Using a Knowledge-based Approach for Data Augmentation

Guilherme Sales[1] and Daniel Brasil Soares[2]

[1]*Instituto Atlântico, Brazil,* guilherme_sales@atlantico.com.br
[2]*Instituto Atlântico, Brazil,* daniel_brasil@atlantico.com.br

Keywords: Text Data Augmentation, Boosting Performance, Computational Grammars, Methods

Despite the stunning achievements of data-driven models on many areas of Natural Language Processing (NLP), there is still a problem when it is necessary to use machine learning techniques with a small data set, this combination normally generates inaccurate models because of the high overfitting. The state-of-the-art solution to this problem has an automated treatment based on random data augmentation systems, such as Easy Data Augmentation (EDA)[1] that uses 4 operations: Synonym Replacement, Random Insertion, Random Swap, and Random Deletion. These methods can in fact improve the performance of the models by expanding the dataset with synthetic sentences, but they can also introduce new problems as cited in [2], falling shorts in producing semantically equivalent texts and sentences acceptable to native speakers. Another way to augment datasets is by using the back-translation technique, that proceeds by translating a sentence from a language x to a language y, and then, translating the sentence back to the language x, possibly resulting in new sentences. This method can generate new sentences but given the inherently nuanced and subjective nature of language, it will often produce very literal and unnatural expressions if it is based on a probabilistic model only. Aiming at a solution to the problem of overfitting in small datasets, using synthetic text data generated from automatic transformations, but preserving the semantic equivalence to the original sentences and their acceptance by natives, we investigated if using methods similar to that used in EDA and back-translation, but guided by knowledge, can help to produce more accurate models or even models that can better categorize, recognize or generate new acceptable sentences. For this investigation, we used computational grammars in the Grammatical Framework (GF) formalism [3] as an alternative to traditional methods of textual data augmentation (TDA). We modeled English and Portuguese linguistic phenomena to generate a controlled corpus of sentences from a dataset of 50 sentences. This modeling included linguistic treatment and manipulations, such as identification of synonymy relationships between words, insertion and deletion of adjuncts, displacement of phrases in the sentence, etc., all based on the knowledge linguists possess of these languages. As a result, we have shown that with the use of computational grammars, the dataset grows quickly without losing data quality.

Bibliography

[1] Jason Wei and Kai Zou. *EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. Association for Computational Linguistics, China, pages 6382—6388, 2019.*

[2] Gan, Wee and Ng, Hwee. *Improving the Robustness of Question Answering Systems to Question Paraphrasing* 01, 2019. pages: 6065-6075

[3] Ranta, Aarne. *Grammatical Framework: Programming with Multilingual Grammars* CSLI Publications, Stanford, 01, 2011.