Graduate School of Science and Technology Master’s Thesis Abstract

Laboratory name (Supervisor)	Natural Language Processing (WATANABE TARO　(Professor))		
Student ID	2211405	Submission date	2024 / 7 / 22
Name	ALI IQRA		
Thesis title	**PSPD: Constructing Dataset for Low Resource Pashto Sentential Paraphrase Detection**		

Abstract

Paraphrase detection is critical for tasks like plagiarism identification and text reuse detection. Despite colossal research in high-resource languages, there has been no research on sentential-level paraphrase identification for the under-resourced Pashto. In our work, we present the fully manually annotated Pashto paraphrase detection dataset, sourced from journalism across ten diverse domains. Our proposed dataset, named PSPD, with a total of 6,000 plus sentences, with 3,000 plus paraphrased and non-paraphrased pairs respectively. It has achieved an inter-annotator agreement score of 90%. Using our corpus, the XLM-RoBERTa model achieved an F1-score of 85% for detecting Pashto paraphrases. In a zero-shot settings, the model achieved F1-scores of 82% on Indonesian and 78% on English paraphrase datasets. Additionally, we also evaluated the performance of GPT-4o via zero-shot prompting on our proposed dataset. Our corpus advances NLP research for Pashto, enabling the development of effective NLP applications for Pashto speakers and contributing to multilingual technologies that bridge language barriers.