ParaCotta: Synthetic Multilingual Paraphrase Corpora from the Most Diverse Translation Sample Pair

Alham Fikri Aji[∗•] **Tirana Noor Fatyanosa**[∗◦] **Radityo Eko Prasojo**[∗⋆] **Philip Arthur**[†]
Suci Fitriany[∗] **Salma Qonitah**[∗] **Nadhifa Zulfa**[∗] **Tomi Santoso**[∗] **Mahendra Data**[‡◦]
[∗] Kata.ai Research Team, [◦] Kumamoto University, [•] University of Edinburgh,
[⋆] Universitas Indonesia, [†] Oracle Digital Assistant, [‡] Brawijaya University
{aji,tirana.fatyanosa,ridho,suci}@kata.ai
{salma.qonitah,dhifa.zulfa,tomi.santoso}@kata.ai
{fatyanosa, mahendra.data}@dbms.cs.kumamoto-u.ac.jp
philip.arthur@oracle.com, mahendra.data@ub.ac.id

Abstract

We release our synthetic parallel paraphrase corpus across 17 languages: Arabic, Catalan, Czech, German, English, Spanish, Estonian, French, Hindi, Indonesian, Italian, Dutch, Romanian, Russian, Swedish, Vietnamese, and Chinese. Our method relies only on monolingual data and a neural machine translation system to generate paraphrases, hence simple to apply. We generate multiple translation samples using beam search and choose the most lexically diverse pair according to their sentence BLEU. We compare our generated corpus with the `ParaBank2`. According to our evaluation, our synthetic paraphrase pairs are semantically similar and lexically diverse.

1 Introduction

Paraphrases are semantically similar sentences or phrases using different expressions (Bhagat and Hovy, 2013). A paraphrase generation system can be developed by training a model, given a dataset of paraphrase parallel texts (Egonmwan and Chali, 2019). Paraphrase parallel corpus is accessible in English (Dolan and Brockett, 2005; Fader et al., 2013; Xu et al., 2015). However, such data for other languages are not as common. (Ganitkevitch and Callison-Burch, 2014) proposed a multilingual paraphrase pairs dataset; however, their corpus is only on phrase-level.

By utilizing a machine translation system, Wieting and Gimpel (2017) proposed synthetic paraphrase corpus by back-translating bilingual text. However, this approach does not consider lexical diversity for the generated paraphrases. Hu et al. (2019a) and Hu et al. (2019b) further improve the method by applying a constraint to generate more diverse paraphrases, then choosing diverse pairs as the synthetic dataset. These methods require bilingual corpus, which might not be easily accessible for certain languages or domains. Our work takes inspiration from selecting the most diverse pair; however, we remove the bilingual text requirement.

We propose a simple way to generate paraphrases by selecting the most diverse pair (in terms of BLEU) from the translation sample. Our approach generates paraphrases from a monolingual text; therefore, not bound to the availability of parallel corpus. We also show that this technique can produce diverse paraphrases, measured in the BLEU score. In addition, we release our generated paraphrase dataset in 17 languages.

2 Generating Paraphrase via Diverse Pairs

We propose a way to construct synthetic paraphrase corpus by utilizing a machine translation system. Our approach involves translating texts from English to the desired language, therefore not limited to the availability of bilingual corpora for back-translation (Hu et al., 2019a). Specifically, given

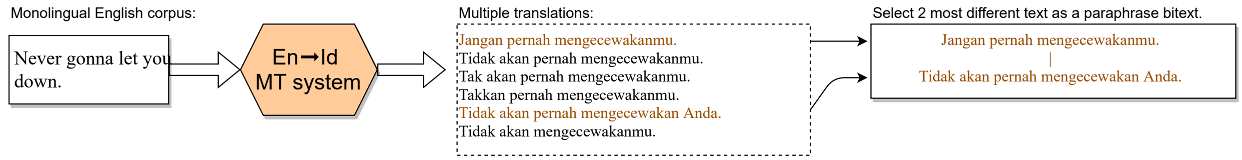


Figure 1: An example of synthetic paraphrase corpus generation using a machine translation system.

Text	Sem.Similarity stsb cosine↑	Lexical Diversity BLEU↓	Lexical Diversity Jaccard↓
He's denied them protection. They're not allowed to do that in a protective shield.	0.630	1.7	0.0
voter representation cannot be guaranteed. It is not possible to guarantee the right to vote.	0.917	2.0	0.0
It is therefore necessary to compensate for business tax failures in the coming years. Therefore, the trade tax losses would have to be compensated in the next few years.	0.796	6.9	0.273
Therefore, unavoidable waiting times may occur. For this reason, there may be inevitable waiting times.	0.866	10.7	0.250
Maintenance-free batteries are supposed to prevent this from happening. Maintenance-free batteries should actually prevent that.	0.921	16.9	0.308
Do taxes need to be raised to finance the stimulus package? Do taxes have to be increased to finance the economic stimulus package?	0.949	21.0	0.615
Small successes for the first time with tested Corona vaccine (9.45 o'clock) Small successes with tested Corona vaccine (9.45 am)	0.959	38.6	0.533
Everything is now clear for the construction of a new ice channel at Barenberg. Now everything is clear for the start of construction of a new ice channel on Barenberg.	0.994	43.6	0.812

Table 1: Synthetic paraphrase corpus example (English)

an input text X, we produce several translation samples $Y_0, ..., Y_N$ with beam-search. Then, we chose two sentences Y_i and Y_j as a paraphrase pair, such that both sentences are the most lexically diverse among other choices. Here, we define the lexical diversity with a BLEU score, where a lower BLEU score denotes a more diverse pair. For more details, see Figure 1.

To produce the synthetic paraphrase corpus for a language L, we use an English to L translation system, as well as monolingual English corpus. It is possible to use a pivot language other than English. However, we argue that it is more difficult to achieve due to the availability of the translation system.

With this method, we generate synthetic paraphrase corpus across 17 languages. For non-English corpus, we translate monolingual English to the desired language. Our English monolingual corpus is sampled from ParaBank2 (Hu et al., 2019b) (3M sent), Wikipedia (1M sent), NewsCrawl (1M sent), and English Tatoeba (1M sent). For the English paraphrase corpus, we translate monolingual German text collected from NewsCrawl (2.5M sent) and German Tatoeba (500k). We are planning to support more languages and use more monolingual data as future work. Examples of English-generated data can be seen in Table 1, alongside their qualitative evaluations, which will be explained in Section 4.

3 Model Configuration

We use a Transformer-based encoder-decoder architecture (Vaswani et al., 2017) for both our translation system and paraphrase generator system. For both systems, we use the same Transformer-base architecture which consists of 6 layers of encoder and decoder, and an embedding size of 512. The input is tokenized with sentence-piece (Kudo and Richardson, 2018).

We rely on NMT system to produce our synthetic paraphrase data. For most of our translation system, we use pre-existing public model available in Huggingface.[1] These MT systems are trained on OPUS parallel corpus. Without losing the generality, we re-train Indonesian MT with the additional dataset from Guntara et al. (2020).

Similarly, we use the same architecture to train our paraphrase generation model. We train our paraphrase system for 10 epochs with Adam optimizer. We use Marian toolkit (Junczys-Dowmunt et al., 2018) to train our model.

4 Evaluation and Analysis

4.1 Evaluation Method

Our objective is to maximize both the lexical diversity and semantic similarity of our paraphrase pairs. The lexical diversity is, by the design of our approach, guarded by the BLEU score. Indeed, using the BLEU score to determine paraphrase originality or diversity has been used in prior work (Mallinson et al., 2017; Hu et al., 2019b; Hu et al., 2019a). However, in our case, we use BLEU exclusively to measure lexical diversity. On top of BLEU, we further evaluate our paraphrase quality using word-level Jaccard Index (Jaccard, 1912) for lexical diversity. For semantic similarity, we rely on using manual evaluation and and sBERT score (Reimers and Gurevych, 2020).

We average the BLEU score for both directions since the reference paraphrase is not defined. Following (Hu et al., 2019b), we also compute the BLEU on lowercased text after stripping the punctuation. We use sacreBLEU (Post, 2018) for calculation. Similarly, we compute the Jaccard index on lowercased and de-punctuated text.

For automatic semantic similarity evaluation, we leverage distilled multilingual sBERT models (Reimers and Gurevych, 2020), in particular the `paraphrase-xlm-r-multilingual-v1` and `stsb-xlm-r-multilingual` models trained on 50+ languages, which scored a high Pearson's ρ on semantic textual similarity (STS) tasks despite being relatively lightweight.

For manual evaluation, we randomly select 100 sentences from the generated corpus. Then, professional annotators[2] are asked to score each of the paraphrase pairs on a 3-point Likert scale system: (1) Inequivalent or unrelated; (2) Roughly equivalent; (3) Completely or mostly equivalent. The scores are then averaged and scaled to 0-100. A more detailed guideline can be found in Appendix A.

[1] https://huggingface.co/Helsinki-NLP

	Semantic Similarity		Lexical Diversity	
Language	stsb↑	para↑	BLEU↓	Jaccard↓
Arabic (ar)	0.926	0.925	25.6	0.357
Catalan (ca)	0.909	0.901	34.3	0.435
Czech (cs)	0.913	0.923	24.7	0.376
German (de)	0.934	0.925	28.0	0.427
English (en)	0.909	0.876	34.6	0.523
Spanish (es)	0.942	0.932	34.0	0.452
Estonian (et)	0.892	0.911	23.2	0.377
French (fr)	0.924	0.914	33.3	0.425
Hindi (hi)	0.894	0.897	39.5	0.604
Indonesian (id)	0.936	0.929	28.1	0.426
Italian (it)	0.931	0.920	31.6	0.421
Dutch (nl)	0.921	0.912	30.4	0.456
Romanian (ro)	0.933	0.927	26.9	0.376
Russian (ru)	0.930	0.921	26.9	0.376
Swedish (sv)	0.916	0.906	29.2	0.428
Vietnamese (vi)	0.933	0.904	40.2	0.517
Chinese (zh)	0.879	0.877	37.8	0.470

Table 2: Corpus statistic across languages. stsb and para are the cosine distance of the embeddings generated by sBERT `stsb` and `paraphrase` models, respectively.

4.2 Synthetic Corpus Evaluation

The data statistic across 17 languages, sampled from 10k sentences per language, can be seen in Table 2. We find that the scores on both models are extremely similar; therefore, we only used the `stsb` model for our later evaluations.

Table 4 shows some examples on our proposed dataset in other languages besides English.

To further analyze our generated dataset, we perform human evaluation on selected languages of English and Indonesian. We also evaluate English ParaBank2 as a comparison. We manually annotate 100 samples from our dataset per language. For ParaBank2, we manually annotate 50 samples. We

[2] This is to control the annotation quality better and to avoid navigating through the ethical concerns of using a crowd-platform (Shmueli et al., 2021). However, this limits our manual evaluation to only two languages in which our annotators are professionally fluent.

Dataset	Semantic Similarity		Lexical Diversity		Semantic Similarity		Lexical Diversity	
	Manual↑	Cosine↑	BLEU↓	Jaccard↓	Manual↑	Cosine↑	BLEU↓	Jaccard↓
	English dataset				**Indonesian dataset**			
ParaBank2 (Hu et al., 2019b)	88.5	0.812	**23.9**	**0.388**	n/a			
Ours (no filter)	95.0	0.876	34.6	0.523	92.5	0.936	28.1	0.426
Ours (BLEU filter 0-80)	95.0	0.909	34.1	0.522	92.3	0.936	28.1	0.426
Ours (BLEU filter 0-60)	94.8	0.908	31.7	0.512	91.2	0.935	**26.4**	**0.420**
Ours (BLEU filter 20-80)	**97.2**	**0.926**	41.7	0.594	**96.6**	**0.953**	40.5	0.566
Ours (BLEU filter 20-60)	97.0	0.924	39.2	0.585	95.2	0.952	38.5	0.573

Table 3: Corpus statistic with human evaluation.

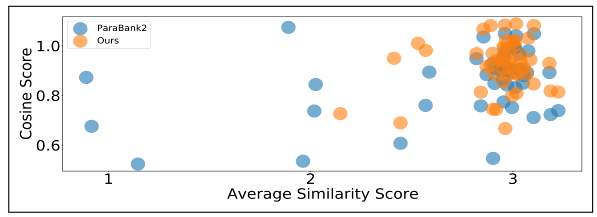
achieve an annotator agreement of 0.5 (weighted kappa) which indicates fair agreement.

As shown in Table 3, our proposed technique is able to generate semantically similar paraphrases. Compared to ParaBank2, we achieve a better semantic similarity score. Unfortunately, our dataset is less lexically diverse. However, our approach uses a monolingual corpus to produce the paraphrase data. Therefore, our approach does not depend on the availability of parallel corpus, which is beneficial for low-resource languages. Note that our approach still requires parallel corpus to build the MT system, although an alternatively zero-shot MT system can be used. Similarly, MT system can be build under low-resource setting with the help of pre-trained language models. In these cases, our paraphrase generation mechanism can be used regardless the availability of the parallel corpus. However, we leave this as future work.

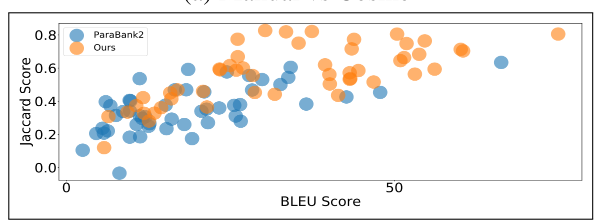
Metric Correlation

To test the relationship between all metrics for all models, we calculated the Spearman correlation with $\alpha = 0.05$ as shown in Table 5. BLEU and Jaccard correlate well to measure lexical diversity, with a 0.803 Spearman coefficient. Similarly, human evaluated semantic similarity correlates with sBERT cosine similarity with 0.304 Spearman coefficient.

In the scatter plots (Figure 2), we visualize the comparison between ParaBank2 and our approach for the English dataset. We can see that all the metrics have a higher density on high-scoring sentences as most of the sentences are given a score of 3 by human annotators. Overall, our proposed approach is not as diverse as ParaBank2 but generally has a higher human annotation value than ParaBank2.



(a) Manual vs Cosine



(b) BLEU vs Jaccard

Figure 2: Scatter plots comparison between ParaBank2 and Ours for metrics correlation. Random gaussian noise $\mathcal{N}(0,0.1)$ and $\mathcal{N}(0,0.05)$ have been added to x-axis and y-axis, respectively.

BLEU-filtering

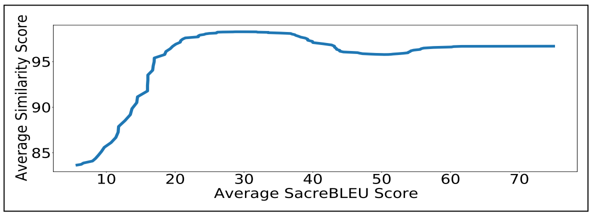
Upon further investigation, we notice a correlation between the BLEU and human-annotated semantic similarity. As shown in Figure 3, less creative paraphrases tend to be more semantically similar. In contrast, more diverse paraphrases are less semantically similar. Based on this observation, we also attempt to filter our generated synthetic pairs based on BLEU. Specifically, high-BLEU paraphrases can be removed to avoid 'lazy and boring' paraphrases, resulting in more diverse datasets. Orthogonally, low-BLEU paraphrases can be removed as they are not as semantically similar. Filtering our corpus can further adjust the overall lexical diversity and semantic similarity, as shown in Table 3.

Lang	Text 1	Text 2
ar	بعدم تحدي (تشارلز ستيوارت) و عدم سؤاله عن أي شيء	بعدم تحدي (تشارلز ستيوارت) أبداً وعدم طلبه أي شيء
	ويجب أن أقنع البابا بأن يبطل زواجنا	وعليّ إقناع البابا بإلغاء زواجنا
	لكن بسبب نزاع في العمل ، عُلّق العمل في سنة ١٩٢٤ .	ولكن بسبب صراع في العمل ، عُلق العمل سنة ١٩٢٤ .
ca	Això demostra que el preu d'exportació a països tercers era substancialment menor.	Això demostra que els preus de l'exportació als països tercers eren substancialment més baixos.
	Mai abans havia estat a prop d'un.	Mai havia estat aprop d'un abans.
	Fem-ho el millor aniversari que hagis tingut.	Fem això el millor aniversari que has tingut mai.
cs	Váš výbor připravuje všestranný návrh zákona.	Vaše komise připravuje zákon o všem.
	Kate, vévodkyně Cambridge, zůstala doma během tohoto prvního setkání.	Kate, vévodkyně z Cambridge, zůstala během prvního setkání doma.
	Oh, dobře, tohle už jsi dělala.	Fajn, už jsi to někdy dělal.
de	Garner führte das Team in Eile und akkumulierte 72 Empfänge.	Garner führte die Mannschaft in Eile und sammelte 72 Empfänge.
	Du hättest nicht so habgierig sein sollen.	Du solltest nicht so gierig sein.
	Später in diesem Jahr gab er sein Debüt bei den Hong Kong Sevens.	Im selben Jahr debütierte er in den Hong Kong Sevens.
es	No encuentro nada que explique tu dolor de cabeza.	No puedo encontrar nada para explicar ese dolor de cabeza tuyo.
	Galileo ofrecerá modelos genéricos para los elementos locales.	Galileo proporcionará modelos genéricos de elementos locales.
	Gracias a su experiencia, tenía ventaja sobre el resto.	Gracias a su experiencia, tuvo una ventaja sobre los demás.
et	Masinad, mis on sama suured kui hooned.	Majasuurused masinad.
	Jah, teise iseseisvusreferendumi korraldamisel on palju emotsioone.	Jah, kui toimub teine iseseisvusreferendum, tekib palju tundeid.
	Ma hüppasin üle logi ja peaaegu kukkus, kuid püütud ise ja jätkas jooksmist.	Hüppasin üle palgi ja peaaegu kukkusin, aga jäin vahele ja põgenesin edasi.
fr	En dehors des trois premiers, les autres luttes.	En dehors de ces trois premières, le reste se bat.
	Donc, tu vois, c'est de ma faute si Hopper revient.	C'était ma faute si Hopper revenait.
	Aujourd'hui, un développement rapide dans la province modifie les modes de vie traditionnels.	Aujourd'hui, le développement rapide de la province modifie le mode de vie traditionnel.
hi	ओलीफैंट ने इस विकास की कल्पना नहीं की थी :	ओलिपहान्ट ने इस विकास का विचार नहीं किया था :
	मैरी अपने आप से घर पहुंच सकती है।	मैरी खुद से घर मिल सकती है।
	एक शो, \$20,000 पुरस्कार, हमने इसे 60/40 में विभाजित किया।	एक प्रदर्शन, \$20,000 पुरस्कार, हम इसे 60/40 विभाजित.
id	Apakah ini berarti kita dapat mengandalkan Anda untuk terakhir kalinya?	Apa ini artinya kami bisa mengandalkanmu untuk terakhir kalinya?
	Mereka tidak ingin berkelahi, tetapi mereka harus.	Mereka tidak ingin bertarung, tapi mereka harus melakukannya.
	Kurt Busch melakukan hal yang sama karena dia mengubah mesin mobilnya.	Kurt Busch juga melakukan hal yang sama karena ia telah mengganti mesin mobilnya.
it	Abbiamo passato tutta la notte inginocchiandoci.	Abbiamo trascorso l'intera notte in ginocchio.
	Il Re delle Scimmie ha distrutto ogni soldato mandato a fermarlo.	Il Re Scimmia ha schiacciato ogni soldato inviato per fermarlo.
	Certo, questa rivelazione era sicura di porre fine a questo sforzo immediatamente.	Naturalmente, questa rivelazione era certa di porre fine immediatamente a questo sforzo.
nl	We zouden meer tijd met elkaar hebben.	Dan hadden we meer tijd samen.
	We namen deel aan de amateur boksbond Berlijn om ons te helpen bij het vinden van nieuwe hoop.	We hebben deelgenomen aan de amateurboksbond Berlijn om ons te helpen nieuwe hoop te vinden.
	Zweer bij God, Lewis, als je die fles aanraakt... trek ik je vingers eraf.	Als je die fles aanraakt, ruk ik je vingers eraf.
ro	corpul apt menţinut în indolenţă senzuală lentă;	corpul capabil să fie menținut în indolență senzuală lent;
	Este la fel ca un vaccin normal.	E ca un vaccin obişnuit.
	Apoi, în genunchi ea cade, plânge, suspină, bate inima ei, lacrimile părul ei, se roagă, blesteme,	Apoi, în genunchi ea cade, plânge, suspină, bate inima ei, lacrimi parul ei, se roaga, blesteme,
ru	Тебе всегда нужно говорить об убийстве людей?	Ты всегда говоришь об убийствах людей?
	Биограф Тэтчера Джон Кэмпбелл заявил, что «этот отчет был частью журналистской ошибки».	Биограф Тэтчера Джон Кэмпбелл утверждал, что "отчёт был частью журналистского грязного дела".
	А вы, вы - комиссар партии, ответственный за моральное состояние экипажа.	А ты, ты - комиссар партии, ответственный за моральный дух команды.
sv	Sen hoppade de på oss på Two Mile Pass.	Då skulle de hoppa över oss vid Two Mile Pass.
	I skolan får pojkarna veta att Terrance har klonat en mänsklig fot.	I skolan lär pojkarna sig att Terrance klonat en människofot.
	Hon kommer att vara där nu.	Hon är där vid det här laget.
vi	Tôi đã từng được gọi là một tay chơi.	Tôi bị gọi là dân chơi.
	Có những giới hạn trong việc nói năng tự do không?	Có giới hạn về việc tự do ngôn luận không?
	(Đó là những gì bạn luôn luôn cho chúng tôi biết dù sao đi nữa.)	(Đó là điều bạn luôn luôn nói với chúng tôi.)
zh	他亲口告诉我的	他自己也跟我说过
	每当你寄支票, 好吧,你必须去"Ta -Ta, girl."	每当你寄支票,好吧, 你得去"塔塔,女孩。"
	可能不想让我给你看我们将要看到的这个片段	可能不想让我给你们看这段我们即将看到的片段

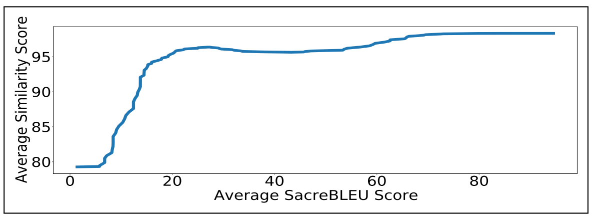
Table 4: ParaCotta example on other languages.

Table 5: Spearman correlation between all scores for all models (Indonesian dataset).

	Manual	Cosine	BLEU	Jaccard
Manual	1	0.304	0.210	0.233
Cosine	0.304	1	0.465	0.605
BLEU	0.210	0.465	1	0.803
Jaccard	0.233	0.605	0.803	1



(a) English dataset



(b) Indonesian dataset

Figure 3: Comparison of BLEU score and manual semantic similarity score.

4.3 Paraphrase System Evaluation

Model	Semantic Similarity		Diversity
	Manual↑	Cosine↑	BLEU↓
Round-trip MT	78.1	86.8	44.7
Ours (no filter)	83.5	86.2	**32.8**
Ours (BLEU filter 20-60)	88.1	91.2	47.0
Ours (BLEU filter 20-80)	**88.9**	**92.1**	48.0

Table 6: Automatic and manual evaluation on 100 sentences across models for single reference (Indonesian dataset).

In this section, we build our paraphrase system by training a Transformer seq2seq model (Vaswani et al., 2017) with our proposed synthetic paraphrase corpus. As another comparison, we also implement round-trip MT, where we translate the input to a pivot language and then translate it back to input language (Mallinson et al., 2017).

Table 6 shows the overall results for lexical diversity and semantic similarity. From the result, Round-trip MT achieves the least semantically similar paraphrases. We argue that since round-trip MT executes the translation two times for both directions, it is more prone to a translation error.

Our proposed scenario achieved lexically diverse paraphrases while maintaining a better semantic similarity than the round-trip translation. Alternatively, filtering the BLEU in our synthetic dataset yields to more semantically similar paraphrase but also sacrifice lexical diversity.

5 Related Work

Prior work has shown that paraphrase can be used to provide additional data (Ma, 2019), which proves to increase model performance, for example in machine translation (Seraj et al., 2015; Marton, 2013), question answering (Dong et al., 2017), relation extraction (Zhang et al., 2015), or text generation (Gao et al., 2020). Additionally, paraphrase has been used to aid NLP evaluation (Thompson and Post, 2020).

There are several well-known English paraphrase corpus, such as ParaBank2 (Hu et al., 2019b), PPDB (Pavlick et al., 2015), Microsoft Research Paraphrase Corpus (MSRP) (Dolan and Brockett, 2005), Microsoft Research Video Description Corpus (Chen and Dolan, 2011), Paralex (Fader et al., 2013), and Paraphrase and Semantic Similarity in Twitter (PIT) (Xu et al., 2015). This paper compares our proposed approach with ParaBank2, which is the synthetically generated and the larger-scale corpus.

6 Conclusion

We proposed a way to generate a synthetic paraphrase corpus by utilizing a monolingual corpus and a translation system. The paraphrase pair is obtained by generating multiple translation samples from an English text and then pick the most diverse pair, denoted with the smallest BLEU score. With this approach, we produce a paraphrase corpus for 17 languages which we release publicly.[3] Our paraphrase is semantically similar, according to human evaluation and sBERT cosine distance evaluation. Nevertheless, our corpus is lexically diverse according to BLEU and Jaccard index.

[3]https://github.com/afaji/paracotta-paraphrase

As future work, it would be interesting to explore a different way to produce translation samples besides the beam search. Adjusting the sample size is another direction to explore, as a higher sample means that we have much more choices, therefore potentially more lexically diverse paraphrases. However, semantic similarity, as well as the computational cost required for higher sample size, must be considered. We also plan to test our approach using different NMT systems and investigate the usefulness of our dataset for downstream NLP tasks. Finally, we left for future work the details and suggestions to consider the trade-off between semantic similarity and lexical diversity.

References

Rahul Bhagat and Eduard Hovy. 2013. What Is a Paraphrase? *Computational Linguistics*, 39(3):463–472, sep.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200, Portland, Oregon, USA, June. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, pages 9–16.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to paraphrase for question answering. *arXiv*, pages 875–886.

Elozino Egonmwan and Yllias Chali. 2019. Transformer and seq2seq model for paraphrase generation. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 249–255, Hong Kong, November. Association for Computational Linguistics.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-driven learning for open question answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1608–1618, Sofia, Bulgaria, August. Association for Computational Linguistics.

Juri Ganitkevitch and Chris Callison-Burch. 2014. The multilingual paraphrase database. In *LREC*, pages 4276–4283. Citeseer.

Silin Gao, Yichi Zhang, Zhijian Ou, and Zhou Yu. 2020. Paraphrase augmented task-oriented dialog generation. *arXiv preprint arXiv:2004.07462*.

Tri Wahyu Guntara, Alham Fikri Aji, and Radityo Eko Prasojo. 2020. Benchmarking multidomain english-indonesian machine translation. In *Proceedings of the 13th Workshop on Building and Using Comparable Corpora*, pages 35–43.

J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin van Durme. 2019a. PARABANK: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of AAAI 2019*, Hawaii; USA.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019b. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. *CoNLL 2019 - 23rd Conference on Computational Natural Language Learning, Proceedings of the Conference*, pages 44–54.

Paul Jaccard. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing revisited with neural machine translation. *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 - Proceedings of Conference*, 1(2003):881–893.

Yuval Marton. 2013. Distributional phrasal paraphrase generation for statistical machine translation. *ACM Transactions on Intelligent Systems and Technology*, 4(3).

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevich, Benjamin Van Durme, and Chris Callison-Burch. 2015. Ppdb 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Association for Computational Linguistics*, Beijing, China, July. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Ramtin Mehdizadeh Seraj, Maryam Siahbani, and Anoop Sarkar. 2015. Improving statistical machine translation with a multilingual Paraphrase Database. *Conference Proceedings - EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, (September):1379–1390.

Boaz Shmueli, Jan Fell, Soumya Ray, and Lun-Wei Ku. 2021. Beyond fair pay: Ethical implications of nlp crowdsourcing. *arXiv preprint arXiv:2104.10097*.

Brian Thompson and Matt Post. 2020. Automatic machine translation evaluation in many languages via zero-shot paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 90–121.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

John Wieting and Kevin Gimpel. 2017. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.

Wei Xu, Chris Callison-Burch, and Bill Dolan. 2015. SemEval-2015 task 1: Paraphrase and semantic similarity in Twitter (PIT). In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 1–11, Denver, Colorado, June. Association for Computational Linguistics.

Congle Zhang, Stephen Soderland, and Daniel S. Weld. 2015. Exploiting Parallel News Streams for Unsupervised Event Extraction. *Transactions of the Association for Computational Linguistics*, 3:117–129.

A Manual Evaluation Guideline

This guideline describes detailed information regarding the concept of annotation for the paraphrasing task. The paraphrase pair should be ranked by its similarity: how much the two sentences are similar semantically. The scores are defined in a 3-point-system. In this guideline, we will show some examples of each score.

A.1 Score 3 - Completely or mostly equivalent

A.1.1 Equivalent meaning/Synonym

The two sentences practically mean the same thing.

Text 1: *The next morning he was found unconscious.*

Text 2: *The next morning he was found passed out.*

Text 1: *I eat rice.*

Text 2: *Rice is eaten by me.*

Text 1: *The head of the local disaster unit, Gyorgy Heizler, said the bus driver failed to notice the red light.*

Text 2: *The bus driver failed to notice the red light, said Gyorgy Heizler, a head of the local disaster unit.*

A.1.2 Identical

Note that the exactly same sentence should be scored 3. We only care about semantic similarity. Creativity will be measured with a different scoring system.

Text 1: *I eat rice*

Text 2: *I eat rice*

A.1.3 Equivalent meaning but using informal form

Different language/style, but equivalent meaning It is acceptable even if the text is not in formal form.

Text 1: *I am delighted.*

Text 2: *I am chuffed.*

A.1.4 Generalization

Subjects and predicates in the main clause are still equivalent or related, the case of pronoun output without additional context is considered generalization.

Text 1: *Uncle has bought a car.*

Text 2: *Uncle has bought a vehicle.*

Text 1: *Jokowi is making a speech.*

Text 2: *He is making a speech.*

Text 1: *Gave a speech nine days ago in St. Petersburg.*

Text 2: *He gave a speech nine days ago in St Petersburg.*

A.1.5 Mostly Similar meaning, but there is additional/missing minor details

Text 1: *The US market is expected to fall 2.1 percent this year.*

Text 2: *The American market is expected to fall 2.1 percent.*

A.1.6 Mostly Similar meaning, but differs in minor details

Text 1: *The US market is expected to fall 2.1 percent this year*

Text 2: *The American market is set to fall 2.1 percent this year.*

A.2 Score 2 - Roughly equivalent

An annotation score of 2 is associated with the medium similarity paraphrases: not identical but similar.

A.2.1 Identical/mostly similar but repeated

The two sentences were almost identical or very similar, but the output model is repeated.

Text 1: *This time it was different, this time it was better.*

Text 2: *This time it was different, this time it was better. This time it was different, this time it was better..*

A.2.2 Roughly similar meaning, but there is additional/missing important information

Text 1: *Richman was irritated by Burne's tone.*

Text 2: *Richman was irritated by Burne's tone. He felt uncomfortable with Burne's attitude.*

A.2.3 Roughly similar meaning, but they differ in important details

Text 1: *How long has it been since I last paid you, Clifton?*

Text 2: *How long have I paid you, Clifton?*

A.3 Score 1 - Inequivalent or unrelated

A.3.1 Not equivalent, but same topics

Text 1: *The Nasdaq composite index rose 10.73 or 0.7 percent to 1,514.77.*

Text 2: *The Nasdaq Composite Index, which is filled with technology stocks, has recently gained about 18 points.*

A.3.2 Very dissimilar and unrelated

Text 1: *I'm eating rice.*

Text 2: *She fell asleep.*