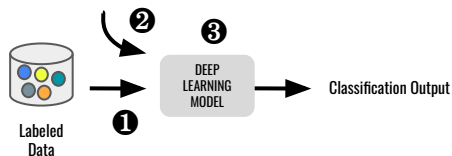


Data Augmentation using Lexicalized PCFGs for Few-Shot Text Classification

Hazel Kim¹, Deacheol Woo¹, Seong Joon Oh², Jeong-Won Cha³, Yo-Sub Han¹

¹Yonsei University, ²NAVER AI Lab, ³Changwon National University



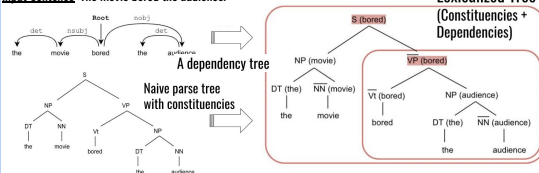
For Higher accuracy, prior work

1. Randomly swaps, inserts, deletes any words or replaces arbitrary synonyms. Translates data to and from a pivot language using pre-trained LMs. Uses masked LMs to generate corrupted and reconstructed data points.
2. Use unlabeled data to inject noise for consistency training.
3. Manipulate loss function to consider noise to generalize the model.



Lexicalized Probabilistic Context Free Grammars

Input Sentence: The movie bored the audience.



Data Augmentation using L-PCFGs (ALP)

For limited data learning, compositionality matters.

Lexical Knowledge

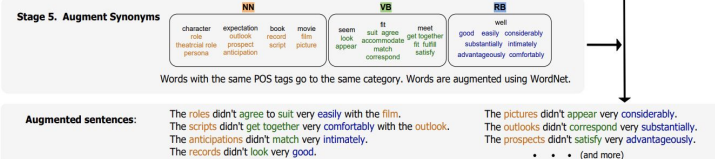
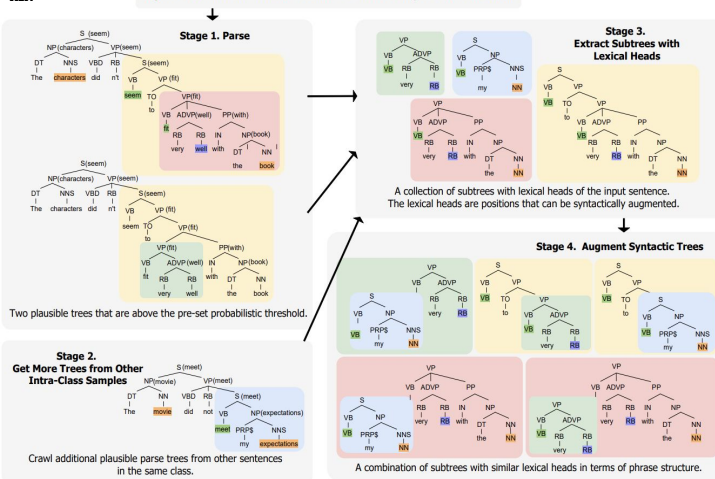
Information that is known about words.

Syntactic Structures

The rules whereby words are combined to form grammatical sentences.

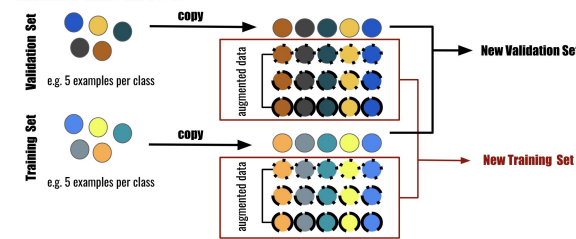
Overview of ALP.

Input sentence: The characters didn't seem to fit very well with the book.



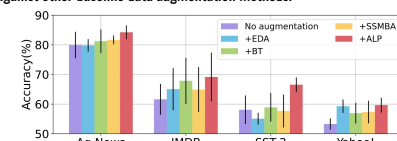
Train-Validation Splitting (AugTrain-Train)

Within the k labeled data, the available labeled data are $2k$



Empirical Studies

1. Comparison against other baseline data augmentation methods.



2. Compatibility with various semi-supervised learning approaches.

Methods	5	10	5	10	5	10	5	10
UST	79.65	83.85	57.13	62.71	63.60	73.63	55.49	63.54
+ BT	81.61	83.43	57.22	67.76	67.14	83.21	61.78	63.91
+ SSMBA	83.05	86.32	48.76	57.00	61.05	66.82	62.81	63.63
+ ALP	84.72	87.41	73.22	78.01	71.32	76.33	61.20	66.89
MixText	81.14	87.11	51.46	50.91	68.09	72.87	66.60	67.40
+ BT	82.04	70.18	51.29	51.78	62.44	74.12	66.19	66.00
+ SSMBA	83.47	69.03	51.95	52.39	54.11	61.36	65.32	67.14
+ ALP	83.50	87.72	52.44	57.06	84.35	84.16	67.31	67.81

3. AugTrain-Train performance gain with ALP as well as other baseline data augmentation methods.

k	Method	No-aug	EDA	BT	SSMBA	ALP
5	AugTrain-Val	58.66	62.73	63.59	63.65	66.45
5	AugTrain-Train	62.28	64.42	67.40	68.09	72.40
10	AugTrain-Val	67.81	66.90	68.90	67.15	72.83
10	AugTrain-Train	70.97	67.31	73.15	68.91	74.89