Monolingual Paraphrase Detection Corpus for Low Resource Pashto Language at Sentence Level

Iqra Ali, Hidetaka Kamigaito, Taro Watanabe
Nara Institute of Science and Technology, Nara, Japan
{ ali.iqra.ai6, kamigaito.h, taro}@is.naist.jp

Abstract

Paraphrase detection is a task to identify if two sentences are semantically similar or not. It plays an important role in maintaining the integrity of written work such as plagiarism detection and text reuse detection. Formerly, researchers focused on developing large corpora for English. However, no research has been conducted on sentence-level paraphrase detection in low-resource Pashto language. To bridge this gap, we introduce the first fully manually annotated Pashto sentential paraphrase detection corpus collected from authentic cases in journalism covering 10 different domains, including Sports, Health, Environment, and more. Our proposed corpus contains 6,727 sentences, encompassing 3,687 paraphrased and 3,040 non-paraphrased. Experimental findings reveal that our proposed corpus is sufficient to train XLM-RoBERTa to accurately detect paraphrased sentence pairs in Pashto with an F1 score of 84%. To compare our corpus with those in other languages, we also applied our fine-tuned model to the Indonesian and English paraphrase datasets in a zero-shot manner, achieving F1 scores of 82% and 78%, respectively. This result indicates that the quality of our corpus is not less than commonly used datasets. It's a pioneering contribution to the field. We will publicize a subset of 1,800 instances from our corpus, free from any licensing issues.

Keywords: Pashto Paraphrase Detection, Corpus Collection, Low Resource NLP

1. Introduction

Paraphrase detection (Vrbanec and Meštrović, 2020; Sameen et al., 2017) identifies the relation between text pairs and categorizes them as paraphrased if they convey the same semantics, in spite of potentially having different wording or syntax, otherwise non-paraphrased. Paraphrase detection is one of the emerging research topics in natural language processing having multiple fundamental applications like natural language understanding (Cho et al., 2019b), plagiarism detection (Mozgovoy et al., 2010), copyright infringement (Clough et al., 2002), natural language generation (Cho et al., 2019a), information extraction (Shinyama and Sekine, 2003), machine translation (Resnik et al., 2010), question answering (Fader et al., 2013), and text recapitulation.

Formerly, the core focus of researchers was inclined towards high-resource languages like English (Dolan and Brockett, 2005), Chinese (Zhang et al., 2019), Spanish (Tamayo et al., 2022), French (Richard et al., 2023), Japanese (Nakagawa and Masuda, 2004), and more European languages. However, there is a need to expand the technologies for natural language processing to all languages in the world for the benefit of any people regardless of language barriers. For instance, Microsoft Research Paraphrase Corpus (Dolan and Brockett, 2005), Clough and Stevenson (Clough and Stevenson, 2011a), Webis Crowd Paraphrase Corpus (Vrbanec and Meštrović, 2020) all contained corpus for English only.

The natural language processing models are based on deep learning (Kenton and Toutanova, 2019) which demands training data to train a model and a benchmark data (Wang et al., 2018) to verify the gains. Paraphrase detection models also follow the same practice, to develop and compare any paraphrase detection models, benchmark corpora are needed. The natural language processing community is unfolding very fast. Recently, researchers started focusing on other low-resource South Asian languages too and some corpora have been created at sentence (Muneer and Nawab, 2022a; Hafeez et al., 2023), phrasal (Muneer and Nawab, 2022b) and document level (Gaizauskas et al., 2001; Sharjeel et al., 2023) for South Asian Urdu language which indicates researcher's interest towards the low resource languages.

According to Haq et al., Pashto remains a relatively unexplored language in NLP research, due to the absence of publicly available corpus and the challenges associated with collecting and annotating Pashto corpora. To the best of our knowledge, no significant research has been channeled on sentence-level paraphrase detection in Pashto. From Table 1, we can see that translation is unsuitable for making a Pashto paraphrase detection corpus, as in English the

LREC-COLING 2024, pages 11574–11581

Sentence 1	Sentence 2
د پښتو ویونکی یم چې په پای کې د پوستکي ارزښت لري. ایا تاسو به د هر څه لپاره خپل پوست قرباني کړئ؟	په پښتو خبرې کوم، چې تر هر څه زیات ارزښت لرم. ایا تاسو چمتو یاست چې د هر څه لپاره خپل صداقت قرباني کړ؟
I am a Pashto speaker who values skin in the end. Would you sacrifice your skin for anything?	I speak Pashto, which I value more than anything else. Are you willing to sacrifice your integrity for anything?

Table 1: In the above example, the idea of "sacrificing one's skin" symbolizes pride, dignity, or commitment indicating cultural connotation. However, the cultural connotation is not fully captured in the English translation, as the word "skin" is used more literally in the translated sentences, and the automatic translation is not suitable for creating a Pashto paraphrase detection corpus, as in English the cultural connotation is not fully transferred during translation.

cultural connotation is not fully captured during translation. To bridge this gap, we are proposing the first monolingual paraphrase detection corpus for Pashto. Our contribution is 2-fold: First, we constructed a corpus comprising 6,727 sentences, encompassing 3,687 paraphrased and 3,040 non-paraphrased instances. All the instances were manually collected from journalism websites with each sentence pair labeled either paraphrased or not by human annotator. The subset [1] of 1,800 instances from our constructed corpus, free from any licensing issues, will be made accessible publicly for research purpose.

Second, we trained Pashto paraphrase detection model by fine-tuning the pre-trained XLM-RoBERTa [2] on our newly proposed corpus, resulting in an F1-score of 84%. In contrast, we also applied our fine-tuned model to the Indonesian and English paraphrase datasets in a zero-shot manner, achieving F1 scores of 82% and 78%, respectively, indicating that the quality of our corpus is not less than commonly used datasets.

2. Related Work

Paraphrase detection is a sphere of interest for researchers in the natural language processing community, mostly the corpora developed in the past are available for the English language which include sentential paraphrase corpus (Alvi et al., 2012), paraphrase for plagiarism (Clough and Stevenson, 2011b), and microsoft research paraphrase corpus (Dolan and Brockett, 2005).

The originality of our study lies from its emphasis on the Pashto language, an area where previous research in paraphrase detection is non-existent. This absence underlines the pioneering nature of the dataset we are developing. The paraphrase datasets for some preeminent South Asian languages like Urdu, Bangla, Punjabi, and Tamil also exists. The corpora created for Urdu language at sentence (Hafeez et al., 2023), phrasal (Muneer and Nawab, 2022b), and document level (Sharjeel et al., 2023), for Punjabi language (Anand Kumar et al., 2018), Tamil (Senthil Kumar et al., 2020), and Bangla (Akil et al., 2022; Ahnaf et al., 2020) shows the sudden interest of researchers towards low-resource South Asian languages.

For methodologies, researchers have used conventional techniques in the past for paraphrase detection tasks. In the era of early 2000's, rule based methods (Lappin and Leass, 1994) were common where the researchers used manual rules for detecting the paraphrase pairs. In the 2000's, the alignment methods (Lin, 1998) and supervised learning (Barzilay and Lee, 2003) gained prominence for paraphrase detection. In late 2000's, distributional semantics (Mikolov et al., 2013) methods like latent semantic analysis and word embeddings became popular among researchers. The rise of deep learning in the 2010's improved the overall paraphrase detection by introducing recurrent neural networks (Ye et al., 2017), and convolutional neural networks (Yin and Schütze, 2015; Zhang et al., 2017). Finally, dragging us to the era of pre-trained language models from 2018 onward revolutionizing the paraphrase detection task (Reimers and Gurevych, 2019), providing the essential embedding and fine-tuned models like GPT (Becker et al., 2023), and BERT (Ta et al., 2022; Peinelt et al., 2020; Khairova et al., 2022) for paraphrase detection.

As far as we are aware, no paraphrase detection corpus at sentence level for Pashto language has been developed previously. This absence underlines the pioneering nature of the dataset we are developing. Moreover, the recent trend in NLP is more inclined towards Transformer-based approaches due to its improve performance, which is the main focus of our methodology.

3. Dataset Creation

Figure 1 shows the overview of our dataset creation process. The next subsections elaborate them.

[1] https://github.com/anonymousrepository11/anonymous.git

[2] https://huggingface.co/xlm-roberta-base

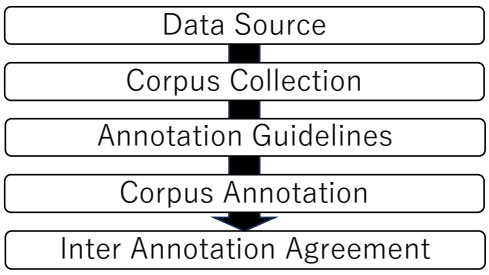

Figure 1: Overview of our dataset creation process.

3.1. Data Source

To develop Pashto paraphrase detection corpus, we needed to identify the source for collecting our corpus. At the start, we thought that collecting Pashto paraphrase corpus from the internet might pose a difficulty. As the digital infiltration is quite limited in Afghanistan and Pakistan and there might not be an abundant amount of data accessible online for Pashto on the internet. After assuring that we had sufficient data online, we identified 10 different but well-known Pashto newspaper websites including Deutsche Welle (DW) [3], Voice of America (VOA) [4], Khyber News [5], Turkish Radio Television (TRT)[6], British Broadcasting Corporation (BBC) [7], and others to collect our data. We chose both international and domestic newspapers for corpus collection. All the newspaper websites contain ample data in Pashto language and are thus a good source for corpus collection.

3.2. Corpus Collection

We manually collected 6,727 sentence pairs to build Pashto Paraphrase Detection corpus from the online Pashto newspapers throughout the duration of 8 to 9 months. In the corpus collection stage, we manually checked a couple of websites and identified the same news occurrence, we presuming that the sentences describing the same news occurrence were paraphrases of each other, otherwise non-paraphrased. Along with the positive instances we also collected non-paraphrase text to have the necessary segment of negative instances in the dataset. The sentences were extracted from the newspaper headlines as well as comprehensive stories published on Pashto newspapers websites. To maintain the diversity of corpus it was collected from 10 different genres as depicted in Table 2.

[3] https://www.dw.com/ps/
[4] https://www.pashtovoa.com/
[5] https://khybernews.tv/pu/
[6] https://www.trt.net.tr/pashto/
[7] https://www.bbc.com/pashto

Genre	No. of Instances
Sports	800
Crime and Law	600
Health	703
Politics and Conflicts	700
Natural Disasters and Accidents	500
Science and Technology	700
Environment	700
Economy and International Affairs	600
Culture and Entertainment	700
Weather	724

Table 2: Genre and No. of Instances

3.3. Corpus Annotation

Upon completion of corpus collection, two human annotators were asked to label each sentences manually. The annotators were graduate with a background in computer science, aged between 20 to 30, and *native speakers* of Pashto. The annotators were provided with complete annotation guidelines described in section 3.4. The labeling was done by the annotators based on the guidelines provided by us. The annotators were told to use their knowledge to label instances as either paraphrased or non-paraphrased along with the given guidelines.

3.4. Annotation Guidelines

Our main objective is to develop a Pashto paraphrase detection corpus that includes two tiers of paraphrasing. The tagging guidelines were prepared in the footsteps of (Sameen et al., 2017; Sharjeel et al., 2017) to label a sentence pair in our corpus:

- **Paraphrased (P)**: A text pair is paraphrased if both texts describe the same news incident or event or article, but using different wording or structures.

  - The key criterion of this semantic equivalence, such as the core information and meaning in both texts should remain the same.

  -There are several techniques that indicate paraphrasing like: the use of synonyms, changes in sentence structure (for instance switching from active to passive voice), and the addition or deletion of words or phrases.

- **Non-Paraphrased (NP)**: A text pair should be called as non-paraphrased when the contents, while potentially relating to the same event or news story or article, are presented in a distinctly different manner without significant overlap in language or structure.

  -This includes situations where the texts have divergent details, perspectives, or stylistic choices that set them apart.

Paraphrased Sentence Pair	
Sentence 1	**Sentence 2**
نوموړي هغه مهال ويلي وو زه نېکمرغه يم چې پيسې مې ګټلې. اوس پر دې ډېر زيات خوښ يم چې يوه برخه يې بېرته ورکوم. He said at that time, I am lucky to have earned money. Now I am more than happy to give back a part of it.	زه نېکمرغه يم چې پيسې مې ګټلې دي. اوس له دې زيات خوښ يم چې يوه برخه يې بېرته ورکړم. ښاغلی لي وايي I am lucky to have earned the money. Now I am more than happy to give some of it back. Mr. Lee says
Non-Paraphrased Sentence Pair	
Sentence 1	**Sentence 2**
په امريکا کې د يوې يتيمې افغانې نجلۍ کيسه وروسته له هغې د رسنيو د پام وړ ګرځېدلې چې په ويرجينيا ايالت کې محکمې د هغې پر ضد د تښتونې دوسيه لغوه کړه. The story of an orphaned Afghan girl in the United States has become the center of media attention after the court in Virginia overturned the kidnapping case against her.	حکومت په افغانستان کې کورنۍ ونه موندله، نو يوې بلې کورنۍ د هغې پالنه وکړه او نجلۍ يې د اخراج په بهير کې امريکا ته يووړه. The government could not find the family in Afghanistan, so another family took care of the girl in America.

Table 3: Paraphrased and Non-Paraphrased Sentence Pair Examples

Total words	13,454	
Total unique words	12,624	
Average length	23.97	
Vocabulary richness (TTR)	0.93	
Label	P	NP
Total pairs	3,687	3,040
Max no. of words	382	473
Min no. of words	3	3
Mean of words	82.3	125.4
Median of words	75.0	117.0

Table 4: Corpus Characteristics

3.4.1. Result of Inter-Annotator Agreement

Annotations were performed in two rounds. In the first stage, based on the annotation guidelines, a random subset of 1000 sentence pairs was annotated by the two annotators. The result of labels tagged by each annotator was compared, and conflicting pairs were discussed with them individually. In the second stage, the remaining corpus was annotated by the annotators. Both annotators agreed on 6,406 and disagreed on 321 sentence pairs. The conflicted 321 pairs are the part of dataset to maintain the complexity of our dataset. We achieved the Inter-Annotator Agreement (IAA) = 90%. The IAA score is good, considering the task and it shows that annotation guidelines were clear and easy to follow.

3.5. Example of a paraphrased and non-paraphrased Text

Table 3 shows the paraphrased example, as the sentence is transformed by incorporating the insertion and deletion of a new text but still conveying the same meaning.

However, Table 3 shows that the sentence pair is describing the same story but do not have any semantic similarity.

3.6. Corpus Statistics

The statistics from Table 4 reflect that the corpus is relatively balanced in terms of paraphrase (54.81%) and non-paraphrase (45.19%) text pairs with total 6,727 sentence pairs. The mean of paraphrased and non-paraphrased sentences is 82.3 and 125.4 words respectively. All statistics indicate that sentences in our corpus are diverse, making the corpus more realistic and challenging for Pashto paraphrase detection task.

4. Experiments

Settings As a multilingual pre-trained model, we used XLM-RoBERTa (Conneau et al., 2019)[8] provided by HuggingFace Transformers (Wolf et al., 2020) in our experiments as it's pre-trained on multiple languages including Pashto. To verify the effectiveness of our created Pashto corpus, we evaluated the model in our corpus based on the following settings:

BERTScore: We also employ BERTScore, a robust cross-lingual capable scoring system, to measure the semantic similarity within our Pashto text data. The BERTScore, as depicted in Table 5, serves as a quantitative tool for assessing the quality of our proposed dataset.

Specifically, the BERTScore depicts that for paraphrase (P) instances the mean F1 score is 88%, indicating a high level similarity between the sentences. In contrast, for non-paraphrase (NP) instances, the F1 score is 85%. This comparative score between P and NP cases hint towards the closeness of the pairs and data complexity in terms of similar and dissimilar sentence pairs. The results of this analysis, illustrated in Figure 2, are instrumental in pinpointing the challenging aspects of paraphrase detection task.

[8] https://huggingface.co/xlm-roberta-base

Label	Precision	Recall	F1
P	0.895	0.880	0.887
NP	0.850	0.851	0.850

Table 5: Results of **Mean BERTScores** for paraphrased (P) and non-paraphrased (NP) instances.

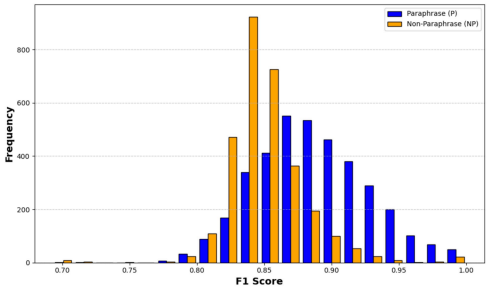

Figure 2: The result of BERTScore across Pashto paraphrase detection dataset.

- **Zero-shot**: This is a baseline setting. In this setting, we fine-tuned XLM-RoBERTa on paraphrase detection corpus of Indonesian "id-paraphrase-detection"[9], and English (Dolan and Brockett, 2005) without using our Pashto corpus.
- **Pashto (our)**: In this setting, we directly fine-tuned XLM-RoBERTa on the training split of our created Pashto corpus. In the above settings, we set the batch size to 32, the number of epochs to 3, and the learning rate to $2e-5$.

Results Table 6 shows the experimental result. From the result we can understand that without training data in Pashto, paraphrase detection performance in Pashto is largely degraded. Our future studies will leverage the benchmark dataset we have developed as a foundational tool for exploring new techniques and methods in the area of paraphrase detection for low resource languages.

5. Conclusion

Paraphrase detection datasets are important for the improvement of algorithms dedicated to identifying semantic equivalences between textual data, thereby facilitating a deeper understanding involved in linguistic paraphrasing. This study introduces an extensive benchmark corpus designed for the detection of sentence-level paraphrases within the Pashto language. The corpus encompasses a total of 6,272 sentence pairs, meticulously collected from various online Pashto newspaper websites, consisting of 3,687 paraphrases and 3,040 non-paraphrases. A notable finding from our research

[9] https://huggingface.co/datasets/jakartaresearch/id-paraphrase-detection

Setting	Precision	Recall	F1
Zero-shot (English)	79	78	78
Zero-shot (Indonesian)	83	82	82
Pashto (Our)	**85**	**85**	**84**

Table 6: Results on each setting. The bold values indicates the best score.

is the effectiveness of the corpus in training and fine-tuning the XLM-RoBERTa model for the task of paraphrase detection in Pashto language.

In future, our research will focus on a comprehensive linguistic analysis of the corpus to examine the predominant paraphrasing techniques employed within these texts. Furthermore, we plan to undertake a rigorous evaluation of the corpus through advanced semantic analysis and the application of cutting-edge deep learning methodologies. This exploration will also include the development of custom trained models leveraging Sentence Transformers, tailored specifically to enhance the performance and accuracy of paraphrase detection within our Pashto corpus.

6. Ethical Considerations and Limitations

Our proposed paraphrase detection corpus encompasses over 5,000 instances, cannot be fully released due to copyright constraints. During the data collection process, we contacted various news agencies for their consent to use their data. While some agencies agreed, others did not grant permission, due to copyright concerns. Consequently, we are limited to publicly releasing only the data for which we have received explicit permission.

To provide further clarity, the 1,800 instances from our total corpus that can be released contains a mix of paraphrased and non-paraphrased sentence pairs, carefully selected to represent the diversity and complexity of the Pashto language in the context of paraphrase detection. The subset has been curated to ensure it is representative of our larger corpus, thus providing researchers with a robust and valuable resource for their future research in Pashto language.

It is also important to note that the data is collected from both international and domestic news sources, and it may contain biased opinions related to political matters. Additionally, the non-paraphrase sentences have been chosen arbitrarily, and that might have some bias as well.

7. Bibliographical References

Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. *arXiv preprint cs/0304006*.

Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2023. Paraphrase detection: Human vs. machine content. *arXiv preprint arXiv:2303.13989*.

Eunah Cho, He Xie, and William M Campbell. 2019a. Paraphrase generation for semi-supervised learning in nlu. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 45–54.

Eunah Cho, He Xie, John P. Lalor, Varun Kumar, and William M. Campbell. 2019b. Efficient semi-supervised learning for natural language understanding by optimizing diversity. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1077–1084.

Paul Clough, Robert Gaizauskas, Scott SL Piao, and Yorick Wilks. 2002. Measuring text reuse. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 152–159.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

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.

Ijazul Haq, Weidong Qiu, Jie Guo, and Peng Tang. 2023. Nlpashto: Nlp toolkit for low-resource pashto language. *International Journal of Advanced Computer Science and Applications*, 14(6).

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Nina Khairova, Anastasiia Shapovalova, Orken Mamyrbayev, Nataliia Sharonova, and Kuralay Mukhsina. 2022. Using bert model to identify sentences paraphrase in the news corpus. In *CEUR Workshop Proceedings*, volume 3171, pages 38–48.

Shalom Lappin and Herbert J Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4):535–561.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 768–774.

Swetha Mandava, Szymon Migacz, and Alex Fit Florea. 2020. Pay attention when required. *arXiv preprint arXiv:2009.04534*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Maxim Mozgovoy, Tuomo Kakkonen, and Georgina Cosma. 2010. Automatic student plagiarism detection: future perspectives. *Journal of Educational Computing Research*, 43(4):511–531.

Animesh Nighojkar and John Licato. 2021. Improving paraphrase detection with the adversarial paraphrasing task. *arXiv preprint arXiv:2106.07691*.

Rodrigo Frassetto Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tbert: Topic models and bert joining forces for semantic similarity detection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7047–7055.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Philip Resnik, Olivia Buzek, Chang Hu, Yakov Kronrod, Alex Quinn, and Benjamin B Bederson. 2010. Improving translation via targeted paraphrasing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 127–137.

Sara Sameen, Muhammad Sharjeel, Rao Muhammad Adeel Nawab, Paul Rayson, and Iqra Muneer. 2017. Measuring short text reuse for the urdu language. *IEEE Access*, 6:7412–7421.

Muhammad Sharjeel, Rao Muhammad Adeel Nawab, and Paul Rayson. 2017. Counter: corpus of urdu news text reuse. *Language resources and evaluation*, 51:777–803.

Yusuke Shinyama and Satoshi Sekine. 2003. Paraphrase acquisition for information extraction. In *Proceedings of the second international workshop on Paraphrasing*, pages 65–71.

Hoang Thang Ta, Abu Bakar Siddiqur Rahman, Lotfollah Najjar, and Alexander Gelbukh. 2022. Gan-bert, an adversarial learning architecture for paraphrase identification. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*.

Tedo Vrbanec and Ana Meštrović. 2020. Corpus-based paraphrase detection experiments and review. *Information*, 11(5):241.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

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*, pages 38–45, Online. Association for Computational Linguistics.

Borui Ye, Guangyu Feng, Anqi Cui, and Ming Li. 2017. Learning question similarity with recurrent neural networks. In *2017 ieee international conference on big knowledge (icbk)*, pages 111–118. IEEE.

Wenpeng Yin and Hinrich Schütze. 2015. Convolutional neural network for paraphrase identification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 901–911.

Xiang Zhang, Wenge Rong, Jingshuang Liu, Chuan Tian, and Zhang Xiong. 2017. Convolution neural network based syntactic and semantic aware paraphrase identification. In *2017 international joint conference on neural networks (ijcnn)*, pages 2158–2163. IEEE.

8. Language Resource References

Adil Ahnaf, Shadhin Saha, and Nahid Hossain. 2020. Closed domain bangla extrinsic monolingual plagiarism detection and corpus creation approach. In *2020 IEEE Region 10 Symposium (TENSYMP)*, pages 146–149. IEEE.

Ajwad Akil, Najrin Sultana, Abhik Bhattacharjee, and Rifat Shahriyar. 2022. Banglaparaphrase: A high-quality bangla paraphrase dataset. *arXiv preprint arXiv:2210.05109*.

Faisal Alvi, El-Sayed M El-Alfy, Wasfi G Al-Khatib, and Radwan E Abdel-Aal. 2012. Analysis and extraction of sentence-level paraphrase sub-corpus in cs education. In *Proceedings of the 13th annual conference on Information technology education*, pages 49–54.

M Anand Kumar, Shivkaran Singh, B Kavirajan, and KP Soman. 2018. Shared task on detecting paraphrases in indian languages (dpil): An overview. In *Text Processing: FIRE 2016 International Workshop, Kolkata, India, December 7–10, 2016, Revised Selected Papers*, pages 128–140. Springer.

Paul Clough and Mark Stevenson. 2011a. Developing a corpus of plagiarised short answers. *Language Resources and Evaluation*, 45:5–24.

Paul Clough and Mark Stevenson. 2011b. Developing a corpus of plagiarised short answers. *Language resources and evaluation*, 45:5–24.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.

Robert Gaizauskas, Jonathan Foster, Yorick Wilks, John Arundel, Paul Clough, and Scott Piao. 2001. The meter corpus: a corpus for analysing journalistic text reuse. In *Proceedings of the corpus linguistics 2001 conference*, volume 1. Lancaster University United Kingdom.

Hamza Hafeez, Iqra Muneer, Muhammad Sharjeel, Muhammad Adnan Ashraf, and Rao Muhammad Adeel Nawab. 2023. Urdu short paraphrase detection at sentence level. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–20.

Iqra Muneer and Rao Muhammad Adeel Nawab. 2022a. Cross-lingual text reuse detection at sentence level for english–urdu language pair. *Computer Speech & Language*, 75:101381.

Iqra Muneer and Rao Muhammad Adeel Nawab. 2022b. Develop corpora and methods for cross-lingual text reuse detection for english urdu language pair at lexical, syntactical, and phrasal levels. *Language Resources and Evaluation*, 56(4):1103–1130.

Hiroshi Nakagawa and Hidetaka Masuda. 2004. Extracting paraphrases of japanese action word of sentence ending part from web and mobile news articles. In *Asia Information Retrieval Symposium*, pages 94–105. Springer.

Ange Richard, Laura Alonzo-Canul, and François Portet. 2023. Fracas: A french annotated corpus of attribution relations in news. *arXiv preprint arXiv:2309.10604*.

B Senthil Kumar, D Thenmozhi, and S Kayalvizhi. 2020. Tamil paraphrase detection using encoder-decoder neural networks. In *Computational Intelligence in Data Science: Third IFIP TC 12 International Conference, ICCIDS 2020, Chennai, India, February 20–22, 2020, Revised Selected Papers 3*, pages 30–42. Springer.

Muhammad Sharjeel, Iqra Muneer, Sumaira Nosheen, Rao Muhammad Adeel Nawab, and Paul Rayson. 2023. Cross-lingual text reuse detection at document level for english-urdu language pair. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Antonio Tamayo, Diego A Burgos, and Alexander Gelbukh. 2022. Using transformers on noisy vs. clean data for paraphrase identification in mexican spanish. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2022)*.

Tedo Vrbanec and Ana Meštrović. 2020. Corpus-based paraphrase detection experiments and review. *Information*, 11(5).

Bowei Zhang, Weiwei Sun, Xiaojun Wan, and Zongming Guo. 2019. Pku paraphrase bank: A sentence-level paraphrase corpus for chinese. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part I 8*, pages 814–826. Springer.