text{-}3_{zero\text{-}shot}$ → Former Oklahoma quarterback Spencer Rattler is transferring to the University of South Carolina.
$GPT\text{-}3_{few\text{-}shot}$ → former OU quarterback Spencer Rattler is transferring to the University of South Carolina.

Source sentence: The Vancouver Canucks have fired head coach Travis Green and hired Bruce Boudreau as a replacement.
Generated sentences:
$T5_{large}$ → Canucks fire Travis Green, hire Bruce Boudreau as new head coach
$GPT\text{-}3_{zero\text{-}shot}$ → The Vancouver Canucks have replaced head coach Travis Green with Bruce Boudreau.
$GPT\text{-}3_{few\text{-}shot}$ → The Vancouver Canucks have let go of head coach Travis Green and replaced him with Bruce Boudreau.

Source sentence: What an incredible tribute to Eddie Kidd on Top Gear.
Generated sentences:
$T5_{large}$ → What a great tribute to Eddie Kidd on Top Gear.
$GPT\text{-}3_{zero\text{-}shot}$ → Top Gear did an amazing job honoring Eddie Kidd.
$GPT\text{-}3_{few\text{-}shot}$ → What a fitting tribute to Eddie Kidd on Top Gear.

Source sentence: How to build immunity for Secondhand Stress
Generated sentences:
$T5_{large}$ → How can you build your immunity against secondhand stress?
$GPT\text{-}3_{zero\text{-}shot}$ → There are many ways to build immunity against secondhand stress. Some methods include yoga, meditation, and aromatherapy.
$GPT\text{-}3_{few\text{-}shot}$ → How to become immune to Secondhand Stress

Source sentence: Twitter is trialling its own 24-hour Stories feature called Fleets
Generated sentences:
$T5_{large}$ → @Twitter is testing a new 24-hour stories feature called Fleets.
$GPT\text{-}3_{zero\text{-}shot}$ → Twitter is testing out a new 24-hour Stories feature called Fleets.
$GPT\text{-}3_{few\text{-}shot}$ → Twitter is testing a new 24-hour Stories feature called Fleets.

Table 16: Paraphrase generation examples by GPT-3 and fine-tuned $T5_{large}$ on MULTIPIT$_{NMR}$.

Source origin: MULTIPIT$_{\text{NMR}}$
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:
$T5_{\text{MULTIPIT}_{\text{AUTO}}} \rightarrow$ Today is Veterans Day, a day to remember those who have served.
$T5_{Quora} \rightarrow$ On this Veterans Day, we salute all the men and women who have served our country.
$T5_{MSCOCO} \rightarrow$ A group of men and women standing next to each other.
$T5_{ParaNMT} \rightarrow$ 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:
$T5_{\text{MULTIPIT}_{\text{AUTO}}} \rightarrow$ What is the job outlook for a food technologist in India?
$T5_{Quora} \rightarrow$ What is the scope of food technology in India?
$T5_{MSCOCO} \rightarrow$ a man in a white shirt is working in a kitchen
$T5_{ParaNMT} \rightarrow$ 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:
$T5_{\text{MULTIPIT}_{\text{AUTO}}} \rightarrow$ A bear is taking a walk in the woods.
$T5_{Quora} \rightarrow$ What is a bear doing in the forest?
$T5_{MSCOCO} \rightarrow$ A brown bear walking through a wooded area.
$T5_{ParaNMT} \rightarrow$ the bear is walking through the forest.
Source origin: ParaNMT
Source sentence: someone speaks in english over there.
Generated sentences:
$T5_{\text{MULTIPIT}_{\text{AUTO}}} \rightarrow$ Someone has to speak english in this room.
$T5_{Quora} \rightarrow$ What is the best way to learn English from someone who speaks fluently?
$T5_{MSCOCO} \rightarrow$ A man standing in front of a wall with a sign on it.
$T5_{ParaNMT} \rightarrow$ someone's speaking english.

Table 17: Paraphrase generation examples by T5$_{\text{large}}$ fine-tuned on MULTIPIT$_{\text{AUTO}}$, 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 MULTIPIT$_{\text{NMR}}$.

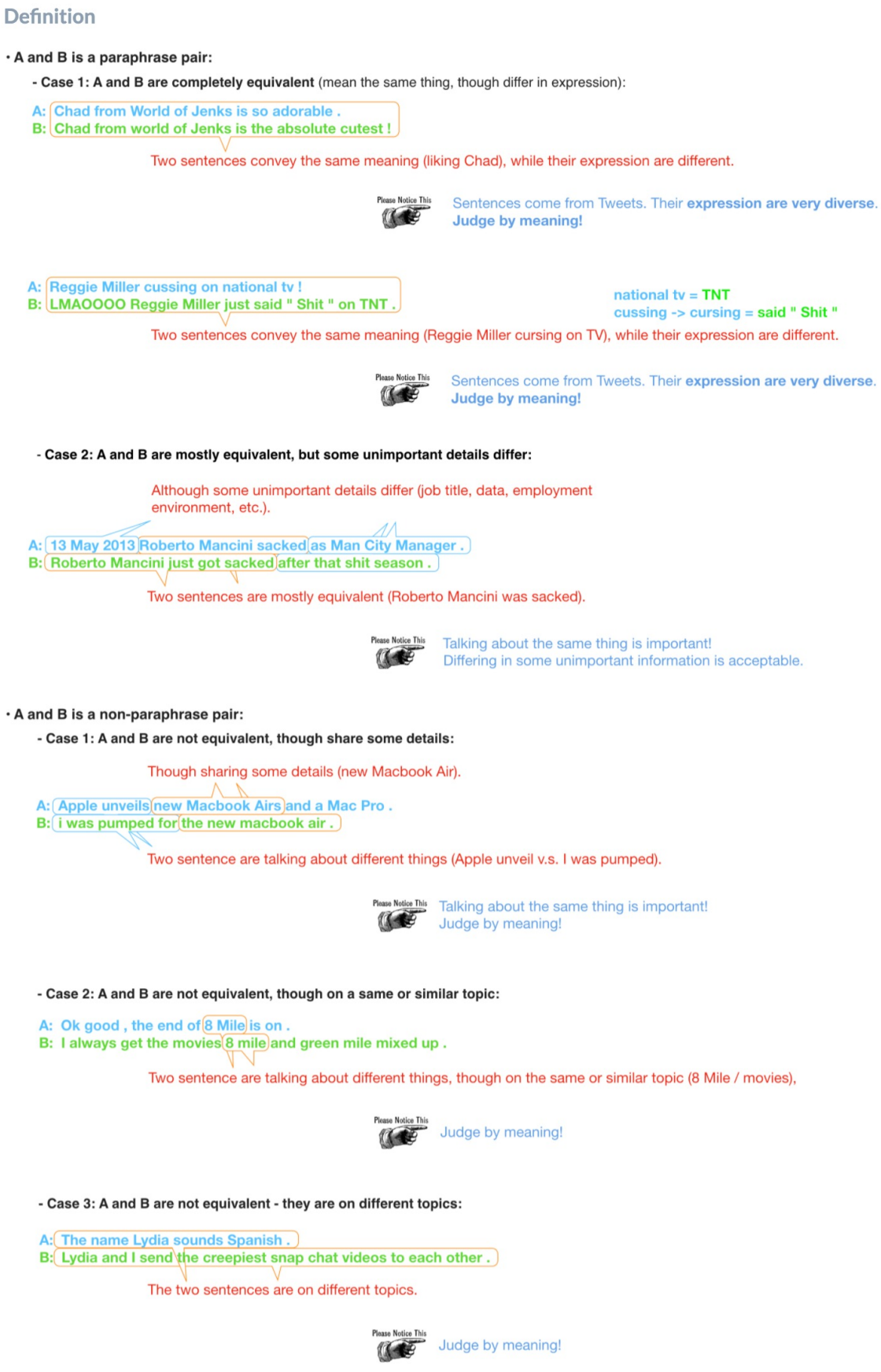

Figure 9: Instruction of our crowdsourcing annotation on the Figure Eight platform for creating MULTIPIT$_{\text{CROWD}}$.

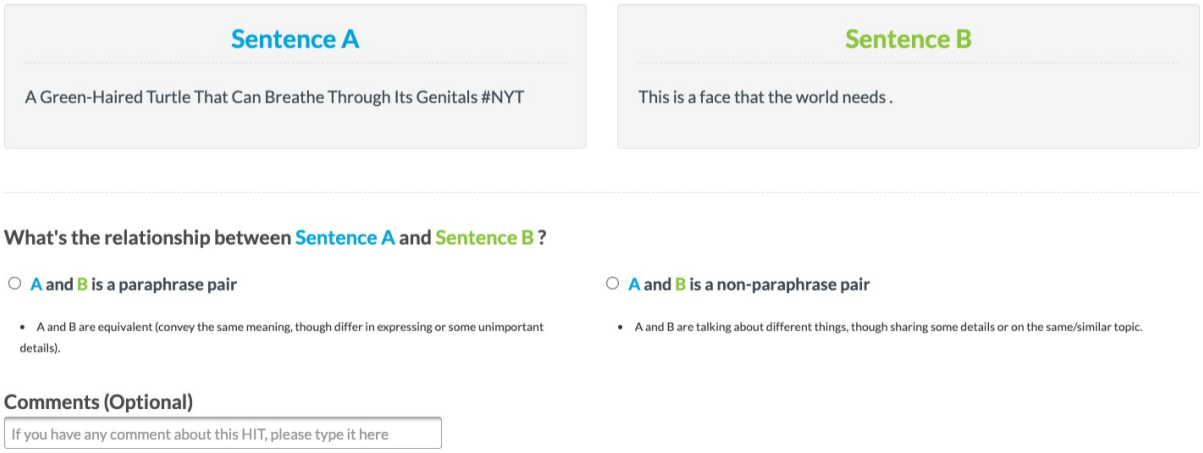

Figure 10: An example question of our crowdsourcing annotation on the Figure Eight platform for creating MULTIPIT$_{\text{CROWD}}$.

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 iteself 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 fack-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 diffrent 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 MULTIPIT$_{\text{EXPERT}}$.

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.*** *Prectising 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 **semanticaly 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 refelect 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.

Note: cases like *"is going to" <-> "will"*, *"wanna" <-> "want to"*, *"gonna" <-> "go to"* are word-wise synonym replacement as well.

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.