

Less Data, More ____?

Data Augmentation and Semi-Supervised Learning for Natural Language Processing

Diyi Yang, Georgia Tech

Ankur P. Parikh, Google Research

Colin Raffel, University of North Carolina, Chapel Hill

“I have an extremely large
collection of clean labeled data”

- No one

Learning from limited labeled data

- Transfer learning
 - Leverage data from a different-but-related task
- Few/zero-shot learning
 - Generalize to new tasks after seeing a few (or no) examples of that task
- Multitask learning
 - Use information learned on different tasks for mutual benefit
- Data augmentation
 - Modify labeled data to with class-preserving transformations
- Semi-supervised learning
 - Learn from labeled and unlabeled data

Data Augmentation

- Token-level augmentation
 - Change individual words
- Sentence-level augmentation
 - Change an entire sentence
- Adversarial augmentation:
 - Change the text to maximally fool the model
- Hidden space augmentation:
 - Change the representations inside the model

Data Augmentation

1. Token-level augmentation:

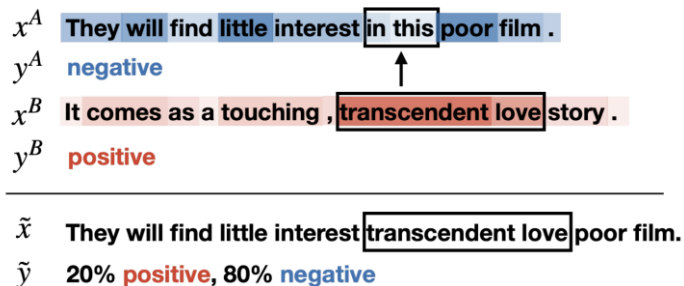
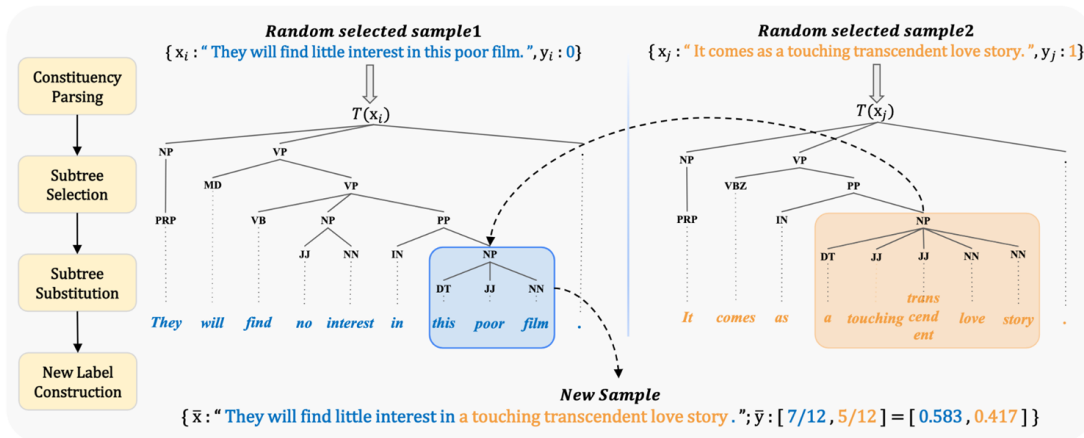
- **Synonym replacement** (Yang et al. 2015, Zhang et al. 2015, Miao et al. 2020)
- **Random insertion, deletion, swapping** (Xie et al. 2019, Wei and Zou 2019)
- **Word replacement via LM** (Wu et al. 2019, Zhu et al. 2019)

Easy Data Augmentation Techniques (EDA)

Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
Synonym replacement	A lamentable , superior human comedy played out on the backward roads of life.
Random insertion	A sad, superior human comedy played out on funniness the back roads of life.
Random swap	A sad, superior human comedy played out on roads back the of life.
Random deletion	A sad, superior human out on the roads of life.

Wei, Jason, and Kai Zou. "EDA: Easy data augmentation techniques for boosting performance on text classification tasks." arXiv preprint arXiv:1901.11196 (2019).

Compositional Augmentation



Saliency based data augmentation where the least salient span from sent A is replaced with the most salient span from sent B (Yoon et al., 2021)

TreeMix: Compositional Constituency-based Data Augmentation for Natural Language Understanding (Zhang et al., 2022)

Token Level Data Augmentation Summary

Methods	Types	News Classification		Topic Classification	
		AG News	20 Newsgroup	Yahoo Answers	PubMed
None	-	78.8(8.9)	65.2(4.8)	56.6(9.4)	63.7(6.1)/49.3(3.9)
SR	Token	79.4(5.9)	66.1(2.5)	56.0(10.1)	62.4(5.7)/48.3(3.9)
LM		76.8(5.1)	60.0(14.4)	56.2(8.4)	60.9(3.0)/47.4(2.5)
RI		79.5(4.9)	66.6(0.6)	57.3(12.0)	63.7(4.2)/49.4(2.1)
RD		79.6(5.0)	66.8(3.0)	58.0(8.3)	63.4(5.0)/49.3(1.5)
RS		79.5(5.3)	64.8(10.8)	57.1(10.3)	63.8(7.4)/49.5(3.3)
WR		79.7(2.0)	67.5(4.2)	59.3(8.9)	64.9(4.9)/49.4(2.5)

Topic Classification and News Classification results with 10 examples. We report the average results across 3 different random seeds with the 95% confidence interval and bold the best results.

Token Level Data Augmentation Summary

Methods	Level	Diversity	Tasks
Synonym replacement	Token	Low	Text classification, Sequence labeling
Random insertion, deletion, swapping	Token	Medium	Text classification, Sequence labeling , Machine translation, Dialogue generation
Word replacement via LM	Token	Low	Text classification, Sequence labeling , Machine translation
Compositional augmentation	Token	High	Text classification, Sequence labeling , Semantic Parsing, Language Modeling, Text Generation

Data Augmentation

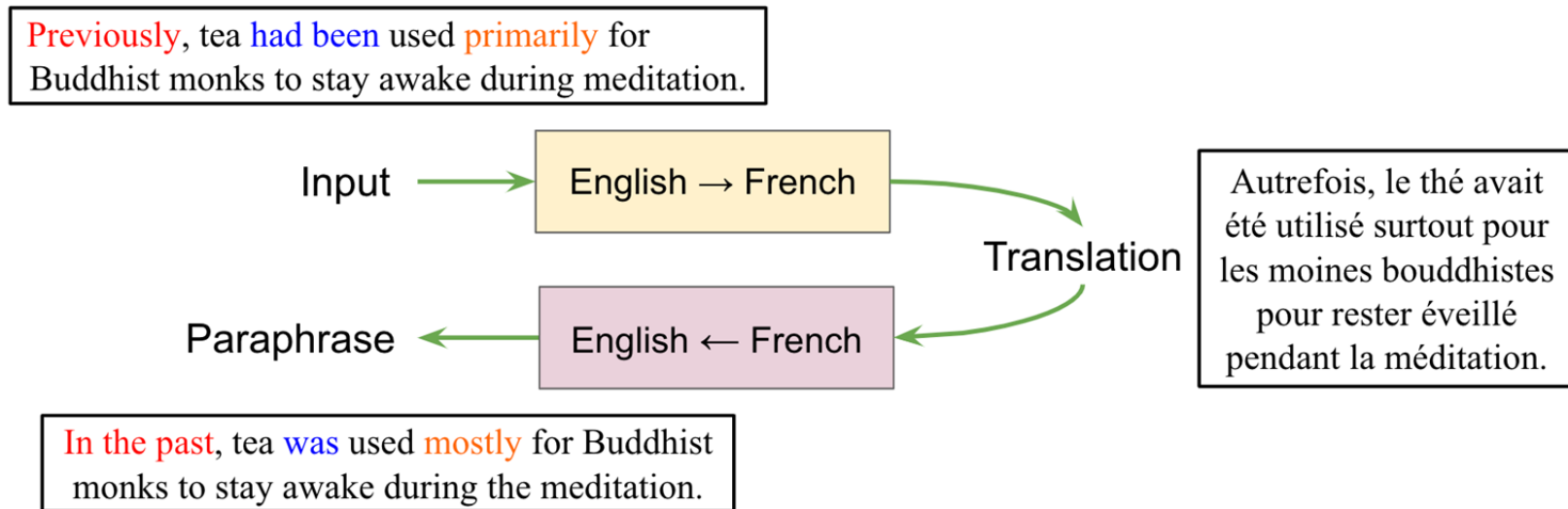
1. Token-level augmentation:

- Synonym replacement (Yang et al. 2015, Zhang et al. 2015, Miao et al. 2020)
- Random insertion, deletion, swapping (Xie et al. 2019, Wei and Zou 2019)
- Word replacement via LM (Wu et al. 2019, Zhu et al. 2019)

2. Sentence-level augmentation:

- Paraphrasing (Xie et al. 2019, Chen et al. 2020)
- Conditional generation (Zhang and Bansal 2019, Yang et al. 2020)

Back-Translation for Data Augmentation (Edunov et al., 2018)



Paraphrasing

Madnani, Nitin, and Bonnie J. Dorr. "Generating phrasal and sentential paraphrases: A survey of data-driven methods." Computational Linguistics 36, no. 3 (2010): 341-387.

template	paraphrase
original	with the help of captain picard , the borg will be prepared for everything .
(SBARQ (ADVP) (,) (S) (,) (SQ))	now , the borg will be prepared by picard , will it ?
(S (NP) (ADVP) (VP))	the borg here will be prepared for everything .
(S (S) (,) (CC) (S) (:) (FRAG))	with the help of captain picard , the borg will be prepared , and the borg will be prepared for everything ... for everything .
(FRAG (INTJ) (,) (S) (,) (NP))	oh , come on captain picard , the borg line for everything .
original	you seem to be an excellent burglar when the time comes .
(S (SBAR) (,) (NP) (VP))	when the time comes , you 'll be a great thief .
(S (`` ) ( UCP ) ( ``) (NP) (VP))	“ you seem to be a great burglar , when the time comes . ” you said .
(SQ (MD) (SBARQ))	can i get a good burglar when the time comes ?
(S (NP) (IN) (NP) (NP) (VP))	look at the time the thief comes .

syntactically controlled paraphrase generation (Iyyer et al., 2018)

Data Augmentation

1. Token-level augmentation:

- Synonym replacement (Yang et al. 2015, Zhang et al. 2015, Miao et al. 2020)
- Random insertion, deletion, swapping (Xie et al. 2019, Wei and Zou 2019)
- Word replacement via LM (Wu et al. 2019, Zhu et al. 2019)

2. Sentence-level augmentation:

- Paraphrasing (Xie et al. 2019, Chen et al. 2020)
- Conditional generation (Zhang and Bansal 2019, Yang et al. 2020)

3. Adversarial augmentation:

- Whitebox methods (Miyato et al., 2017; Zhu et al., 2020; Jiang et al., 2019; Chen et al., 2020d)
- Blackbox methods (Ren et al. 2019; Garg and Ramakrishnan, 2020)

White-box Attack

HotFlip uses the model gradient to identify the most important letter in the text (Ebrahimi et al., 2018)

$$\max \nabla_x J(\mathbf{x}, \mathbf{y})^T \cdot \vec{v}_{ijb} = \max_{ijb} \frac{\partial J^{(b)}}{\partial x_{ij}} - \frac{\partial J^{(a)}}{\partial x_{ij}}$$

Loss of the
model on input
x with label y

Flip vector: flip of
the j-th character of the
i-th word (a → b)

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism.
57% **World**

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a moo**P** of optimism.
95% **Sci/Tech**

Chancellor Gordon Brown has sought to quell speculation over who should run the Labour Party and turned the attack on the opposition Conservatives.
75% **World**

Chancellor Gordon Brown has sought to quell speculation over who should run the Labour Party and turned the attack on the o**B**position Conservatives.
94% **Business**

Adversarial examples with a single character change, which will be misclassified by a neural classifier.

White-box Attack

HotFlip uses the model gradient to identify the most important letter in the text (Ebrahimi et al., 2018)

$$\max \nabla_x J(\mathbf{x}, \mathbf{y})^T \cdot \vec{v}_{ijb} = \max_{ijb} \frac{\partial J^{(b)}}{\partial x_{ij}} - \frac{\partial J^{(a)}}{\partial x_{ij}}$$

Find the flip vector with biggest increase in loss

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a mood of optimism.
57% **World**

South Africa's historic Soweto township marks its 100th birthday on Tuesday in a moo**P** of optimism.
95% **Sci/Tech**

Chancellor Gordon Brown has sought to quell speculation over who should run the Labour Party and turned the attack on the opposition Conservatives.
75% **World**

Chancellor Gordon Brown has sought to quell speculation over who should run the Labour Party and turned the attack on the o**B**position Conservatives.
94% **Business**

Adversarial examples with a single character change, which will be misclassified by a neural classifier.

Black-box Attack

ORIGINAL

The government made a quick decision

BAE - R 

The **MASK** made a quick decision

judge , doctor , captain

BAE - I 

The **MASK** government made a quick decision

state , british , federal

The government **MASK** made a quick decision

officials , then , immediately

40-80% accuracy drop!

Model	Adversarial Attack	Datasets			
		Amazon	Yelp	IMDB	MR
wordLSTM	Original	88.0	85.0	82.0	81.16
	TextFooler	31.0 (0.747)	28.0 (0.829)	20.0 (0.828)	25.49 (0.906)
	BAE-R	21.0 (0.827)	20.0 (0.885)	22.0 (0.852)	24.17 (0.914)
	BAE-I	17.0 (0.924)	22.0 (0.928)	23.0 (0.933)	19.11 (0.966)
	BAE-R / I	16.0 (0.902)	19.0 (0.924)	8.0 (0.896)	15.08 (0.949)
	BAE-R+I	4.0 (0.848)	9.0 (0.902)	5.0 (0.871)	7.50 (0.935)
wordCNN	Original	82.0	85.0	81.0	76.66
	TextFooler	42.0 (0.776)	36.0 (0.827)	31.0 (0.854)	21.18 (0.910)
	BAE-R	16.0 (0.821)	23.0 (0.846)	23.0 (0.856)	20.81 (0.920)
	BAE-I	18.0 (0.934)	26.0 (0.941)	29.0 (0.924)	19.49 (0.971)
	BAE-R / I	13.0 (0.904)	17.0 (0.916)	20.0 (0.892)	15.56 (0.956)
	BAE-R+I	2.0 (0.859)	9.0 (0.891)	14.0 (0.861)	7.87 (0.938)
BERT	Original	96.0	95.0	85.0	85.28
	TextFooler	30.0 (0.787)	27.0 (0.833)	32.0 (0.877)	30.74 (0.902)
	BAE-R	36.0 (0.772)	31.0 (0.856)	46.0 (0.835)	44.05 (0.871)
	BAE-I	20.0 (0.922)	25.0 (0.936)	31.0 (0.929)	32.05 (0.958)
	BAE-R / I	11.0 (0.899)	16.0 (0.916)	22.0 (0.909)	20.34 (0.941)
	BAE-R+I	14.0 (0.830)	12.0 (0.871)	16.0 (0.856)	19.21 (0.917)

Use BERT-MLM to predict masked tokens in the text for generating adversarial examples.
(Garg and Ramakrishnan, 2020)

Adversarial Attack Augmentation Summary

Methods	Level	Diversity	Tasks
White-box attack	Token	Low	Text classification, Sequence labeling, Machine translation
Black-box attack	Token	Medium	Text classification, Sequence labeling, Machine translation, Textual entailment, Dialogue generation, Text Summarization