T5

22ND – 27TH MAY | 60TH MEETING | DUBLIN

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).

Word Replacement via Language Modeling

	IWSLT			WMT
	De → En	Es → En	He → En	En → De
Base	34.79	41.58	33.64	28.40
$+LM_{sample}$	35.40	42.09	34.31	28.73
Ours	**35.78**	**42.61**	**34.91**	**29.70**

Contextual data augmentation:

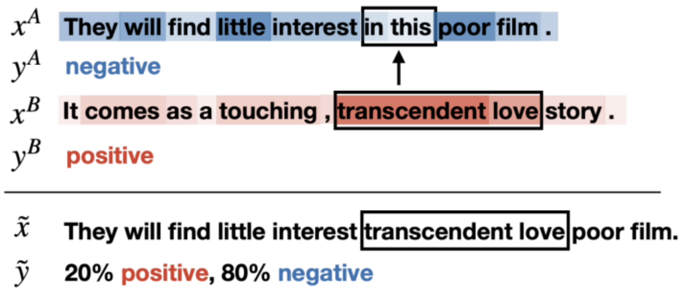
when a sentence "*the actors are fantastic*" is augmented by replacing only *actors* with words predicted based on the context (Kobayashi, 2018)

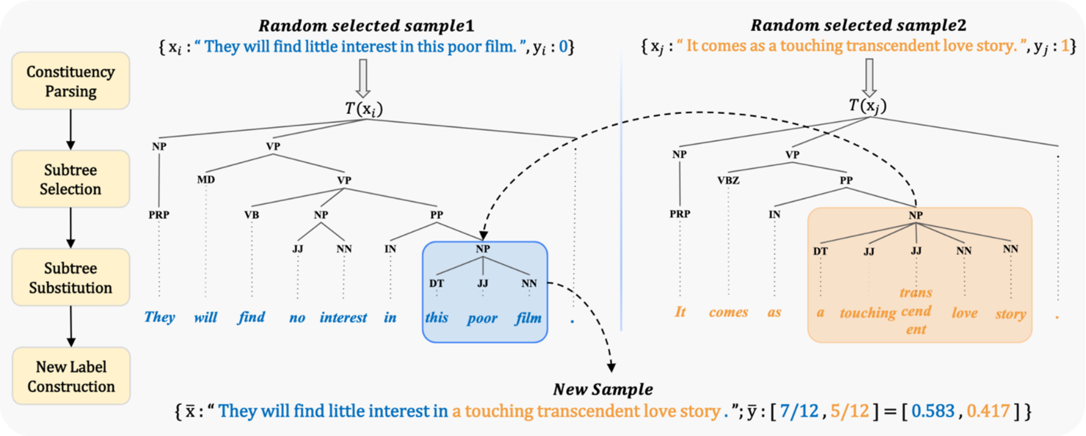
Soft contextual data augmentation

(Gao et al., 2019)

$$e_w = P(w)E = \sum_{j=0}^{|V|} p_j(w)E_j$$

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	Token	76.8(5.1)	60.0(14.4)	56.2(8.4)	60.9(3.0)/47.4(2.5)
RI	Token	79.5(4.9)	66.6(0.6)	57.3(12.0)	63.7(4.2)/49.4(2.1)
RD	Token	79.6(5.0)	66.8(3.0)	58.0(8.3)	63.4(5.0)/49.3(1.5)
RS	Token	79.5(5.3)	64.8(10.8)	57.1(10.3)	63.8(7.4)/49.5(3.3)
WR	Token	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)

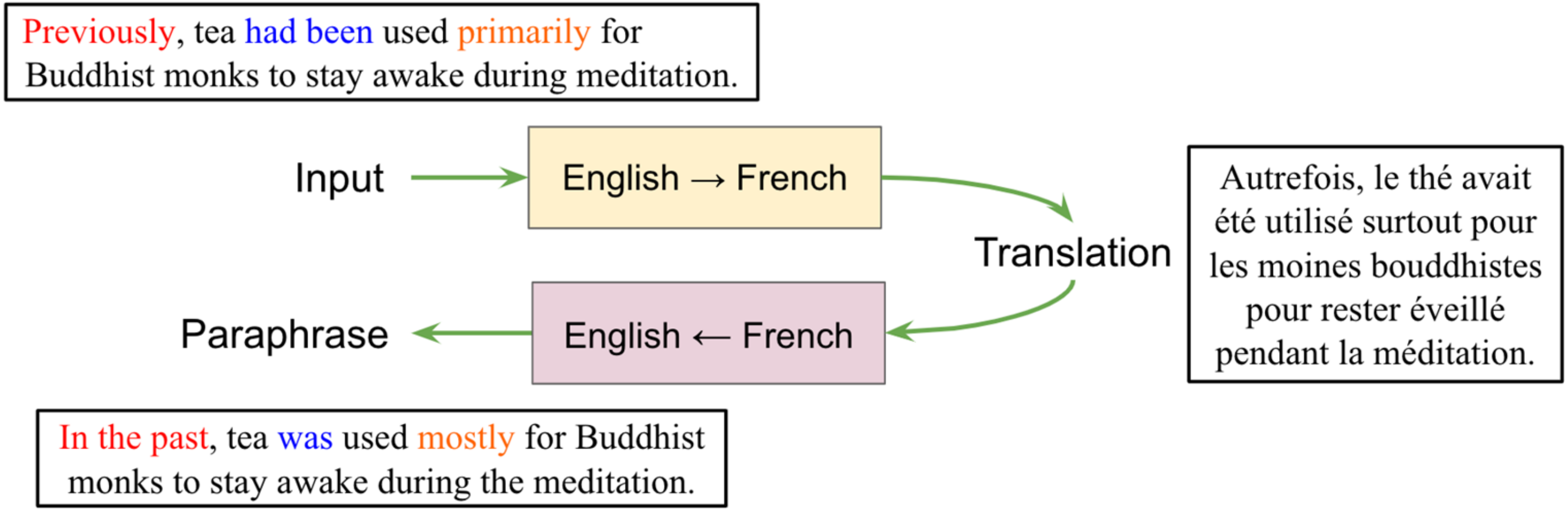


Image credit to https://github.com/vietai/dab

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.

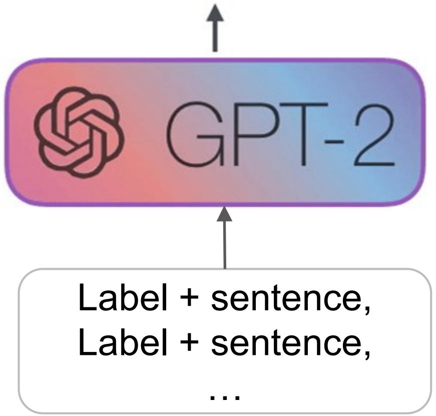
template	**paraphrase**
original	with the help of captain picard , the borg will be prepared for everything .
(SBARQ(ADVP)(,)(S)(,)(SQ))	now , the borg will be prepared by picard , will it ?
(S(NP)(ADVP)(VP))	the borg here will be prepared for everything .
(S(S)(,)(CC)(S) (:)(FRAG))	with the help of 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)

Conditional Generation

Language model based data augmentation (LAMBADA) using GPT (Anaby-Tavor et al., 2019)

Class label	Sentences
Flight time	what time is the last flight from san francisco to washington dc on continental
Aircraft	show me all the types of aircraft used flying from atl to dallas
City	show me the cities served by canadian airlines



Sentence Level 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	Token	76.8(5.1)	60.0(14.4)	56.2(8.4)	60.9(3.0)/47.4(2.5)
RI	Token	79.5(4.9)	66.6(0.6)	57.3(12.0)	63.7(4.2)/49.4(2.1)
RD	Token	79.6(5.0)	66.8(3.0)	58.0(8.3)	63.4(5.0)/49.3(1.5)
RS	Token	79.5(5.3)	64.8(10.8)	57.1(10.3)	63.8(7.4)/49.5(3.3)
WR	Token	79.7(2.0)	**67.5(4.2)**	**59.3(8.9)**	**64.9(4.9)/49.4(2.5)**
RT	Sentence	**80.1(4.3)**	65.1(7.9)	57.1(9.6)	60.2(5.1)/46.3(6.4)

Methods	Diversity	Tasks
Paraphrase	High	Text classification, Machine translation, Question answering, Generation
Conditional Generation	High	Text classification, Question answering

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)

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

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)

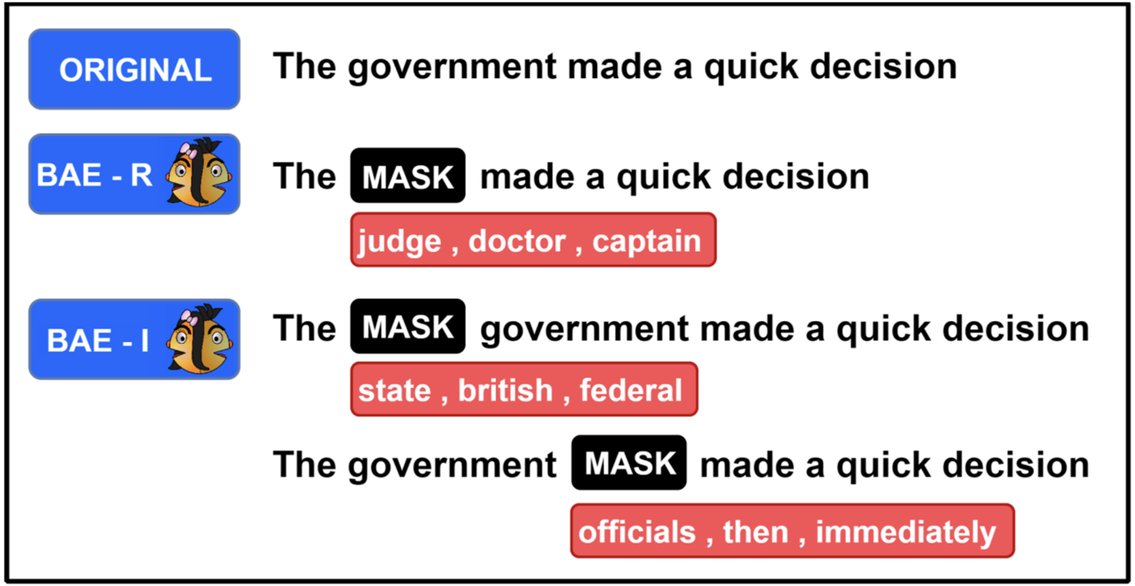
$$\max \nabla_x J(\mathbf{x}, \mathbf{y})^T \cdot \vec{v}_{ijb} = \max_{ijb} \frac{\partial J}{\partial x_{ij}}^{(b)} - \frac{\partial J}{\partial x_{ij}}^{(a)}$$

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



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

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)
4. Hidden space augmentation:
    - Mixup (Zhang et al., 2019, Chen et al. 2020)

Hidden-space Augmentation via Perturbation

Manipulating the hidden representations

- Through perturbations such as adding noises
- Or performing interpolations with other data points

Interpolation: **mixup** for text data

A Generalized View of Text Mixup: linguistically informed interpolations

label for x

$$\tilde{y} = mix(y, y') = \lambda\, y \oplus (1 - \lambda)\, y'$$
$$\tilde{x} = mix(x, x') = \lambda\, x \oplus (1 - \lambda)\, x'$$

label for x' (e.g., sentiment, named entity)

another sampled data point from the corpus (e.g., sentence, token, subtree)

a given data point (e.g., sentence, token, subtree)

Mix parameter (e.g., sampled from a Beta distribution $\lambda \sim \mathrm{Beta}(\alpha, \alpha)$)

Interpolation: **mixup** for text data

A Generalized View of Text Mixup

$$\tilde{y} = mix(y, y') = \lambda\, y \oplus (1-\lambda)\, y'$$
$$\tilde{x} = mix(x, x') = \lambda\, x \oplus (1-\lambda)\, x'$$

feature

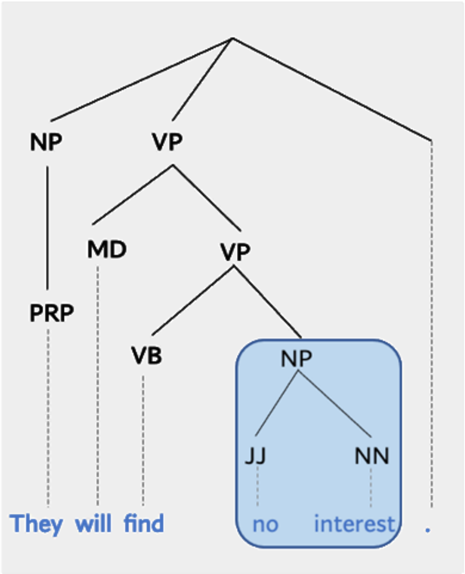
feature

discrete

A sentence:
"I like this movie"

A token:
Pfizer-BioNTech ORG or COVID-19 DISEASE

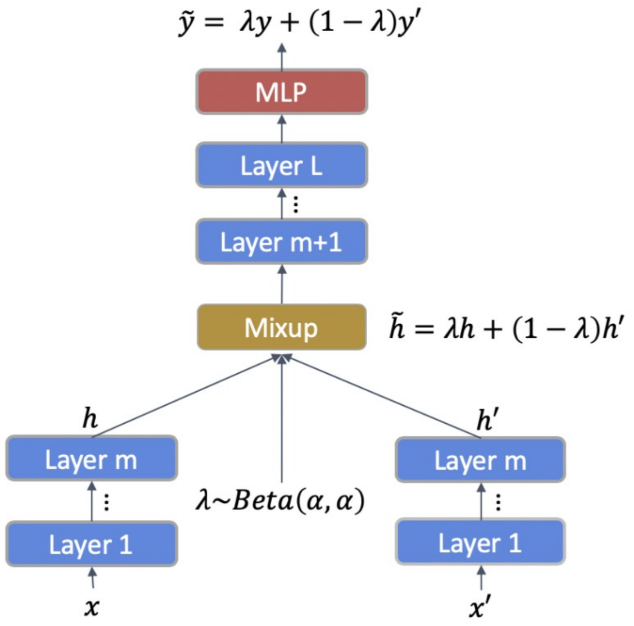
A subtree:



Interpolation: **mixup** in textual hidden space

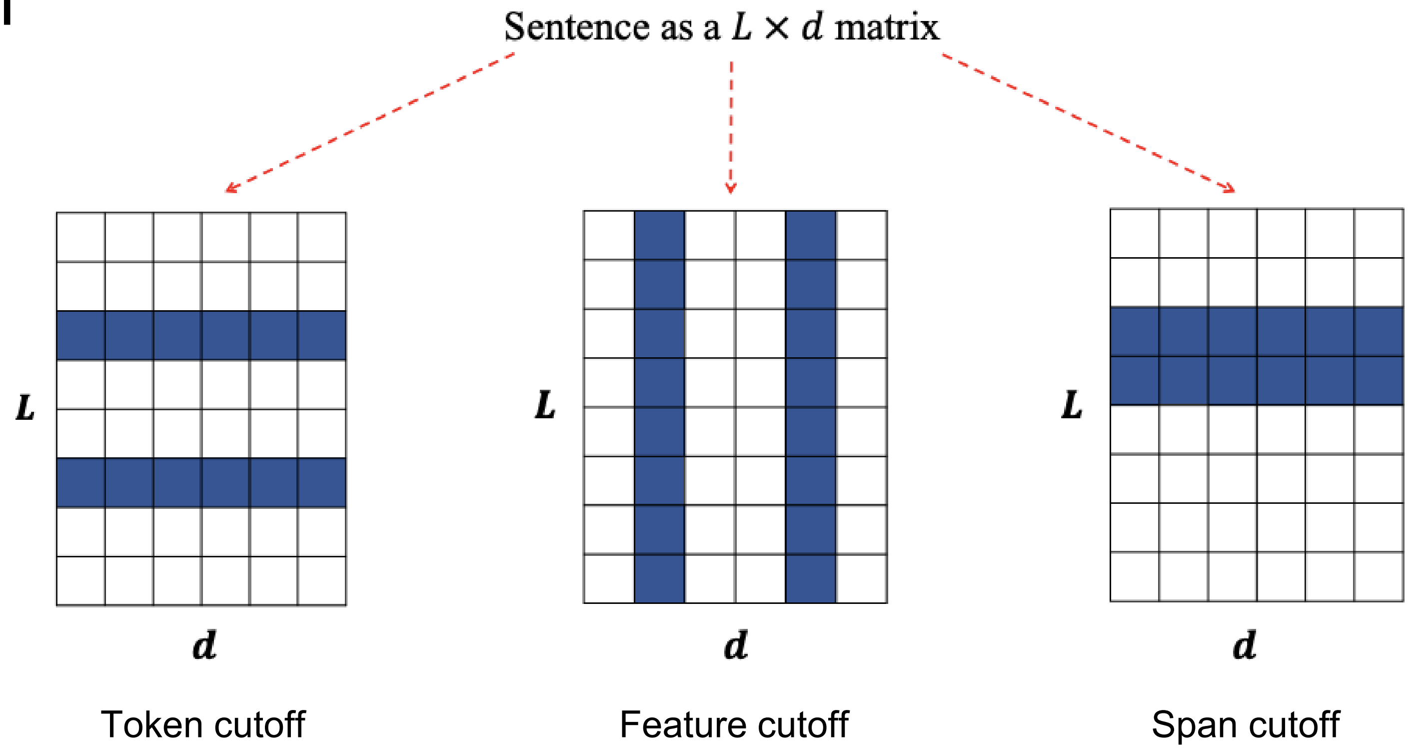
$$\tilde{\mathbf{x}} = \text{mix}(\mathbf{x}_i, \mathbf{x}_j) = \lambda \mathbf{x}_i + (1 - \lambda)\mathbf{x}_j$$
$$\tilde{\mathbf{y}} = \text{mix}(\mathbf{y}_i, \mathbf{y}_j) = \lambda \mathbf{y}_i + (1 - \lambda)\mathbf{y}_j$$

$$\lambda \sim \text{Beta}(\alpha, \alpha)$$



Chen, Jiaao, Zichao Yang, and Diyi Yang. "MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification." ACL 2020.

Cutoff



Shen, Dinghan, Mingzhi Zheng, Yelong Shen, Yanru Qu, and Weizhu Chen. "A simple but tough-to-beat data augmentation approach for natural language understanding and generation." arXiv preprint arXiv:2009.13818 (2020).

Cutoff

Closely related to **multi-view learnir**

Can be apply to both text classification a

Model	BLEU score
Actor-critic (Bahdanau et al., 2016)	28.5
Transformer Base (Vaswani et al., 2017)	34.4
Adversarial training (Wang et al., 2019)	35.2
Data Diversification (Nguyen et al., 2019)	37.2
MAT (Fan et al., 2020)	36.2
Mixed Representations (Wu et al., 2020)	36.4
MAT+Knee (Iyer et al., 2020)	36.6
Transformer Base & Cutoff (w/o JS loss)	36.7
Transformer Base & Cutoff (w/ JS loss)	**37.6**

$$\mathcal{L} = \mathcal{L}_{\text{ce}}(x, y) + \alpha \sum_{i=1}^{N} \mathcal{L}_{\text{ce}}(x^{i}_{\text{cutoff}}, y)$$

$$+ \ \beta \mathcal{L}_{\text{divergence}}(x, x^{1}_{\text{cutoff}}, x^{2}_{\text{cutoff}}, \ldots, x^{N}_{\text{cutoff}}, y)$$

Hidden Space Augmentation Summary

Methods	Level	Diversity	Tasks
Hidden-space perturbation	Token or Sentence	High	Text classification, Sequence labeling, Speech recognition
Interpolation	Token or Sentence	High	Text classification, Sequence labeling, Machine translation

Hidden Space Augmentation Summary

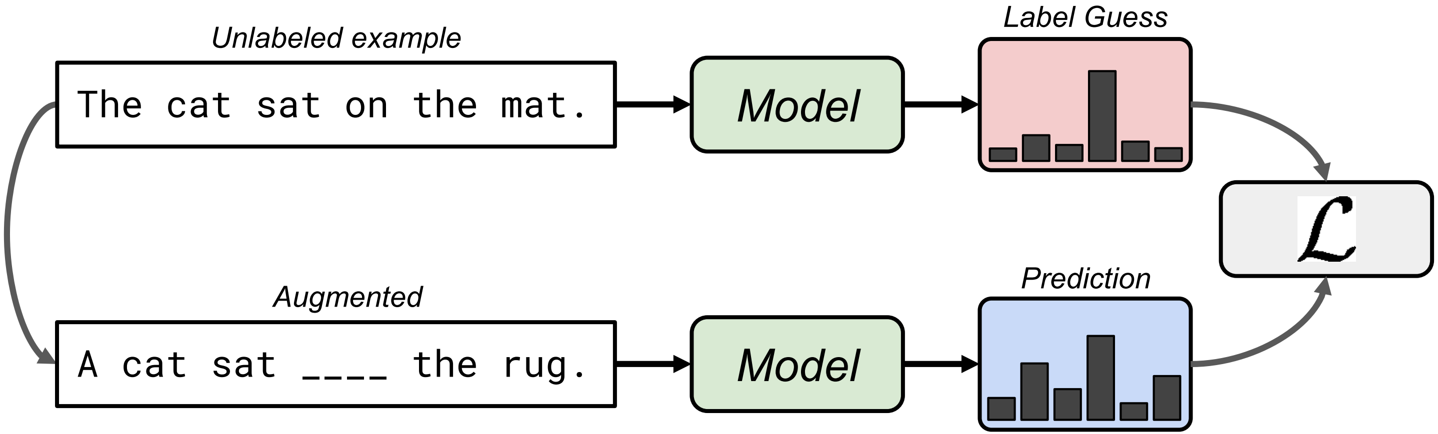
Methods	Types	News Classification		Topic Classification	
		AG News	20 Newsgroup	Yahoo Answers	PubMed
None	-	78.8(8.9)	65.2(4.8)	56.6(9.4)	63.7(6.1)/49.3(3.9)
SR	Token	79.4(5.9)	66.1(2.5)	56.0(10.1)	62.4(5.7)/48.3(3.9)
LM		76.8(5.1)	60.0(14.4)	56.2(8.4)	60.9(3.0)/47.4(2.5)
RI		79.5(4.9)	66.6(0.6)	57.3(12.0)	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)**
RT	Sentence	**80.1(4.3)**	65.1(7.9)	57.1(9.6)	60.2(5.1)/46.3(6.4)
ADV	Hidden	78.2 (5.3)	65.5(1.6)	53.8(4.89)	37.4(2.6)/19.9(10.6)
Cutoff		79.3(5.0)	66.6(1.4)	57.3(9.3)	60.5(8.3)/46.6(9.4)
Mixup		80.0 (6.52)	65.9(3.1)	57.8(4.19)	51.4(19.3)/39.8(3.2)

Hidden Space Augmentation Summary

	Methods	Types	Inference			Paraphrase		Single Sentence	
			MNLI	QNLI	RTE	QQP	MRPC	SST-2	CoLA
Supervised	None								
	SR								
	LM								
	RI								
	RD								
	RS								
	WR								
	RT								
	ADV		33.3(4.7)	49.7(1.8)	48.3(12.1)	57.5(24.7)	61.5(21.5)	53.3(13.07)	1.37(4.66)
	Cutoff	Hidden	35.1(2.3)	51.4(8.3)	52.2(3.6)	62.6(8.8)	61.0(21.2)	**63.5(8.45)**	**12.4(9.58)**
	Mixup		32.6(3.5)	49.9(1.4)	49.8(9.2)	63.0(0.3)	62.1(19.8)	62.3(12.3)	4.03(8.68)

- No single augmentation works the best for every task.
- Augmentation does not always improve performance, and can sometimes hurt performances.
- Token-level augmentations work well in general for supervised learning, especially with limited labeled data

Consistency regularization



$$\mathbb{E}_x - \hat{p}_\theta(y|x) \log p_\theta(y|x)$$

$$\hat{p}_\theta(y|x) = p_\theta(y|x')$$

“Unsupervised Data Augmentation” (UDA)

Initialization	**UDA**	IMDb (20)	Yelp-2 (20)	Yelp-5 (2.5k)	Amazon-2 (20)	Amazon-5 (2.5k)	DBpedia (140)
Random	✗	43.27	40.25	50.80	45.39	55.70	41.14
	✓	25.23	8.33	41.35	16.16	44.19	7.24
$\text{BERT}_{\text{BASE}}$	✗	18.40	13.60	41.00	26.75	44.09	2.58
	✓	5.45	2.61	33.80	3.96	38.40	1.33
$\text{BERT}_{\text{LARGE}}$	✗	11.72	10.55	38.90	15.54	42.30	1.68
	✓	4.78	2.50	33.54	3.93	37.80	1.09

Xie, Qizhe, et al. "Unsupervised data augmentation for consistency training." NeurIPS 2020.

“Unsupervised Data Augmentation” (UDA)

Initialization	UDA	IMDb (20)	Yelp-2 (20)	Yelp-5 (2.5k)	Amazon-2 (20)	Amazon-5 (2.5k)	DBpedia (140)
Random	✗	43.27	40.25	50.80	45.39	55.70	41.14
	✓	25.23	8.33	41.35	16.16	44.19	7.24
$BERT_{BASE}$	✗	18.40	13.60	41.00	26.75	44.09	2.58
	✓	5.45	2.61	33.80	3.96	38.40	1.33
$BERT_{LARGE}$	✗	11.72	10.55	38.90	15.54	42.30	1.68
	✓	4.78	2.50	33.54	3.93	37.80	1.09

Pre-training helps

Xie, Qizhe, et al. "Unsupervised data augmentation for consistency training." NeurIPS 2020.

“Unsupervised Data Augmentation” (UDA)

Initialization	**UDA**	IMDb (20)	Yelp-2 (20)	Yelp-5 (2.5k)	Amazon-2 (20)	Amazon-5 (2.5k)	DBpedia (140)
Random	✗	43.27	40.25	50.80	45.39	55.70	41.14
	✓	25.23	8.33	41.35	16.16	44.19	7.24
$\text{BERT}_{\text{BASE}}$	✗	18.40	13.60	41.00	26.75	44.09	2.58
	✓	5.45	2.61	33.80	3.96	38.40	1.33
$\text{BERT}_{\text{LARGE}}$	✗	11.72	10.55	38.90	15.54	42.30	1.68
	✓	4.78	2.50	33.54	3.93	37.80	1.09

SSL is complementary

Xie, Qizhe, et al. "Unsupervised data augmentation for consistency training." NeurIPS 2020.

SSL or just augmentation?

	Methods	Types	News Classification		Topic Classification	
			AG News	20 Newsgroup	Yahoo Answers	PubMed
Supervised	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)**
	RT	Sentence	**80.1(4.3)**	65.1(7.9)	57.1(9.6)	60.2(5.1)/46.3(6.4)
	ADV	Hidden	78.2 (5.3)	65.5(1.6)	53.8(4.89)	37.4(2.6)/19.9(10.6)
	Cutoff		79.3(5.0)	66.6(1.4)	57.3(9.3)	60.5(8.3)/46.6(9.4)
	Mixup		80.0 (6.52)	65.9(3.1)	57.8(4.19)	51.4(19.3)/39.8(3.2)
Semi Supervised	SR	Token	69.6(29.3)	65.7(1.8)	51.4(9.4)	59.3(5.9)/43.1(11.9)
	LM		68.5(13.7)	68.3(2.1)	53.2(6.3)	61.5(6.6)/46.4(4.4)
	RI		65.8(5.5)	66.7(1.1)	50.5(3.2)	61.4(11.3)/44.4(17.4)
	RD		73.2(14.0)	66.1(3.3)	51.5(7.5)	59.3(7.1)/46.0(3.8)
	RS		71.6(16.6)	65.0(2.0)	51.1(7.1)	64.2(12.1)/46.7(11.5)
	WR		74.1(12.3)	**69.3(2.5)**	55.6(5.9)	60.4(7.5)/43.7(14.2)
	RT	Sentence	82.1(8.2)	68.8(2.4)	59.8(3.9)	**64.3(1.2)/49.8(1.9)**
	ADV	Hidden	**82.3(2.33)**	66.8(5.9)	55.9(3.89)	62.2(10.8)/46.2(9.8)
	Cutoff		79.9(5.5)	67.9(0.8)	**60.1(1.0)**	62.7(9.0)/48.1(3.2)

Augmentation alone helps

Chen, Jiaao, et al. "An empirical survey of data augmentation for limited data learning in NLP." arXiv preprint arXiv:2106.07499 (2021).

SSL or just augmentation?

	Methods	Types	News Classification		Topic Classification	
			AG News	20 Newsgroup	Yahoo Answers	PubMed
Supervised	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	Token	76.8(5.1)	60.0(14.4)	56.2(8.4)	60.9(3.0)/47.4(2.5)
	RI	Token	79.5(4.9)	66.6(0.6)	57.3(12.0)	63.7(4.2)/49.4(2.1)
	RD	Token	79.6(5.0)	66.8(3.0)	58.0(8.3)	63.4(5.0)/49.3(1.5)
	RS	Token	79.5(5.3)	64.8(10.8)	57.1(10.3)	63.8(7.4)/49.5(3.3)
	WR	Token	79.7(2.0)	**67.5(4.2)**	**59.3(8.9)**	**64.9(4.9)/49.4(2.5)**
	RT	Sentence	**80.1(4.3)**	65.1(7.9)	57.1(9.6)	60.2(5.1)/46.3(6.4)
	ADV	Hidden	78.2 (5.3)	65.5(1.6)	53.8(4.89)	37.4(2.6)/19.9(10.6)
	Cutoff	Hidden	79.3(5.0)	66.6(1.4)	57.3(9.3)	60.5(8.3)/46.6(9.4)
	Mixup	Hidden	80.0 (6.52)	65.9(3.1)	57.8(4.19)	51.4(19.3)/39.8(3.2)
Semi Supervised	SR	Token	69.6(29.3)	65.7(1.8)	51.4(9.4)	59.3(5.9)/43.1(11.9)
	LM	Token	68.5(13.7)	68.3(2.1)	53.2(6.3)	61.5(6.6)/46.4(4.4)
	RI	Token	65.8(5.5)	66.7(1.1)	50.5(3.2)	61.4(11.3)/44.4(17.4)
	RD	Token	73.2(14.0)	66.1(3.3)	51.5(7.5)	59.3(7.1)/46.0(3.8)
	RS	Token	71.6(16.6)	65.0(2.0)	51.1(7.1)	64.2(12.1)/46.7(11.5)
	WR	Token	74.1(12.3)	**69.3(2.5)**	55.6(5.9)	60.4(7.5)/43.7(14.2)
	RT	Sentence	82.1(8.2)	68.8(2.4)	59.8(3.9)	**64.3(1.2)/49.8(1.9)**
	ADV	Hidden	**82.3(2.33)**	66.8(5.9)	55.9(3.89)	62.2(10.8)/46.2(9.8)
	Cutoff	Hidden	79.9(5.5)	67.9(0.8)	**60.1(1.0)**	62.7(9.0)/48.1(3.2)

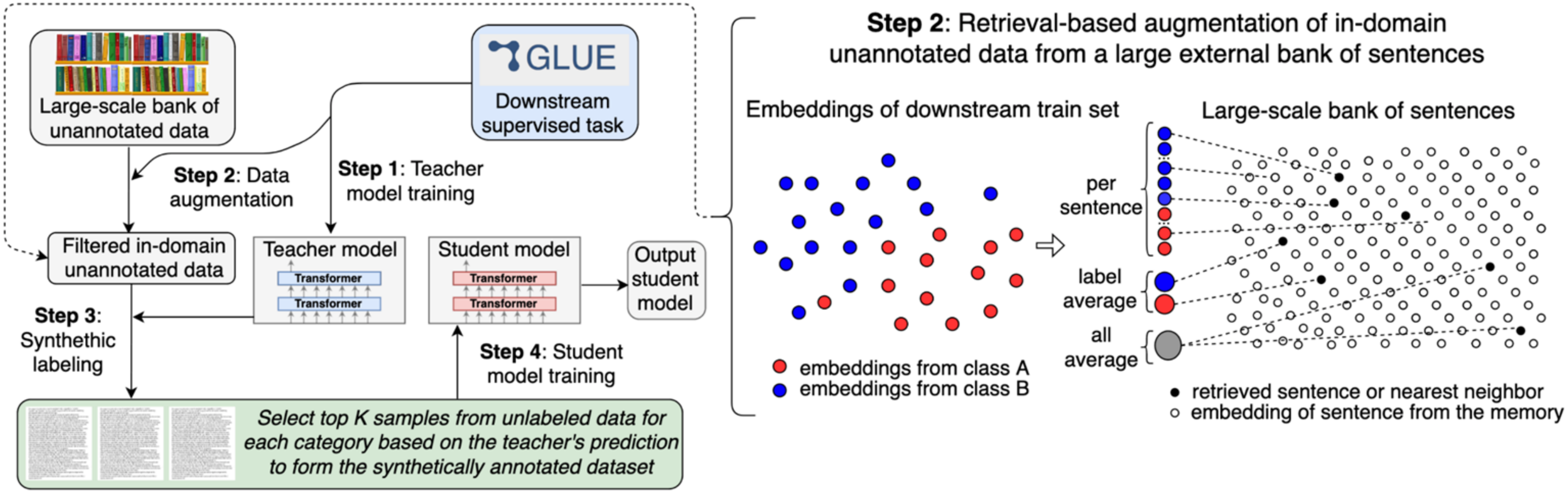
No “best” augmentation

Chen, Jiaao, et al. "An empirical survey of data augmentation for limited data learning in NLP." arXiv preprint arXiv:2106.07499 (2021).

The problem with unlabeled data…

- Some problems (e.g. *machine translation*) are meant to be applied to any text; unlabeled data is abundant
- Some problems (e.g. *sentiment analysis*) only apply to certain kinds of text (e.g. all product reviews but not all tweets)
- For some problems (e.g. *natural language inference*), it is unreasonable to expect that a large amount of unlabeled data is available – it's nearly as hard to collect data as it is to label it.

SentAugment



Du, Jingfei, et al. "Self-training Improves Pre-training for Natural Language Understanding." NAACL 2021.

SentAugment

BioNLP query: A single gene on chromosome 7 makes a protein called the cystic fibrosis transmembrane conductance regulator (CFTR). **Nearest neighbor**: Cystic Fibrosis A mutation in the gene cystic fibrosis transmembrane conductance regulator (CFTR) in chromosome 7.
Financial Query: Google has entered into an agreement to buy Nest Labs for $3.2 billion. **Nearest neighbor**: In January Google (NASDAQ:GOOG) reached an agreement to buy Nest Labs for $3.2 billion in cash.
Hate-speech Query: *Average sentence embeddings of the "hateful" class of IMP* **Nearest neighbor**: fuzzy you are such a d* f* piece of s* just s* your g* d* mouth. – All you n* and s* are fucking ret*
Movie review Query: *Average sentence embeddings of the "bad movie" class of SST-5* **Nearest neighbor**: This movie was terribly boring, but so forgettable as well that it didn't stand out for how awful it was..
Product review Query: *Average sentence embeddings of the "positive" class of CR* **Nearest neighbor**: The phone is very good looking with superb camera setup and very lightweight.
Question type Query: *Average sentence embeddings of the "location" class of TREC* **Nearest neighbor**: Lansing is the capital city of which state?

Du, Jingfei, et al. "Self-training Improves Pre-training for Natural Language Understanding." NAACL 2021.