Public Report - Automatic Dataset Creation for NLP

Introduction

This study develops a novel PLN dataset by combining established data augmentation techniques with robust performance metrics, thereby enabling a comprehensive interpretation and enhancement of the underlying data. The work began with a critical examination of state-of-the-art augmentation methods and evolved into the creation of a methodology tailored to address the unique challenges posed by natural language processing. Notably, the investigation underscored that techniques successful in computer vision are not directly transferable to NLP due to the intricate complexities of language.

Data Augmentation Techniques

Rule-Based Methods

One approach employs rule-based methods that apply fixed linguistic transformations—such as synonym replacement, token manipulation, and graph-based inferences—to efficiently generate new text samples, though these techniques tend to offer only modest improvements.

Example Interpolation

Another approach, example interpolation, merges elements from different sentences within a continuous embedding space to produce natural-sounding hybrid text. This strategy demands careful balancing to ensure that grammaticality and clarity are maintained throughout the generated content.

Model-Based Strategies

A third strategy leverages advanced language models, using techniques such as back-translation and prompt-driven paraphrasing to create contextually appropriate text. Although this method incurs a higher computational cost, it enhances the contextual fidelity of the augmented samples. An important consideration during this phase is the monitoring of latent space representations to maintain semantic alignment with the original data—a factor particularly critical for low-resource languages like Portuguese.

Evaluation Metrics for Data Augmentation

The move toward transformer-generated embeddings has necessitated the development of robust evaluation metrics to address challenges in directly comparing experimental

outcomes.

Assessing Lexical Variety

To measure global and local lexical variation, a variety of metrics were employed, including Jaccard similarity, Unique Trigram Ratio (UTR), self-BLEU, Type-Token Ratio (TTR), and RWORDS. These metrics help assess diversity while mitigating issues such as mode collapse, thereby preserving the coherence of the augmented text.

Ensuring Semantic Integrity and Fluency

In tandem with lexical variety measures, the study utilized methods like the BPRO algorithm to ensure that the core meaning of the original text is maintained. Additional metrics focusing on fluency—such as perplexity and the syntactic log-odds ratio—complemented these methods. Moreover, sentiment consistency was evaluated to confirm that the emotional tone of the text remained intact following augmentation. This integrated evaluation framework illuminates gaps in current methodologies and underscores the necessity of a dynamic, metric-informed strategy to yield meaningful diversity, especially in single-class datasets.

BoostEDA: A Metric-Informed Augmentation Approach

Building on these insights, the study introduces BoostEDA, a refined method that directly integrates quantitative metrics into the text augmentation process. BoostEDA addresses the limitations of conventional EDA techniques by ensuring that all modifications made to the text remain both contextually relevant and linguistically fluent.

Implementation as a Python Pipeline

BoostEDA is implemented as an end-to-end Python pipeline that leverages pre-trained models such as BERT, SentenceTransformer, and Qwen. This approach operates on clusterized datasets, allowing for nuanced evaluations at both a global level and within specific clusters.

Iterative Augmentation with Dynamic Thresholding

The augmentation process in BoostEDA is guided by a comprehensive array of metrics. These metrics inform iterative operations—such as swapping, deleting, inserting, and replacing text segments—and are further refined through dynamic thresholding based on cluster centroids to maintain semantic consistency.

Dimensionality Reduction and Comparative Analysis

To manage dataset size and uncover underlying data relationships, dimensionality reduction techniques like UMAP and PCA are employed. Comparative experiments, which included

clustering methods such as HDBSCAN and a facility location greedy algorithm, demonstrated the delicate balance required between preserving semantic fidelity and achieving lexical diversity.

Experiments Results

The research undertakes an assessment of text augmentation methods using a paired query dataset, examining how well the generated texts maintain the semantic intent and fluency of the originals while introducing new lexical variations. The methodology integrates several quantitative measures, including lexical overlap (Jaccard similarity), structural consistency (BPRO), fluency indicators from language modeling (LM_metric, Perplexity, and SLOR), and semantic alignment via cosine measures (Centroid Similarity), with a fixed diversity parameter represented by the Unique Trigram Ratio. This multi-dimensional framework enables a fine-grained evaluation of augmented outputs.

Assessment Methodology

The evaluation framework applies a combination of quantitative measures to capture multiple aspects of the augmented text. Metrics such as lexical overlap, structural consistency, and semantic alignment work together to form a comprehensive picture of augmentation performance, ensuring that both diversity and fluency are rigorously accounted for.

Observations and Findings

The findings reveal a complex landscape where some transformations, exemplified by instances with high structural and semantic fidelity, succeed in preserving meaning despite significant modifications. In contrast, other cases exhibit notable semantic drift, high prediction uncertainty, and reduced readability due to problematic reordering and awkward phrasing. These observations highlight the need to balance quantitative metrics with qualitative judgment to fully assess coherence and clarity.

Practical Challenges

The experiments underscore several practical challenges, including substantial computational demands. Processing 406 augmentation instances over several hours on advanced hardware illustrates that, while the rule-based framework is informative, there exists significant room for improvement through more efficient, hybrid approaches.

Next Steps

Looking forward, the study outlines several future directions to address current limitations. Enhancements may involve integrating large language models with generative adversarial strategies to refine grammatical and semantic patterns while reducing resource demands. Additionally, further research will seek to develop a more integrated system that combines

automated metrics with human evaluative judgment, aiming to effectively mitigate issues such as semantic drift and unnatural phrasing.