OSPT: European Portuguese Paraphrastic Dataset with Machine Translation

Afonso Sousa[1,2](✉) and Henrique Lopes Cardoso[1,2]

[1] Faculdade de Engenharia, Universidade do Porto, Porto, Portugal
ammlss@fe.up.pt, hlc@fe.up.pt

[2] Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC), Porto, Portugal

Abstract. We describe OSPT, a new linguistic resource for European Portuguese that comprises more than 1.5 million Portuguese-Portuguese sentential paraphrase pairs. We generated the pairs automatically by using neural machine translation to translate the non-Portuguese side of a large parallel corpus. We hope this new corpus can be a valuable resource for paraphrase generation and provide a rich semantic knowledge source to improve downstream natural language understanding tasks. To show the quality and utility of such a dataset, we use it to train paraphrastic sentence embeddings and evaluate them in the ASSIN2 semantic textual similarity (STS) competition. We found that semantic embeddings trained on a small subset of OSPT can produce better semantic embeddings than the ones trained in the finely curated ASSIN2's training data. Additionally, we show OSPT can be used for paraphrase generation with the potential to produce good data augmentation systems that *pseudo*-translate from Brazilian Portuguese to European Portuguese.

Keywords: Paraphrastic dataset · Semantic embeddings · Paraphrase generation · European portuguese

1 Introduction

Paraphrase generation[1] transforms a natural language text into a new text with the same semantic meaning but a different syntactic or lexical surface form [7]. This is a challenging problem commonly approached using supervised learning [2,17].

While this task has been extensively explored for English, few works have been developed for other languages, namely Portuguese. We are aware of one work exploring paraphrase generation for (Brazilian) Portuguese [28]. There is no existing work targeting paraphrase generation for European Portuguese, and

[1] The code and data are available at https://github.com/afonso-sousa/pt_para_gen.

N. Moniz et al. (Eds.): EPIA 2023, LNAI 14115, pp. 454–466, 2023.
https://doi.org/10.1007/978-3-031-49008-8_36

only two small phrasal datasets of aligned paraphrases are available [4,5], which are not publicly accessible. For English, however, approaches have been developed for generating freely-available datasets with millions of sentential paraphrase pairs [12,37].

In this paper, we describe the creation of a dataset containing more than 1.5 million sentential paraphrase pairs. We use neural machine translation (NMT) to translate the English side of a large English-Portuguese parallel corpus, namely OpenSubtitles [23]. We pair the Portuguese translations with the European Portuguese references to form paraphrase pairs. We call this dataset OSPT, as an abbreviation of OpenSubtitles for Portuguese. This dataset covers a broad range of paraphrase phenomena (we cover this analysis in more detail in Sect. 3).

We show the utility of the dataset by using it to train paraphrastic sentence embeddings. We primarily evaluate our sentence embeddings on the ASSIN2 [29] semantic textual similarity (STS) competition. Despite being built for Brazilian Portuguese, for a lack of a better alternative, we deem this competition a good option to evaluate the quality of our data intrinsically. We compare sentence embeddings trained on the official training set from the competition against sentence embeddings trained with a small subset of OSPT. We found the embeddings trained with our dataset outperform those trained from a curated training split.

Lastly, we show that our dataset can be used in paraphrase generation. Having the European Portuguese sentences as targets in fine-tuning a multilingual pre-trained language model produces a *pseudo*-translation effect. The generations are much more European Portuguese-like than the sources, which exhibited Brazilian-like features.

We release our dataset, trained sentence embeddings, paraphrase generators, and all the code to do so. As far as we know, OSPT is the most extensive collection of Portuguese sentential paraphrases released to date. We hope it can motivate new research directions in Portuguese and be used to create powerful Natural Language Processing models while adding robustness to existing ones by incorporating paraphrastic knowledge.

2 Related Work

We discuss work in automatically building paraphrase corpora using parallel text for learning sentence embeddings and similarity functions, and paraphrase generation in Portuguese.

Paraphrase Discovery and Generation
Many methods have been developed for generating or finding paraphrases, including using multiple translations of the same source material [6], using comparable articles from multiple news sources [10], crowdsourcing [18], using diverse machine translation systems to translate a single source sentence [34], and using tweets with matching URLs [21].

Besides all these techniques, the most influential prior work uses bilingual corpora. Bannard, and Callison-Burch [3] used methods from statistical machine translation to find lexical and phrasal paraphrases in parallel text. Ganitkevitch et al. [13] scaled up these techniques to produce the Paraphrase Database (PPDB), which has then been extended for many languages [12] since it only needs parallel text. Wieting et al. [38] used NMT to translate the non-English side of sentential parallel texts to get English-English paraphrase pairs and claimed their data quality to be on par with manually-written English paraphrase pairs. The same authors then scale up the method to produce a larger dataset [37]. We intend to do the same but produce Portuguese-Portuguese paraphrase pairs.

Sentence Embeddings
As in Wieting and Gimbel's work [37,38], we train sentence embeddings to demonstrate the quality of the dataset. These works trained models on noisy paraphrase pairs and evaluated them primarily on semantic textual similarity (STS) tasks. Prior work in learning general sentence embeddings has used autoencoders [16], encoder-decoder architectures [11], and other learning frameworks [1,9,27]. More recently, there are approaches leveraging the embeddings of pretrained language models, like SimCSE [14] or Sentence-BERT (SBERT) [30]. We use the latter for our STS task.

Parallel Text for Learning Embeddings
Prior work has shown that parallel text, and resources built from parallel text like NMT systems and PPDB, can be used to learn word and sentence embeddings. Some works have used PPDB as a knowledge resource for training or improving embeddings [26,36]. Others have used NMT architectures and training settings to obtain better embeddings, like Mallinson et al. [25] that adapted trained NMT models to produce sentence similarity scores in semantic evaluations, or Wieting and Gimpel [37] that proposed mega-batches to expand the search space for selecting negative examples for each paraphrase pair to then compute a margin triple loss [30]. In this work, we opt to use a multiple negative loss [15] because we do not have negative examples. This loss assumes that every other target sentence (aside from the target sentence from the pair being evaluated) in the batch is a negative example.

3 The Dataset

To create our dataset, we used back-translation [38]. We used an English-Portuguese NMT system to translate English sentences from the training data into Portuguese. We paired the translations with the European Portuguese references to form Portuguese-Portuguese paraphrase pairs (i.e., $\langle MixedPortuguese, EuropeanPortuguese \rangle$ pairs).

Throughout the document, we refer to Portuguese as a mixture of European and Brazilian Portuguese, as most pre-trained multilingual models do not distinguish between the two variants. To refer to a specific variant, we explicitly say so.

Table 1. Examples from source dataset machine-translated sentences that build into paraphrase pairs for our dataset. Each entry consists of the original English sentence ("en-XX"), its Portuguese machine translation ("MT pt-XX") and the European Portuguese reference ("pt-PT"). These pairs have varying lexical diversity.

en-XX	MT pt-XX	pt-PT
That's for someone else to judge	É para outra pessoa julgar.	Não é a nós que cabe julgar isso
What are you doing with those people, I wondered	O que estão a fazer com essas pessoas, perguntei-me	O que fazes com estas pessoas, perguntei-me eu
You wouldn't want me to pretend	Vocês não querem que eu fingue.	Vais querer que eu finja?
But I was able to find out that her area of expertise was gerontology	Mas pude descobrir que a sua área de especialidade era a geronologia.	Mas eu consegui descobrir que a sua área profissional era a gerontologia
You all right?	Está bem?	Estás bem?
Guys, it was like a circus out there	Rapaces, era como um circo lá fora	Rapazes, estava muita confusão

Because pivot translation can potentially diminish the fidelity of the information forwarded into the target language, we chose parallel data containing text in European Portuguese, from which we can translate the side which is not European Portuguese. This is the approach from [37]. Additionally, in [38], the authors found little difference among Czech, German, and French as source languages for back-translation from English. As for Portuguese, we did not find prior work focusing on the best source language to translate from. As such, to maximize performance, we chose English as our language to translate from and an English-centric multilingual pre-trained language model, such as mBART-50 [35]. This model extends the original mBART [24] to encompass more languages, including Portuguese.

3.1 Choosing a Data Source

As far as we know, the two primary publicly available datasets with European Portuguese bitext are Europarl [20] and OpenSubtitles [23]. As per the study conducted in [37], Europarl exhibits low diversity in terms of rare word usage, vocabulary entropy, and parse entropy, mainly due to the formulaic and repetitive nature of speech in a Parliament. In [37], the authors chose the CzEng dataset [8], of which a significant portion is movie subtitles which tend to use a vast vocabulary and have a diversity of sentence structures. This serves as a strong motivation for conducting our experiments using OpenSubtitles.

The OpenSubtitles dataset has over 33 million English-European Portuguese bitext pairs. Because of the computational expense of translating such an extensive dataset, we sample 3 million entries. When translating the English sentences to Portuguese, we used beam search with a beam size of 5 and selected the highest-scoring translation. We show illustrative examples in Table 1. Note the matching is not always perfect, mainly because the original bitext pairs not being perfect translations (there are instances where the meaning is significantly different). The translations are of very high quality, with sporadic errors like gender-mismatch due to English having no gendered nouns, or translations failing to discern whether a second person pronoun ("you") is singular or plural.

3.2 Automatic Quality Assessment, Cleaning, and Filtering

As manually evaluating such an extensive dataset is very expensive and time-consuming, we resort to automatic mechanisms to assess the dataset's quality and clean and filter uninteresting information.

We found recurring problems on manual inspection, like close captions, start hyphenation, and sentence misalignment. For example, "(vomita) Tu queres saber o que é de loucos?" has a close caption that should be removed. Similarly, in "- Deem-me dois minutos.", the hyphen should be removed to match the target sentence. An example of the misalignment is 'E Dr^{a}. Lin, tente não me chamar." → "Sim.", where the two sentences do not share the same meaning. To find these pairs, we search for big differences in token size between source and target. We use the following equation to prune heavily uneven word counts while normalizing for text sizes:

$$|n\_tokens_{src} - n\_tokens_{tgt}| / \max(n\_tokens_{src}, n\_tokens_{tgt}) > 0.5$$

We arbitrate the threshold value to be 0.5 based on a few empirical experiments. For a random sample of 100 000 entries, we find around 3 500 entries that do not match the above equation (are deemed unfit to keep). The mean SBERT score for this sample is 81.69, a low value for SBERT, indicating that these pairs with heavily uneven word counts have a low semantic similarity.

Finally, we remove sentence pairs that are exactly the same. This behavior occurs most prominently for very small sentences (<4 tokens).

3.3 Data Analysis

We further analyze the relevance of the data. As per Li et al. [22], relevance regards how semantically close the paraphrase text is to the original text. We study the semantic similarity resorting to Sentence BERT [30] (SBERT). Specifically, we conduct preliminary testing with multilingual SBERT (mSBERT) [31], and a Brazilian Portuguese SBERT[2] trained on ASSIN2 [29]. Despite being more general-purpose, we found mSBERT performs better than the latter. Using

[2] This model can be found on the HuggingFace as "ricardo-filho/bert-base-portuguese-cased-nli-assin-2".

mSBERT and normalizing the scores in the range of [0, 1], we get an average value of 87.724, which suggests the majority of the pairs have high semantic similarity between them. Nonetheless, we prune pairs with semantic scores lower than 80. From empirical assessment, from this threshold on, most sentence pairs are misaligned.

We do not conduct any particular study regarding fluency (the syntactic and grammar correctness of the paraphrased text [22]), relying on the assumption that pre-trained language models are inherently good grammar inductors [19].

OSPT has 1 519554 pairs. For reference, two widely used English sentential parallel paraphrase datasets, QQP[3] and PAWS [39] have respectively 1 49263 and 2 8904 paraphrase pairs. TaPaCo [32], a corpus of sentential paraphrases for various languages, has 3 6451 Brazilian Portuguese paraphrase pairs. The OSPT averages 8 words for both source and target sentences, as subtitles are rarely long. QQP averages around 11 words per sentence, PAWS around 21, and TaPaCo around 7.

4 Learning Sentence Embeddings

We assess the quality of the dataset intrinsically, using it to train sentence embeddings.

4.1 Experimental Setup

We fine-tune a mSBERT [31] model. We train the model for 10 epochs with a batch size of 64, a learning rate of 2e-5, AdamW optimizer and a linear scheduler with 100 warmup steps. As referred to in Sect. 2, the training loss we use allows for training good quality sentence embeddings without negative examples. The training data for the loss consists of sentence pairs $[(a_1, b_1), \ldots, (a_n, b_n)]$ where we assume that (a_i, b_i) are similar sentences and (a_i, b_j) are dissimilar sentences for $i \neq j$. It minimizes the distance (cosine similarity) between a_i and b_i while maximizing the distance between a_i and b_j for all $i \neq j$.

We evaluate sentence embeddings using the ASSIN2 semantic textual similarity (STS) tasks [29]. Given two sentences, the aim of the STS tasks is to predict their similarity on a 0-5 scale, where 0 indicates the sentences are on different topics and 5 means they are entirely equivalent. To fairly compare OSPT with ASSIN2's official training data (with 6 500 pairs), we randomly sampled a subset of 6.5K pairs from our dataset. We further compare with a 6500-pair subset of the TaPaCo dataset.

[3] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs.

4.2 Results

In Table 2, we report the scores for the official task's evaluation metrics.

Table 2. Results for STS on the ASSIN2 test set. We compare three fine-tuned SBERT models, one using the ASSIN2 training data (has 6 500 pairs), other using a random subset of 6 500 samples from TaPaCo [32], and the other using a random subset of 6 500 samples from OSPT. We report the official metrics from the STS tasks of the ASSIN2 competition. The best results are in **bold**.

	Pearson's r	MSE
SBERT-ASSIN2	0.711	0.03
SBERT-TaPaCo	0.763	**0.02**
SBERT-OSPT	**0.780**	**0.02**

The results reported compare the same model trained under the same conditions, and with the same amount of data, only changing the data source. The mSBERT trained with a subset of OSPT performed the best for the task, achieving the highest Pearson's r and MSE values. Assuming the randomly sampled subsets to be good representations of the data as a whole (which is hard to assess if this is true for sentences), we can conclude the data to be of good quality, or at least, to be good enough to produce good quality sentence embeddings.

5 Paraphrase Generation

Besides creating state-of-the-art paraphrastic sentence embeddings, we show our dataset can help produce interesting paraphrase generators for data augmentation.

5.1 Experimental Setup

We fine-tune three mBART [24] models, two on subsets of OSPT and another on the TaPaCo dataset. Since our dataset is so large, it is computationally demanding to train paraphrase generation models in its entirety. As such, we filter the data to create a training set of 240K samples, 30K samples for validation, and 30K for testing. Additionally, we build a subset of OSPT of 36451 training pairs (the same size as TaPaCo) for fair comparison. We train both models for four epochs, with a batch size of 64, a learning rate of 1e-4, AdamW optimizer, and a linear scheduler with 100 warmup steps.

Following recent work [17], we use as our primary evaluation metric the **iBLEU** [33] score:

$$\begin{aligned}\text{iBLEU} = {} & \alpha \cdot \text{BLEU}(outputs, references) \\ & - (1 - \alpha) \cdot \text{BLEU}(outputs, inputs)\end{aligned}$$

iBLEU measures the fidelity of generated outputs to reference paraphrases as well as the level of diversity introduced. We set α as 0.7 per the original paper [33]. Additionally, to probe the semantic retention of the generations, we measure the semantic similarity using mSBERT [30]. We chose this metric because it was found to have the lowest coupling between semantic similarity and linguistic diversity [2].

5.2 Results

We evaluate paraphrase generation using the ASSIN2 competition's test set.

Table 3. Top-1 results for automatic evaluation on the ASSIN2 test set. The **Source as prediction** baseline serves as a dataset quality indicator. The naming convention matches the number of pairs used to train the models. The best results are in **bold**.

	iBLEU↑	SBERT↑
Source as prediction	−9.9	74.876
mBart-OSPT-240k	−2.5	70.476
mBart-OSPT-36k	**−2.3**	69.324
mBart-TaPaCo	−3.6	**71.048**

Table 3 shows the performance of the two mBART-based models we fine-tuned. The results are bound to the basic statistics of the data, hence why we report the *source as prediction*, that is, using the source sentences as predictions. The ASSIN2 pairs have high word overlap, expressed as a low iBLEU score in the *source as prediction* baseline. Consequently, models trained on that data will produce sentences similar to the sources. That is why the iBLEU scores are low across the board. These iBLEU values could be made higher by increasing the α hyper-parameter, but we would be reducing the contribution of lexicon diversity for the results. Nevertheless, we can see that we can improve diversity by having more diverse generations (expressed as a higher iBLEU score) with a drop in the semantics (even though the metric is not fully decoupled from the vocabulary used). The model trained on OSPT-36k achieves the highest diversity but at the cost of some semantic preservation. Ramping up the number of training examples to 240k has a minimal decrease in diversity with increased semantic fidelity, much closer to the model trained on TaPaCo. Note that we did not fiddle with hyper-parameters, and four epochs may not be sufficient for achieving optimal performance considering the complexity and size of our model, hence why the larger model is not clearly better than the smaller one. Notice that TaPaCo is a Brazilian Portuguese dataset, such as ASSIN2, making it likely to perform better in this specific context, as we are trying to produce European Portuguese text. Moreover, this ASSIN2 test set contains texts with low syntactic diversity and many uses of the gerund form of the verbs, a pattern most prevalent in Brazilian Portuguese.

Table 4. Example generations from the mBART-OSPT-240k model on the ASSIN2 test set illustrating the *pseudo*-translation.

Original	Generation
Alguém está tocando um piano	Está alguém a tocar piano
O homem está falando ao telefone	O homem está a falar ao telefone
Um homem negro está andando no pavimento	Está um negro a caminhar no chão
Duas mulheres estão dançando	Duas mulheres dançam

Table 4 shows some examples of these sentences and the respective generations from the mBart-OSPT-240k model. We can produce European Portuguese paraphrases by building the training pairs with the European Portuguese as targets, even when paraphrasing from Brazilian Portuguese. Our model performs a *pseudo*-translation from Brazilian Portuguese to European Portuguese.

Future work could use the properties mentioned above of the paraphrase generator to further denoise the dataset we present in this paper. We could use the generations of this paraphrase generator to convert the source sentences of our dataset into European-like Portuguese. We can also consider generalizing the approach and employing this technique to convert any Brazilian Portuguese text into European Portuguese.

6 Conclusion

We described the creation of a dataset of more than 1.5M Portuguese sentential paraphrase pairs. We showed how to use this dataset to train paraphrastic sentence embeddings that outperform systems trained with other data on STS tasks, as well as how it can be used for generating paraphrases for purposes of data augmentation and *pseudo*-translate from Brazilian Portuguese to European Portuguese.

The key advantage of our approach is that it only requires parallel text and a translation system. There are hundreds of millions of parallel sentence pairs, and more are being generated continually. Our procedure immediately applies to the wide range of languages for which we have parallel text. Additionally, the quality of the datasets generated using this approach will increase in parallel with improvements in Machine Translation.

We release our dataset, code, and pre-trained sentence embeddings.[4]

This work is supported by LIACC, funded by national funds through FCT/MCTES (PIDDAC), with reference UIDB/00027/2020.

Acknowledgments. The first author is supported by a PhD studentship with reference 2022.13409.BD from Fundação para a Ciência e a Tecnologia (FCT).

[4] We will release code and embeddings under the permissive MIT license.

References

1. Arora, S., Liang, Y., Ma, T.: A simple but tough-to-beat baseline for sentence embeddings. In: International Conference on Learning Representations (2017)
2. Bandel, E., Aharonov, R., Shmueli-Scheuer, M., Shnayderman, I., Slonim, N., Ein-Dor, L.: Quality controlled paraphrase generation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 596–609. Association for Computational Linguistics, Dublin, Ireland, May 2022. https://doi.org/10.18653/v1/2022.acl-long.45
3. Bannard, C., Callison-Burch, C.: Paraphrasing with bilingual parallel corpora. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pp. 597–604. Association for Computational Linguistics, Ann Arbor, Michigan, Jun 2005. https://doi.org/10.3115/1219840.1219914
4. Barreiro, A., Mota, C.: e-pact: esperto paraphrase aligned corpus of en-ep/bp translations. Traduçao em Revista **1**(22), 87–102 (2017)
5. Barreiro, A., Mota, C., Baptista, J., Chacoto, L., Carvalho, P.: Linguistic resources for paraphrase generation in portuguese: a lexicon-grammar approach. Lang. Resour. Eval. (2021)
6. Barzilay, R., McKeown, K.R.: Extracting paraphrases from a parallel corpus. In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics, pp. 50–57. Association for Computational Linguistics, Toulouse, France, Jul 2001. https://doi.org/10.3115/1073012.1073020
7. Bhagat, R., Hovy, E.: Squibs: what is a paraphrase? Comput. Linguist. **39**(3), 463–472 (2013)
8. Bojar, O., Dušek, O., Kocmi, T., Libovickỳ, J., Novák, M., Popel, M., Sudarikov, R., Variš, D.: Czeng 1.6: enlarged czech-english parallel corpus with processing tools dockered. In: Text, Speech, and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, 12–16 Sept. 2016, Proceedings 19. pp. 231–238. Springer (2016)
9. Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A.: Supervised learning of universal sentence representations from natural language inference data. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 670–680. Association for Computational Linguistics, Copenhagen, Denmark, Sept. 2017. https://doi.org/10.18653/v1/D17-1070
10. Dolan, W.B., Brockett, C.: Automatically constructing a corpus of sentential paraphrases. In: Proceedings of the Third International Workshop on Paraphrasing (IWP2005) (2005)
11. Gan, Z., Pu, Y., Henao, R., Li, C., He, X., Carin, L.: Learning generic sentence representations using convolutional neural networks. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2390–2400. Association for Computational Linguistics, Copenhagen, Denmark, Sept. 2017. https://doi.org/10.18653/v1/D17-1254
12. Ganitkevitch, J., Callison-Burch, C.: The multilingual paraphrase database. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 4276–4283. European Language Resources Association (ELRA), Reykjavik, Iceland, May 2014
13. Ganitkevitch, J., Van Durme, B., Callison-Burch, C.: PPDB: the paraphrase database. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 758–764. Association for Computational Linguistics, Atlanta, Georgia, Jun 2013

14. Gao, T., Yao, X., Chen, D.: SimCSE: simple contrastive learning of sentence embeddings. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 6894–6910. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, Nov 2021. https://doi.org/10.18653/v1/2021.emnlp-main.552
15. Henderson, M., Al-Rfou, R., Strope, B., Sung, Y.H., Lukács, L., Guo, R., Kumar, S., Miklos, B., Kurzweil, R.: Efficient natural language response suggestion for smart reply (2017). arXiv:1705.00652
16. Hill, F., Cho, K., Korhonen, A.: Learning distributed representations of sentences from unlabelled data. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1367–1377. Association for Computational Linguistics, San Diego, CA, June 2016. https://doi.org/10.18653/v1/N16-1162
17. Hosking, T., Tang, H., Lapata, M.: Hierarchical sketch induction for paraphrase generation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2489–2501. Association for Computational Linguistics, Dublin, Ireland, May 2022. https://doi.org/10.18653/v1/2022.acl-long.178
18. Jiang, Y., Kummerfeld, J.K., Lasecki, W.S.: Understanding task design trade-offs in crowdsourced paraphrase collection. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 103–109. Association for Computational Linguistics, Vancouver, Canada, Jul 2017. https://doi.org/10.18653/v1/P17-2017
19. Kim, T., Choi, J., Edmiston, D., Goo Lee, S.: Are pre-trained language models aware of phrases? Simple but strong baselines for grammar induction. In: International Conference on Learning Representations (2020)
20. Koehn, P.: Europarl: a parallel corpus for statistical machine translation. In: Proceedings of Machine Translation Summit X: Papers, Phuket, Thailand, pp. 79–86, 13–15 Sept. 2005
21. Lan, W., Qiu, S., He, H., Xu, W.: A continuously growing dataset of sentential paraphrases. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1224–1234. Association for Computational Linguistics, Copenhagen, Denmark, Sept. 2017. https://doi.org/10.18653/v1/D17-1126
22. Li, Z., Jiang, X., Shang, L., Li, H.: Paraphrase generation with deep reinforcement learning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3865–3878. Association for Computational Linguistics, Brussels, Belgium, Oct.–Nov. 2018. https://doi.org/10.18653/v1/D18-1421 ¡error l="308" c="Invalid
command: paragraph not started." /¿
23. Lison, P., Tiedemann, J.: OpenSubtitles2016: extracting large parallel corpora from movie and TV subtitles. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 923–929. European Language Resources Association (ELRA), Portorož, Slovenia, May 2016
24. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Trans. Assoc. Comput. Linguist. **8**, 726–742 (2020)
25. Mallinson, J., Sennrich, R., Lapata, M.: Paraphrasing revisited with neural machine translation. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pp. 881–893. Association for Computational Linguistics, Valencia, Spain, Apr 2017

26. Mrkšić, N., Ó Séaghdha, D., Thomson, B., Gašić, M., Rojas-Barahona, L.M., Su, P.H., Vandyke, D., Wen, T.H., Young, S.: Counter-fitting word vectors to linguistic constraints. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 142–148. Association for Computational Linguistics, San Diego, CA, June 2016. https://doi.org/10.18653/v1/N16-1018
27. Pagliardini, M., Gupta, P., Jaggi, M.: Unsupervised learning of sentence embeddings using compositional n-gram features. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 528–540. Association for Computational Linguistics, New Orleans, Louisiana, June 2018. https://doi.org/10.18653/v1/N18-1049
28. Pellicer, L.F.A.O., Pirozelli, P., Costa, A.H.R., Inoue, A.: PTT5-paraphraser: diversity and meaning fidelity in automatic portuguese paraphrasing. In: Computational Processing of the Portuguese Language: 15th International Conference, PROPOR 2022, Fortaleza, Brazil, 21–23 Mar. 2022, Proceedings, pp. 299–309. Springer (2022)
29. Real, L., Fonseca, E., Oliveira, H.G.: The ASSIN 2 shared task: a quick overview. In: International Conference on Computational Processing of the Portuguese Language, pp. 406–412. Springer (2020)
30. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Nov. 2019
31. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Nov. 2020
32. Scherrer, Y.: TaPaCo: a corpus of sentential paraphrases for 73 languages. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 6868–6873. European Language Resources Association, Marseille, France, May 2020
33. Sun, H., Zhou, M.: Joint learning of a dual SMT system for paraphrase generation. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 38–42. Association for Computational Linguistics, Jeju Island, Korea, July 2012
34. Suzuki, Y., Kajiwara, T., Komachi, M.: Building a non-trivial paraphrase corpus using multiple machine translation systems. In: Proceedings of ACL 2017, Student Research Workshop, pp. 36–42. Association for Computational Linguistics, Vancouver, Canada, Jul 2017
35. Tang, Y., Tran, C., Li, X., Chen, P.J., Goyal, N., Chaudhary, V., Gu, J., Fan, A.: Multilingual translation with extensible multilingual pretraining and finetuning (2020). arXiv:2008.00401
36. Wieting, J., Bansal, M., Gimpel, K., Livescu, K.: From paraphrase database to compositional paraphrase model and back. Trans. Assoc. Comput. Linguist. **3**, 345–358 (2015)
37. Wieting, J., Gimpel, K.: ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 451–462. Association for Computational Linguistics, Melbourne, Australia, July 2018. https://doi.org/10.18653/v1/P18-1042

38. Wieting, J., Mallinson, J., Gimpel, K.: Learning paraphrastic sentence embeddings from back-translated bitext. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 274–285. Association for Computational Linguistics, Copenhagen, Denmark, Sept. 2017. https://doi.org/10.18653/v1/D17-1026
39. Zhang, Y., Baldridge, J., He, L.: PAWS: Paraphrase adversaries from word scrambling. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 1298–1308. Association for Computational Linguistics, Minneapolis, Minnesota, June 2019. https://doi.org/10.18653/v1/N19-1131