**TEXT AUGMENTATION LITERATURE REPORT**

Joshua S Raju | 22BAI1213 | Augmentation

**Report Date:** 11.03.24

**EDA: Easy Data Augmentation**

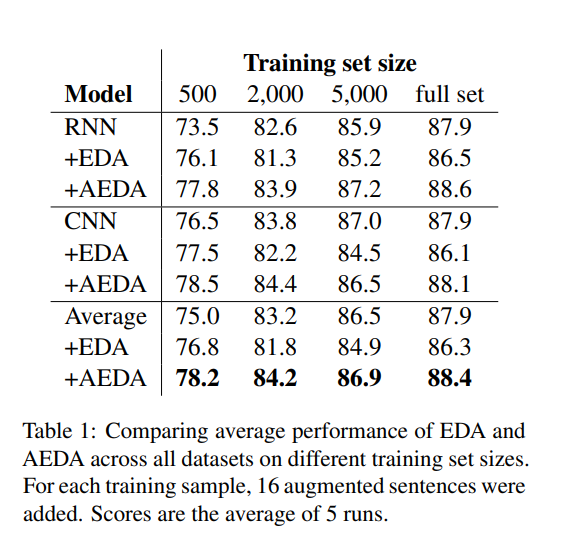
***Key Points:***

* Paper proposes 4 simple random operations: insertion, deletion, swap and synonym replacement.
* Demonstrated strong results, especially on smaller datasets. Avg. accuracy w/o augmentation noted was 88.3% at using 100% of the training dataset. Avg. accuracy of 88.6% was noted for EDA models using 50% of the same dataset.
* Little is given about if EDA conserves the true labels. But according to the paper ‘for the most part, sentences augmented with EDA conserved the labels of their original sentences.’
* However, on large and complete datasets, the performance gain is marginal (less than 1%).
* Doesn’t use LM (language model) or ext. datasets
* Besides synonym replacement, other 3 EDA operations haven’t been fully explored in depth yet. However, upon conducting tests, paper has concluded that all 4 operations contribute significantly.

**AEDA: An Easier Data Augmentation**

***Key Points:***

* While EDA proposes 4 different random operations, AEDA proposes to use only random insertion of punctuation marks {".", ";", "?", ":", "!", ","}.
* Punctuation marks maintains order, but changes positions of the words; claimed to leading to better generalized performance.
* EDA’s deletion operation leads to loss of info, while AEDA preserves all input info.
* No. of insertions = A number randomly is chosen between 1 and 1/3rd of sequence length. This ensures that at least, or not too many punctuation marks are inserted which can lead to noise.
* The position where the mark is inserted is also decided randomly.
* While EDA wasn’t boosting performance on large datasets, AEDA was consistent in boosting the performance in all various dataset size; outperforming EDA.



* An absolute improvement of 1.5-2.5% is noted for all dataset sizes by a single augmentation. Increasing the number of augmentations created a significant impact only on the smaller datasets, but remained stagnant for the full datasets (<1% gain).

**CoSDA-ML: Multi-Lingual Code-Switching Data Augmentation for Zero-Shot Cross-Lingual NLP**

***Key Points***