T5

ACL 2022
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

Diyi Yang
Georgia Tech

Ankur Parikh
Google

Colin Raffel
UNC-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

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

Outline

- [Introduction]: Overview (Colin)
- [Session 1]: Data Augmentation (Diyi)
- [Session 2]: Semi-supervised Learning (Colin)
- [Session 3]: Applications to Multilinguality (Ankur)
- [Conclusion]: Moving Forward (Diyi)

Outline

- [Introduction]: Overview (Colin)
- [Session 1]: Data Augmentation (Diyi)
- [Session 2]: Semi-supervised Learning (Colin)
- [Session 3]: Applications to Multilinguality (Ankur)
- [Conclusion]: Moving Forward (Diyi)

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

Semi-supervised learning

- Consistency regularization
  - Train the model to output consistent predictions after augmentation
- Entropy regularization
  - Train the model to output confident predictions
- Self-training
  - Train the model to predict its own outputs
- How to find unlabeled data?
  - Mine unstructured text corpora for task-specific data
- Leveraging the pre-training format
  - Pre-training on downstream data and framing tasks as cloze problems

Applications to Multilinguality

- What should we do when we have limited data in some languages?
- Multilingual Pre-training
  - Pre-train the model on a large multilingual corpus
- Back-Translation for Machine Translation
  - Generate additional data through paraphrasing
- Zero shot Translation
  - Translate between unseen language pairs
- Unsupervised Machine Translation
  - Translate without any paired data

Outline

- [Introduction]: Overview (Colin)
- [Session 1]: Data Augmentation (Diyi)
- [Session 2]: Semi-supervised Learning (Colin)
- [Session 3]: Applications to Multilinguality (Ankur)
- [Conclusion]: Moving Forward (Diyi)

Data Augmentation

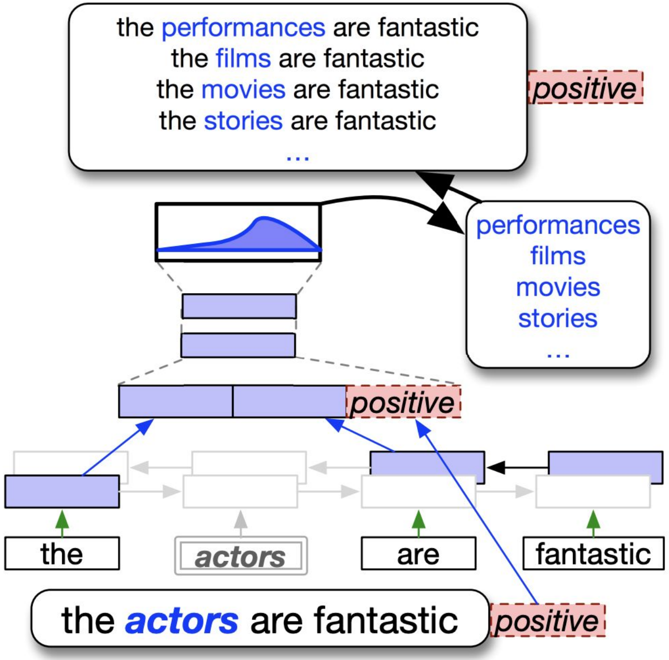
1. Token-level augmentation:
    - Synonym replacement (Yang et al. 2015, Zhang et al. 2015, Miao et al. 2020)
    - Random insertion, deletion, swapping (Xie et al. 2019, Wei and Zou 2019)
    - Word replacement via LM (Wu et al. 2019, Zhu et al. 2019)
2. Sentence-level augmentation:
    - Paraphrasing (Xie et al. 2019, Chen et al. 2020)
    - Conditional generation (Zhang and Bansal 2019, Yang et al. 2020)
3. Adversarial augmentation:
    - Paraphrasing (Xie et al. 2019, Chen et al. 2020)
    - Conditional generation (Zhang and Bansal 2019, Yang et al. 2020)
4. Hidden space augmentation:
    - Mixup (Zhang et al., 2019, Chen et al. 2020)

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** road 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



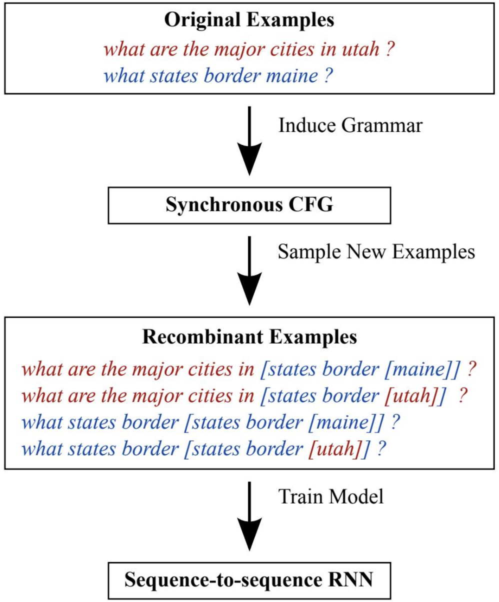
Contextual augmentation, when a sentence "*the actors are fantastic*" is augmented by replacing only actors with words predicted based on the context (Kobayashi, 2018)

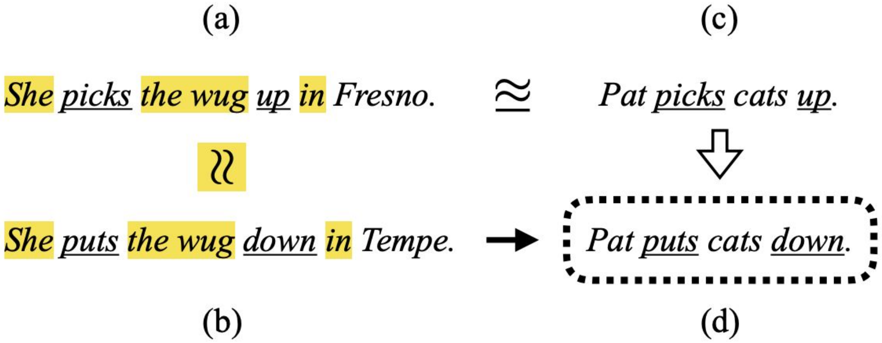
Soft contextual data augmentation

(Gao et al., 2019)

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

Compositional Augmentation

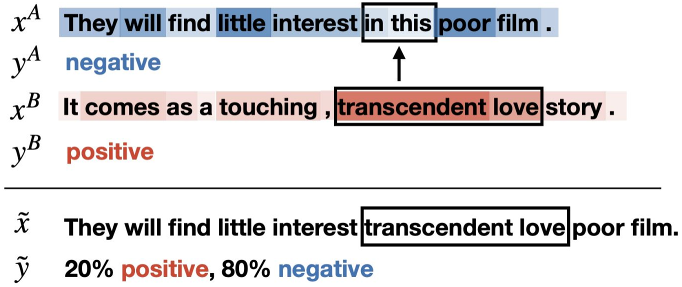


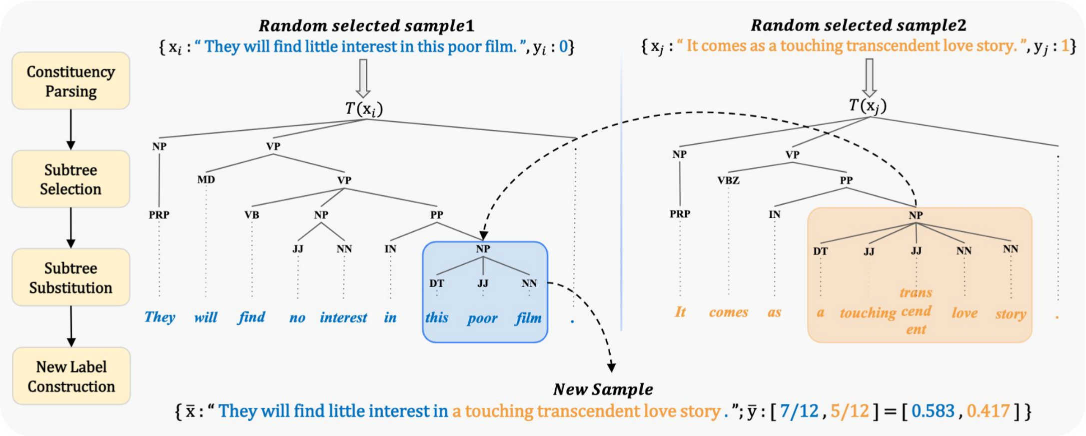


Induce a high-precision synchronous context-free grammar, and then sample from this grammar to generate new "recombinant" examples (Jia and Liang, 2016)

Two discontinuous sentence fragments (a–b, underlined) which appear in similar environments (a–b, highlighted) are identified. Additional sentences in which the first fragment appears (c) are used to synthesize new examples (d) by substituting in the second fragment (Andreas, 2020)

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

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

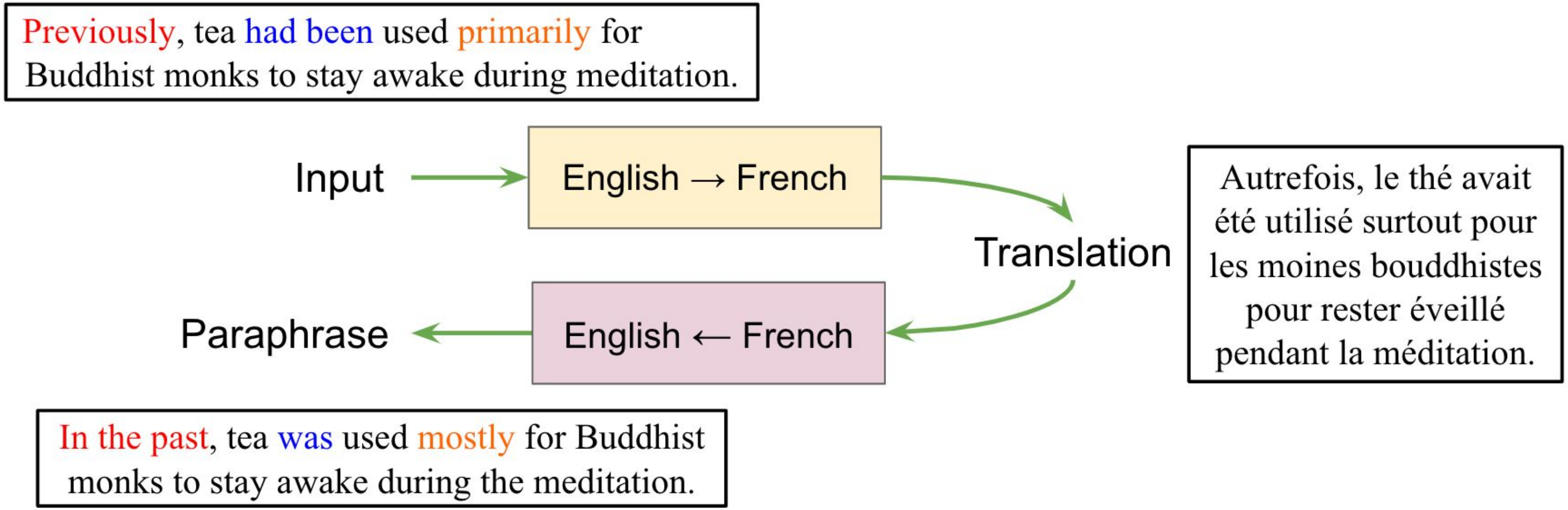


Image credit to https://github.com/vietai/dab

Paraphrasing

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 .

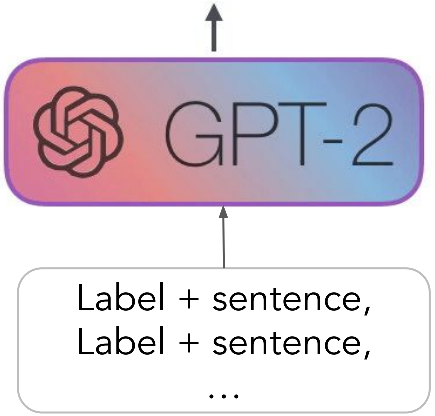
syntactically controlled paraphrase generation (Iyyer et al., 2018)

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

White-box Attack

HotFlip uses the model gradient to identify the most important letter in the text

(Ebrahimi et al., 2018)

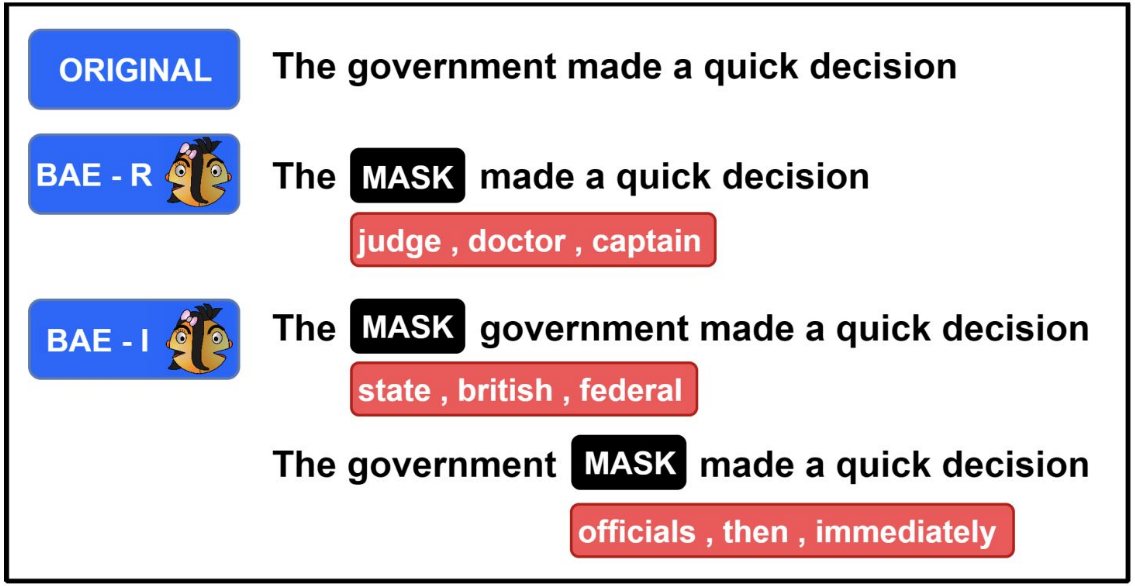
$$\max \nabla_x J(\mathbf{x}, \mathbf{y})^T \cdot \vec{v}_{ijb} = \max_{ijb} \frac{\partial J}{\partial x_{ij}}^{(b)} - \frac{\partial J}{\partial x_{ij}}^{(a)}$$

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



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 or Sentence	Medium	Text classification, Sequence labeling, Machine translation
Black-box attack	Token or Sentence	Medium	Text classification, Sequence labeling, Machine translation, Textual entailment, Dialogue generation, Text Summarization

Hidden-space Augmentation via Perturbation

Manipulating the hidden representations

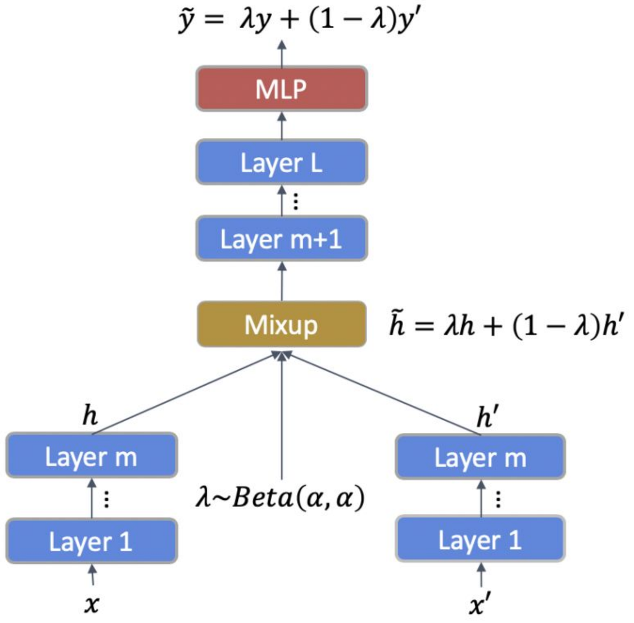
- Through perturbations such as adding noises
- Or performing interpolations with other data points

Interpolation: mixup in textual hidden space

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

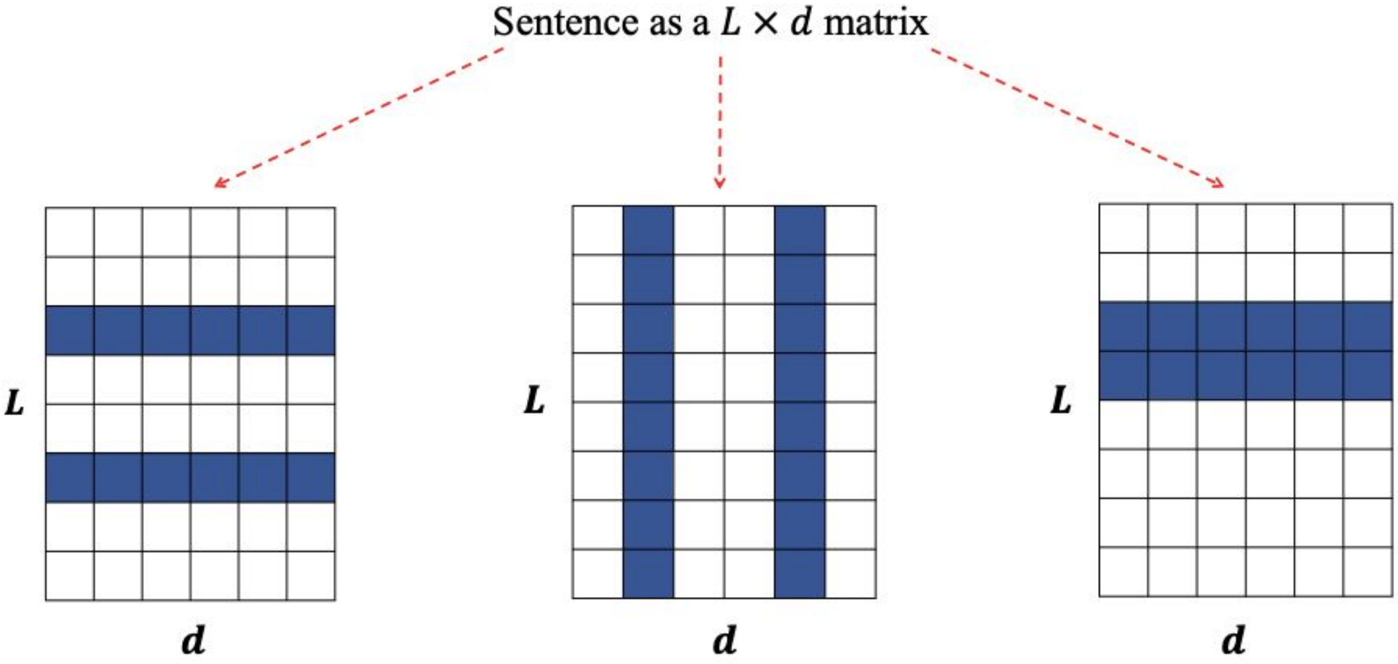
$$\tilde{\mathbf{y}} = \text{mix}(\mathbf{y}_i, \mathbf{y}_j) = \lambda\mathbf{y}_i + (1-\lambda)\mathbf{y}_j$$

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



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 learning

Cutoff removes the information from the input embedding matrix

$$\mathcal{L} = \mathcal{L}_{\text{ce}}(x, y) + \alpha \sum_{i=1}^{N} \mathcal{L}_{\text{ce}}(x_{\text{cutoff}}^{i}, y)$$
$$+ \ \beta \mathcal{L}_{\text{divergence}}(x, x_{\text{cutoff}}^{1}, x_{\text{cutoff}}^{2}, \ldots, x_{\text{cutoff}}^{N}, y)$$

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).

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)

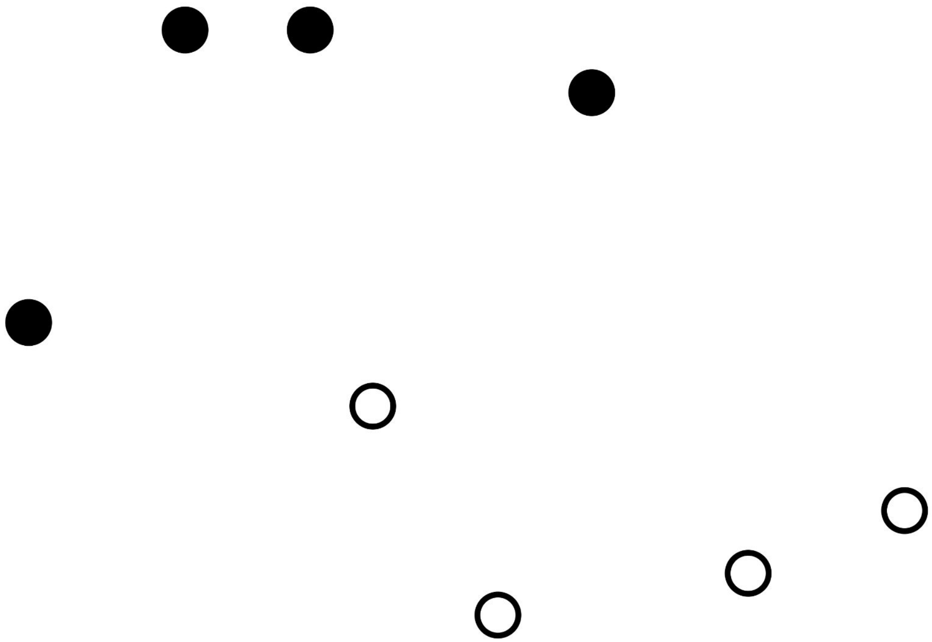
Outline

- [Introduction]: Overview (Colin)
- [Session 1]: Data Augmentation (Diyi)
- [Session 2]: Semi-supervised Learning (Colin)
- [Session 3]: Applications to Multilinguality (Ankur)
- [Conclusion]: Moving Forward (Diyi)

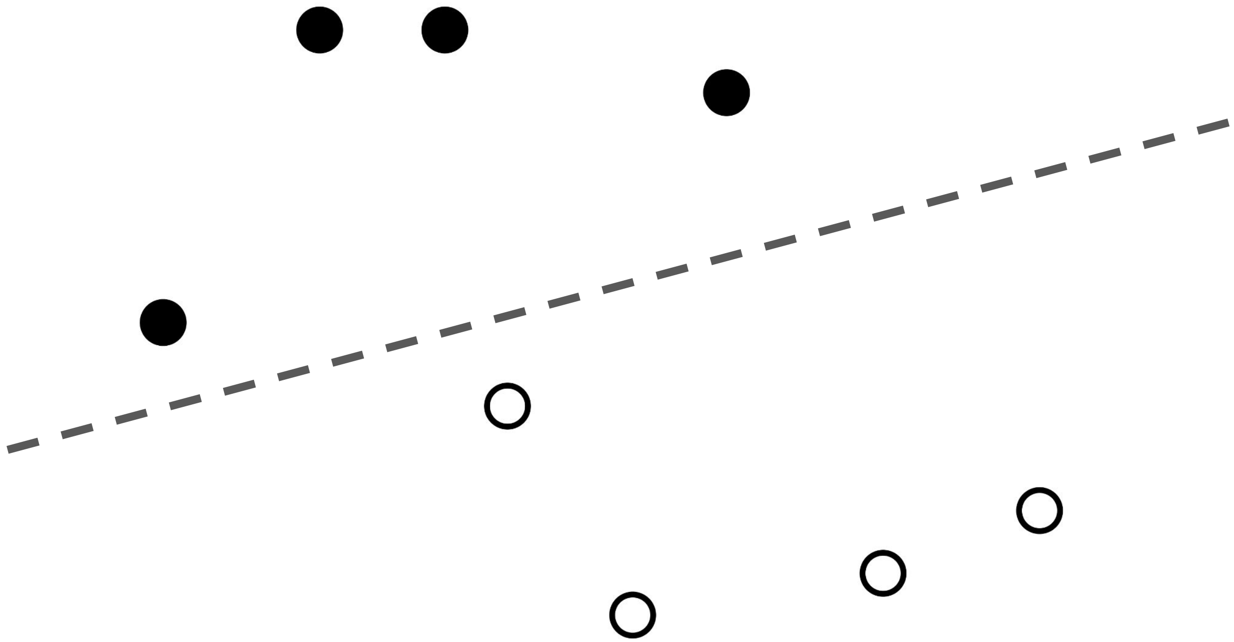
Semi-Supervised Learning

- What is semi-supervised learning?
- Consistency regularization
- Entropy minimization
- Self-training
- Finding unlabeled data
- Continued pre-training
- Pattern-exploiting training

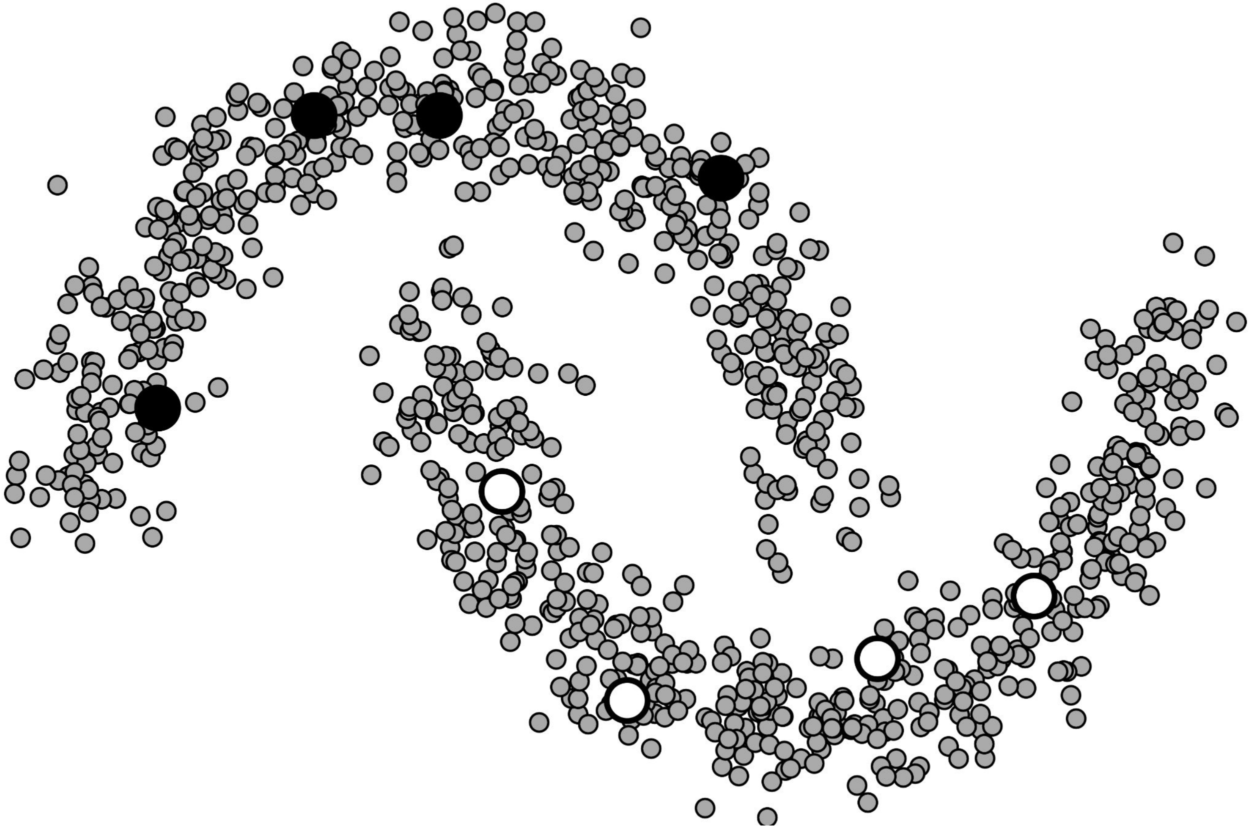
Semi-Supervised Learning



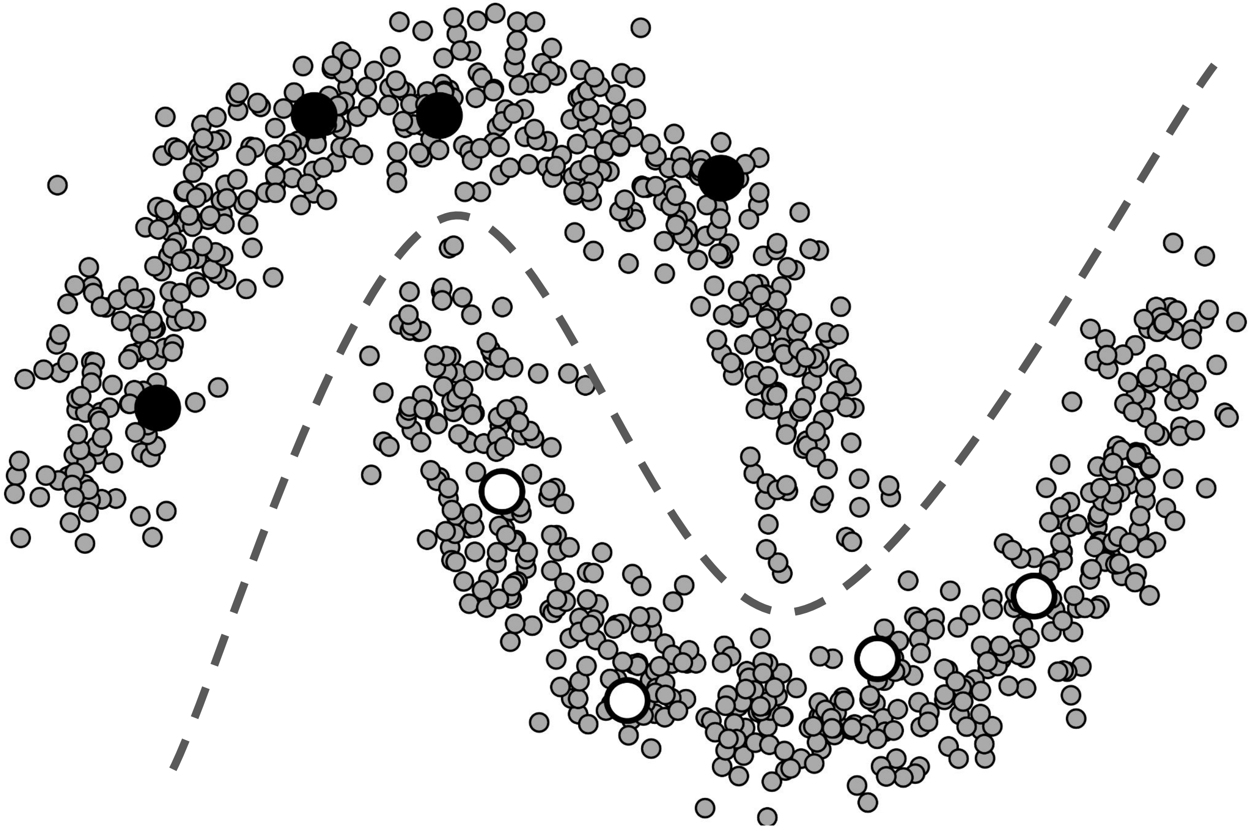
Semi-Supervised Learning



Semi-Supervised Learning



Semi-Supervised Learning



Supervised Learning

$$x, y \sim p(x, y)$$

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

Semi-Supervised Learning

$$x, y \sim p(x, y)$$

and

$$x \sim p(x)$$

Transfer Learning

$$x, y \sim p(x, y)$$

and

$$x, y \sim q(x, y) \text{ or } x \sim q(x)$$

How to use unlabeled data?

$$x, y \sim p(x, y)$$

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

How to use unlabeled data?

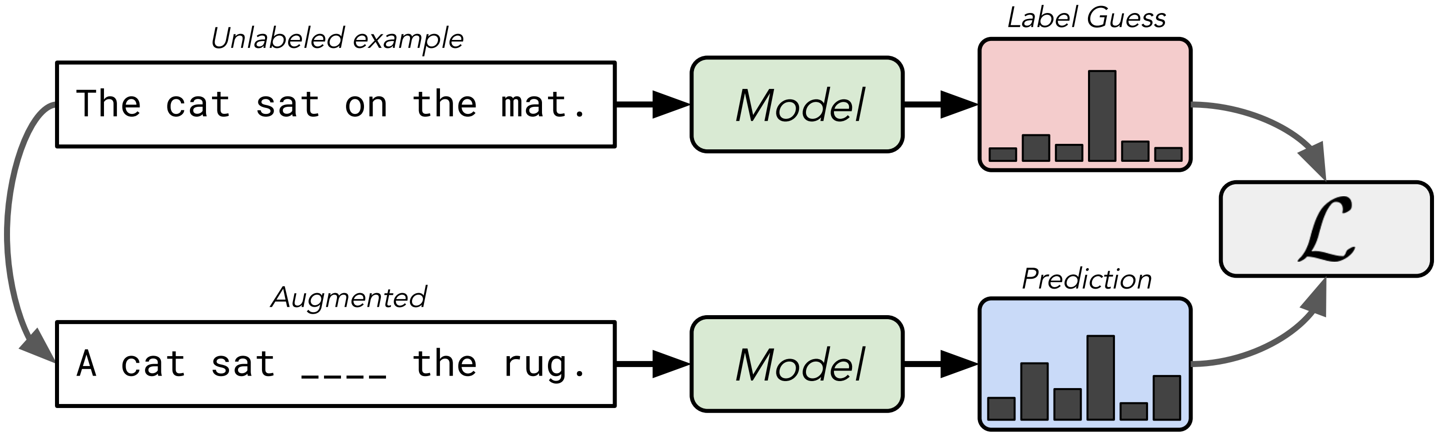
$$x \sim p(x)$$

Use a proxy-label/pseudo-label/label guess

$$x \sim p(x)$$

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

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
$\text{BERT}_{\text{BASE}}$	✗	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

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		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)

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).

Entropy regularization

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

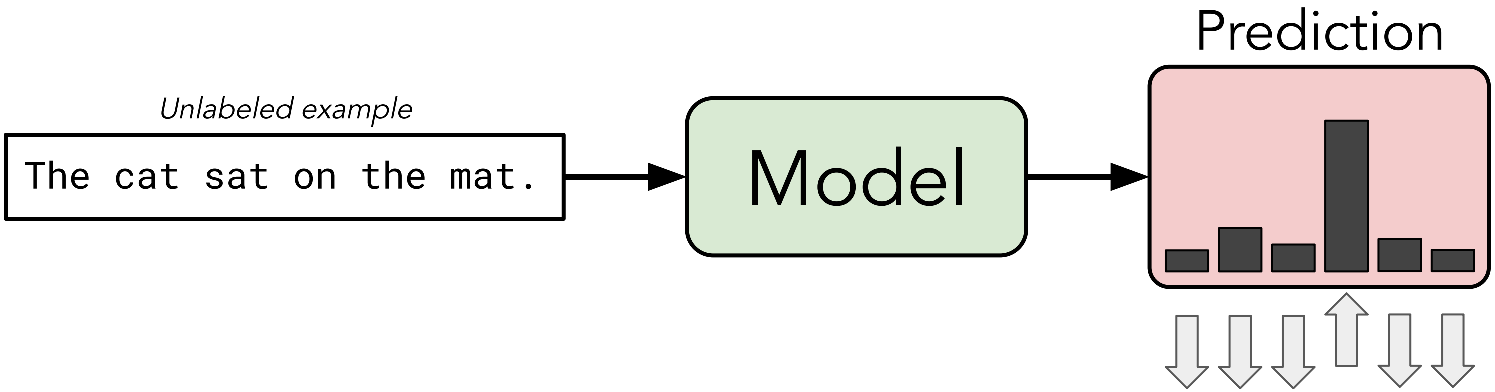
$$\hat{p}_\theta(y|x) = p_\theta(y|x)$$

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

Entropy!

Grandvalet, Yves, and Yoshua Bengio. "Semi-Supervised Learning by Entropy Minimization." NeurIPS 2004.

Entropy regularization

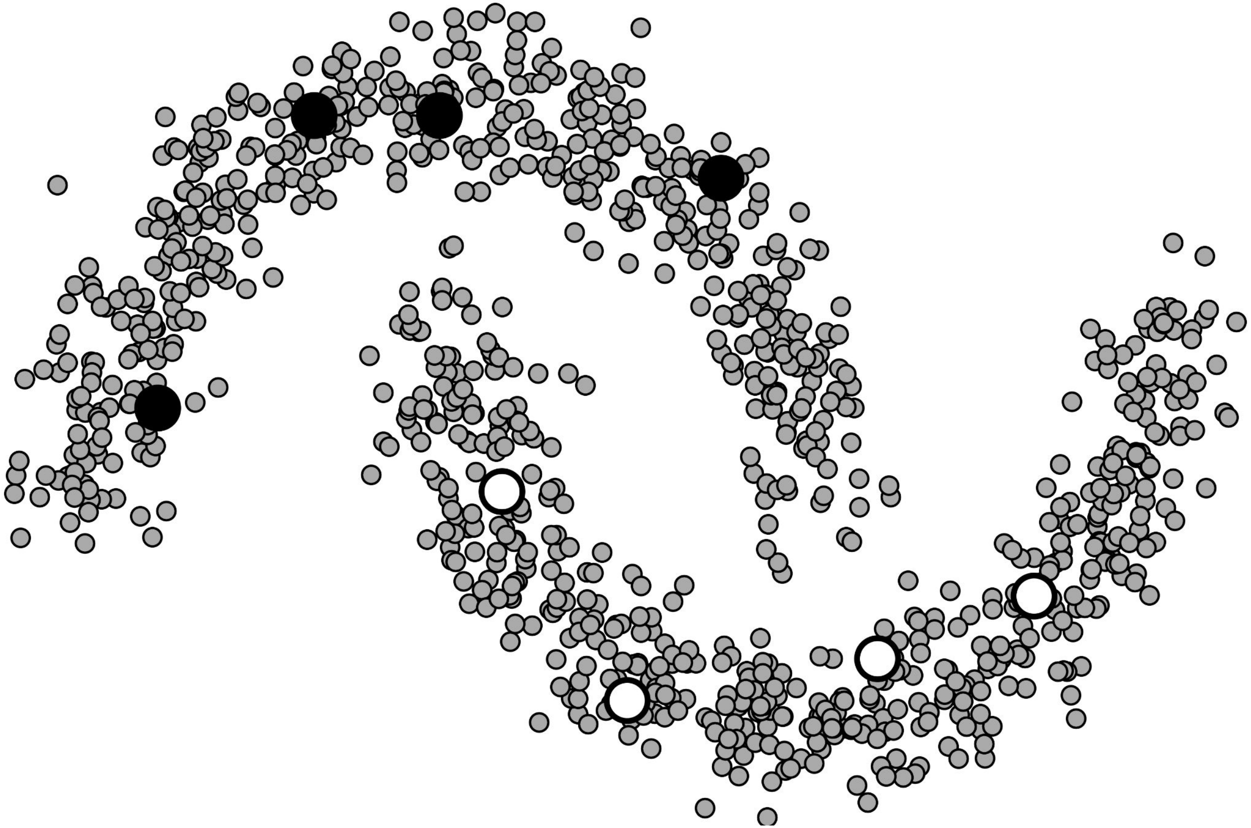


$$\mathbb{E}_x - p_\theta(y|x) \log p_\theta(y|x)$$

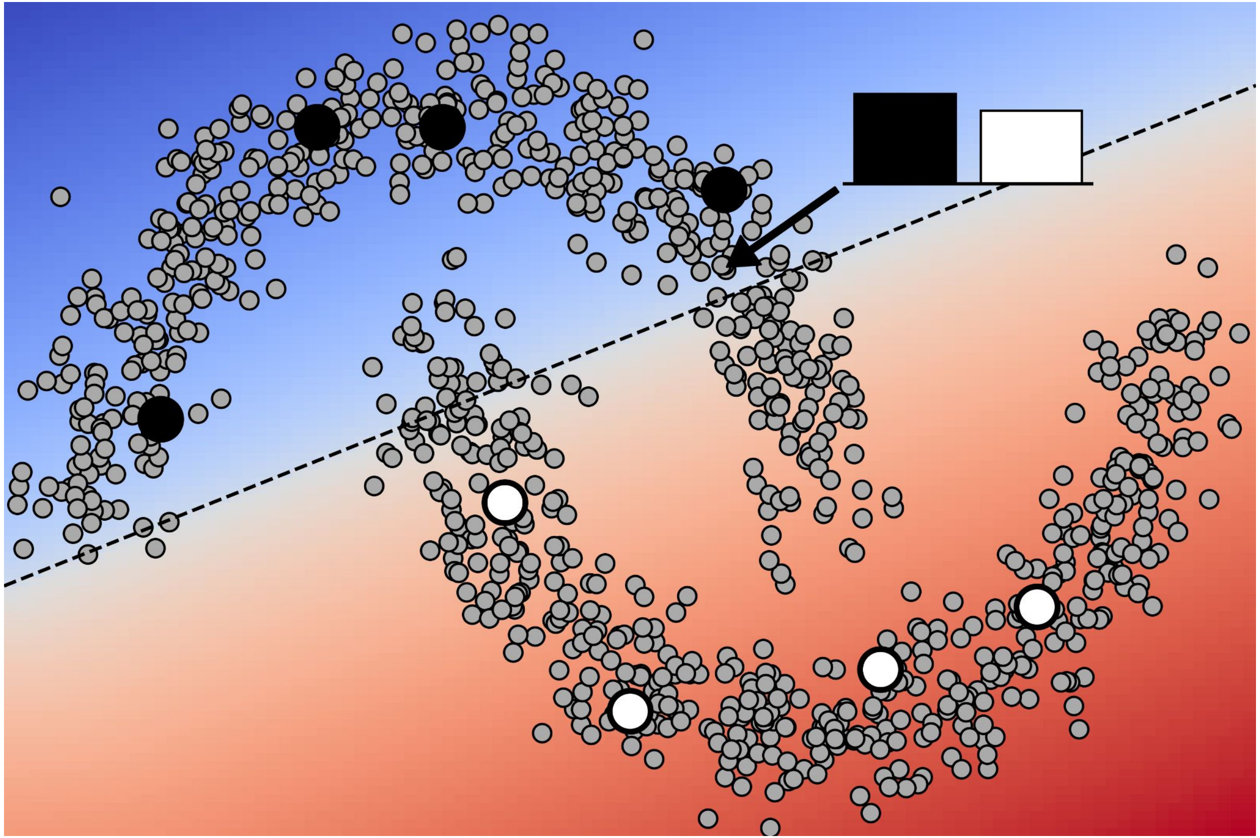
Entropy!

Grandvalet, Yves, and Yoshua Bengio. "Semi-Supervised Learning by Entropy Minimization." NeurIPS 2004.

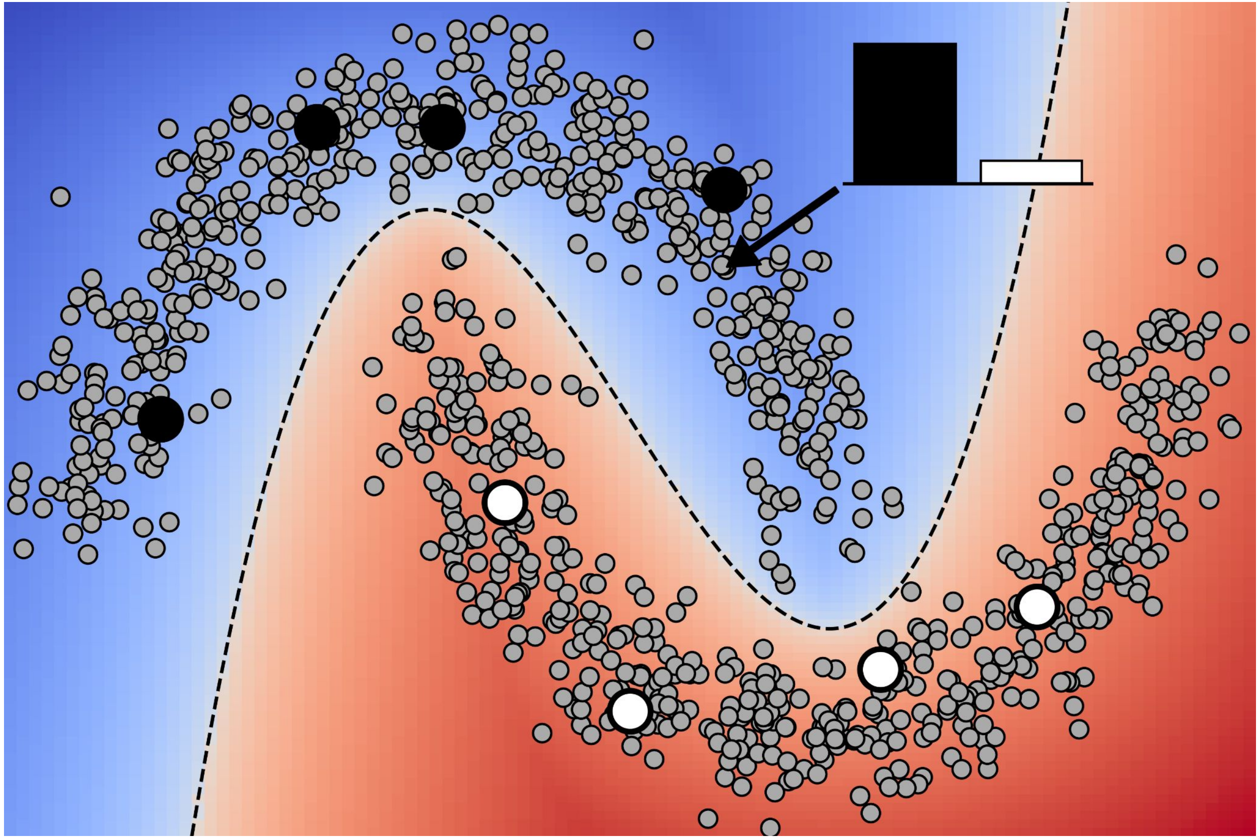
Why does(n't) entropy minimization work?



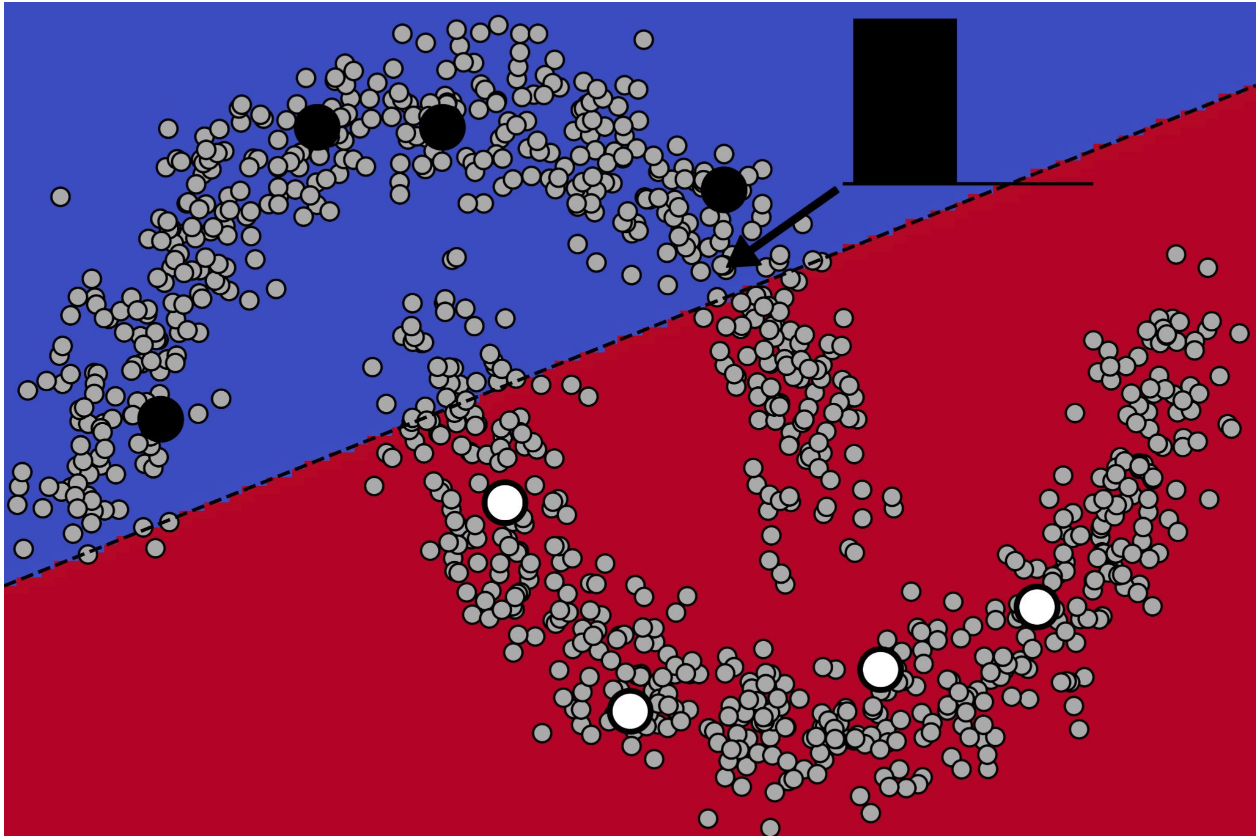
Why does(n't) entropy minimization work?



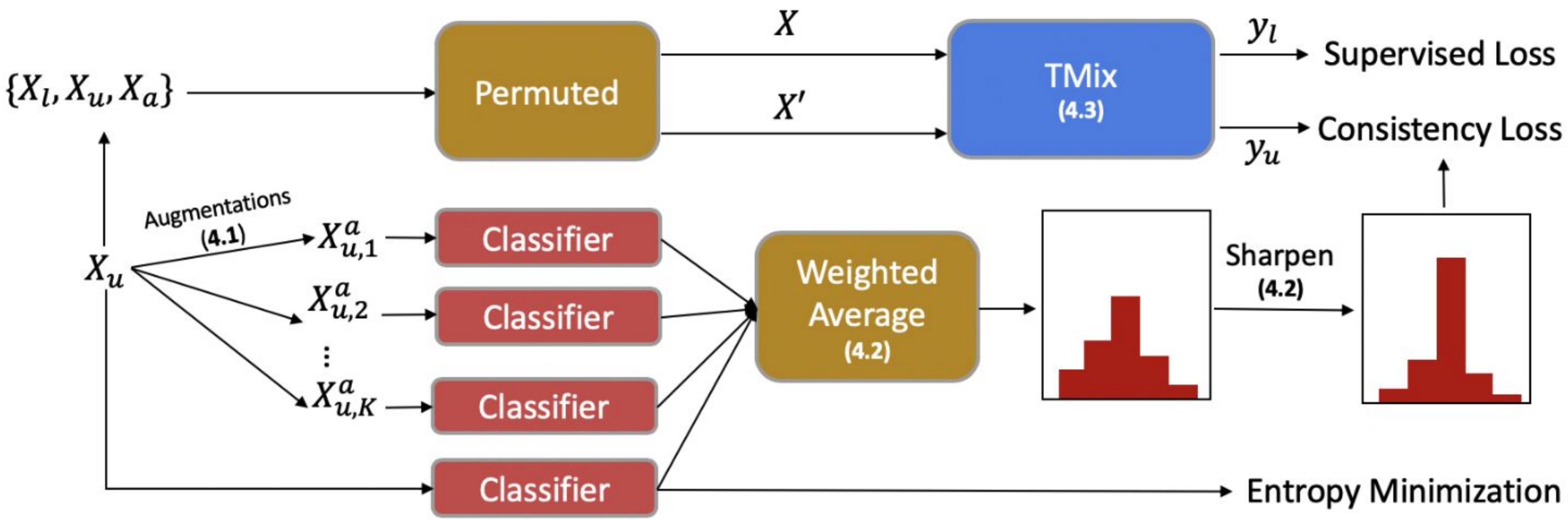
Why does(n't) entropy minimization work?



Why does(n't) entropy minimization work?

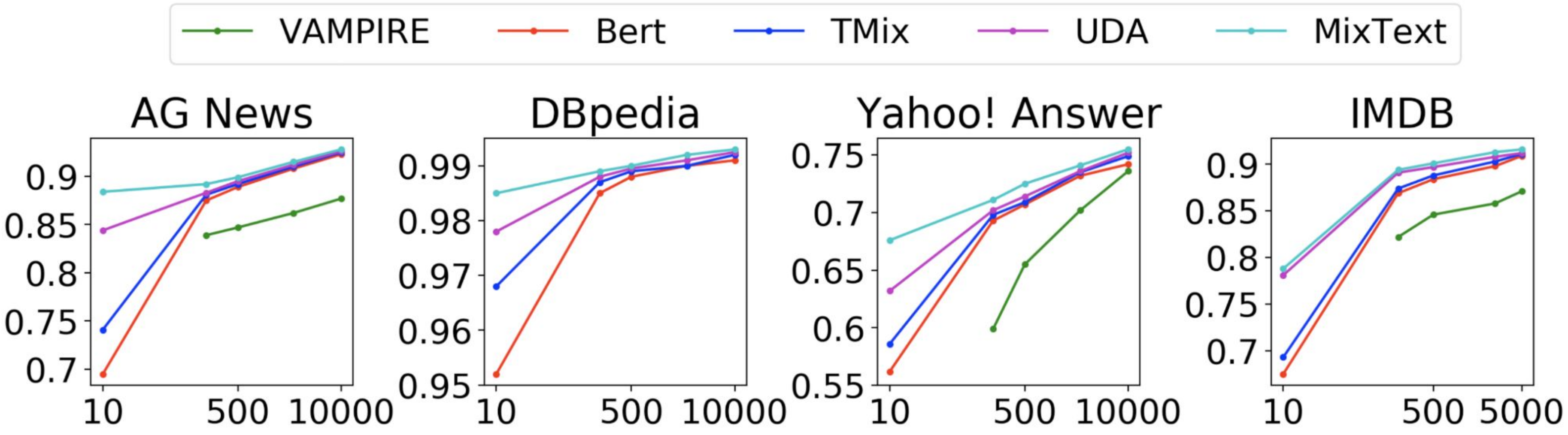


MixText



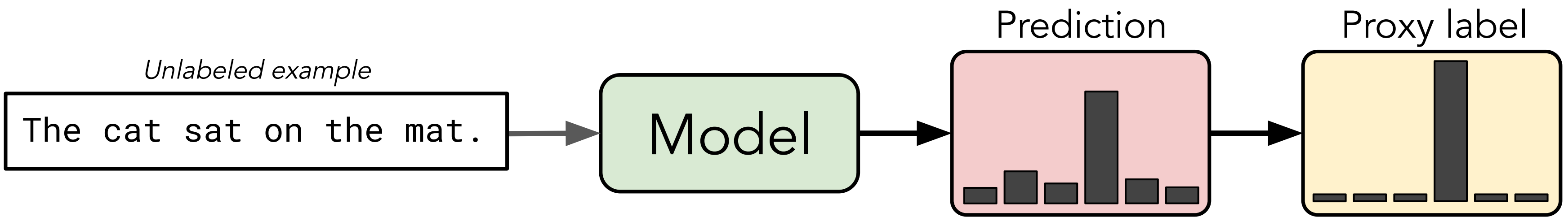
Chen, Jiaao, Zichao Yang, and Diyi Yang. "MixText: Linguistically-Informed Interpolation of Hidden Space for Semi-Supervised Text Classification." ACL 2020.

MixText



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

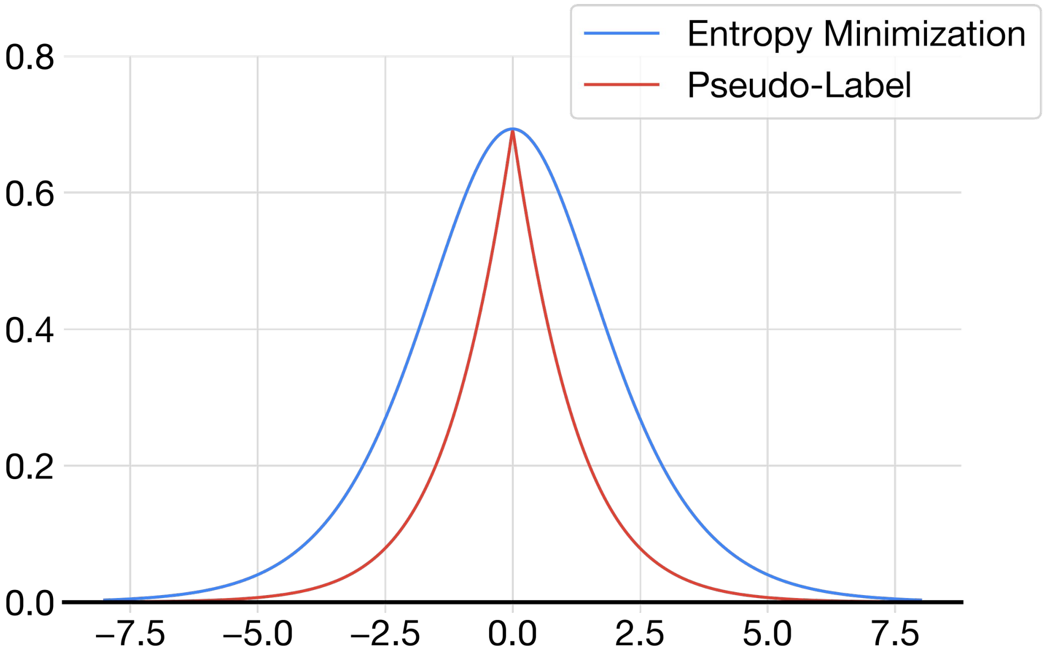
Self-training



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

$$\hat{p}_\theta(y|x) = \arg\max_y [p_\theta(y|x)]$$

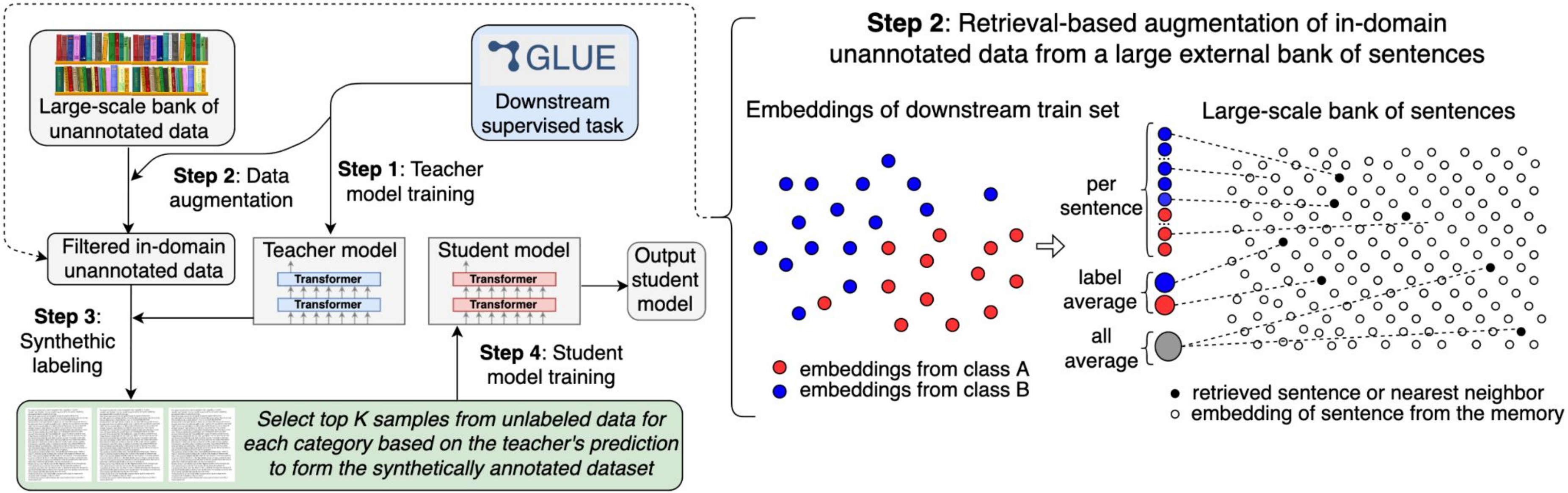
Self-training vs. entropy minimization



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.

SentAugment

Model	SST-2	SST-5	CR	IMP	TREC	NER	Avg
$\text{RoBERTa}_{\text{Large}}$	96.5	57.8	94.8	84.6	**97.8**	92.7	87.4
$\text{RoBERTa}_{\text{Large}}$ + ICP	93.9	55.1	93.7	84.4	97.8	92.1	86.2
$\text{RoBERTa}_{\text{Large}}$ + ST	**96.7**	**60.4**	**95.7**	**87.7**	**97.8**	**93.3**	**88.6**

Table 2: Results of self-training on natural language understanding benchmarks. We report a strong RoBERTa-Large baseline, as well as in-domain continued pretraining of this model (ICP) and our self-training approach (ST).

Model	SST-2	SST-5	CR	IMP	TREC	NER	Avg
Num samples	40	100	40	40	120	200	-
$\text{RoBERTa}_{\text{Large}}$	83.6±2.7	42.3±1.6	88.9±1.7	77.3±2.8	90.9±2.5	49.0±1.7	72.0±2.2
$\text{RoBERTa}_{\text{Large}}$ + ST	**86.7±2.3**	**44.4±1.0**	**89.7±2.0**	**81.9±1.4**	**92.1±2.4**	**58.4±1.4**	**75.5±1.8**

Table 3: Results of self-training for few-shot learning, using only 20 samples per class.

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

SentAugment

Model	SST-2	SST-5	CR	IMP	TREC	NER	Avg
RoBERTa$_{\text{Large}}$	96.5	57.8	94.8	84.6	**97.8**	92.7	87.4
RoBERTa$_{\text{Large}}$ + ICP	93.9	55.1	93.7	84.4	97.8	92.1	86.2
RoBERTa$_{\text{Large}}$ + ST	**96.7**	**60.4**	**95.7**	**87.7**	**97.8**	**93.3**	**88.6**

Works better than continued pretraining

Table 2: Results of self-training on natural language understanding benchmarks. We report a strong RoBERTa-Large baseline, as well as in-domain continued pretraining of this model (ICP) and our self-training approach (ST).

Model	SST-2	SST-5	CR	IMP	TREC	NER	Avg
Num samples	40	100	40	40	120	200	-
RoBERTa$_{\text{Large}}$	83.6±2.7	42.3±1.6	88.9±1.7	77.3±2.8	90.9±2.5	49.0±1.7	72.0±2.2
RoBERTa$_{\text{Large}}$ + ST	**86.7±2.3**	**44.4±1.0**	**89.7±2.0**	**81.9±1.4**	**92.1±2.4**	**58.4±1.4**	**75.5±1.8**

Table 3: Results of self-training for few-shot learning, using only 20 samples per class.

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

SentAugment

Model	SST-2	SST-5	CR	IMP	TREC	NER	Avg
$\text{RoBERTa}_{\text{Large}}$	96.5	57.8	94.8	84.6	**97.8**	92.7	87.4
$\text{RoBERTa}_{\text{Large}}$ + ICP	93.9	55.1	93.7	84.4	97.8	92.1	86.2
$\text{RoBERTa}_{\text{Large}}$ + ST	**96.7**	**60.4**	**95.7**	**87.7**	**97.8**	**93.3**	**88.6**

Table 2: Results of self-training on natural language understanding benchmarks. We report a strong RoBERTa-Large baseline, as well as in-domain continued pretraining of this model (ICP) and our self-training approach (ST).

Model	SST-2	SST-5	CR	IMP	TREC	NER	Avg
Num samples	40	100	40	40	120	200	-
$\text{RoBERTa}_{\text{Large}}$	83.6±2.7	42.3±1.6	88.9±1.7	77.3±2.8	90.9±2.5	49.0±1.7	72.0±2.2
$\text{RoBERTa}_{\text{Large}}$ + ST	**86.7±2.3**	**44.4±1.0**	**89.7±2.0**	**81.9±1.4**	**92.1±2.4**	**58.4±1.4**	**75.5±1.8**

Bigger gains
in few-shot

Table 3: Results of self-training for few-shot learning, using only 20 samples per class.

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

SentAugment

Model	SST-2	SST-5	CR	IMP	TREC	NER	Avg
$\text{RoBERTa}_{\text{Large}}$	96.5	57.8	94.8	84.6	**97.8**	92.7	87.4
$\text{RoBERTa}_{\text{Large}}$ + ICP	93.9	55.1	93.7	84.4	97.8	92.1	86.2
$\text{RoBERTa}_{\text{Large}}$ + ST	**96.7**	**60.4**	**95.7**	**87.7**	**97.8**	**93.3**	**88.6**

Let's revisit this

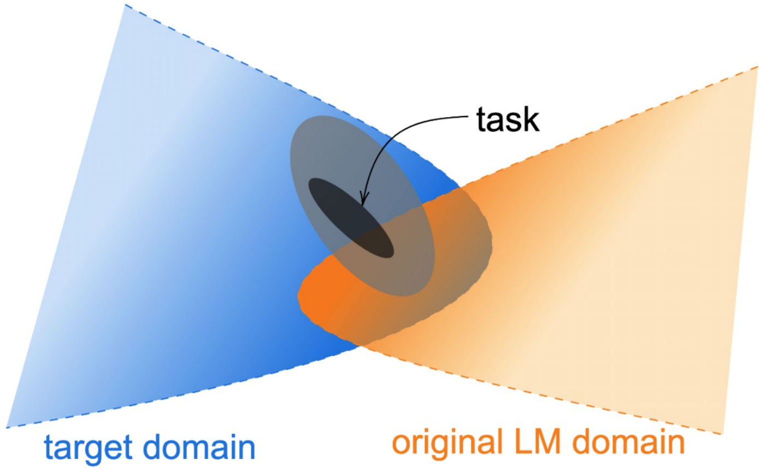
Table 2: Results of self-training on natural language understanding benchmarks. We report a strong RoBERTa-Large baseline, as well as in-domain continued pretraining of this model (ICP) and our self-training approach (ST).

Model	SST-2	SST-5	CR	IMP	TREC	NER	Avg
Num samples	40	100	40	40	120	200	-
$\text{RoBERTa}_{\text{Large}}$	83.6±2.7	42.3±1.6	88.9±1.7	77.3±2.8	90.9±2.5	49.0±1.7	72.0±2.2
$\text{RoBERTa}_{\text{Large}}$ + ST	**86.7±2.3**	**44.4±1.0**	**89.7±2.0**	**81.9±1.4**	**92.1±2.4**	**58.4±1.4**	**75.5±1.8**

Table 3: Results of self-training for few-shot learning, using only 20 samples per class.

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

Domain/Task-adaptive pre-training



Domain	Pretraining Corpus	# Tokens	Size	$\mathcal{L}_{\text{ROB.}}$	$\mathcal{L}_{\text{DAPT}}$
BIOMED	2.68M full-text papers from S2ORC (Lo et al., 2020)	7.55B	47GB	1.32	0.99
CS	2.22M full-text papers from S2ORC (Lo et al., 2020)	8.10B	48GB	1.63	1.34
NEWS	11.90M articles from REALNEWS (Zellers et al., 2019)	6.66B	39GB	1.08	1.16
REVIEWS	24.75M AMAZON reviews (He and McAuley, 2016)	2.11B	11GB	2.10	1.93

Gururangan, Suchin, et al. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." ACL 2020.

Domain/Task-adaptive pre-training

Pretraining	Steps	Docs.	Storage	F_1
RoBERTa	-	-	-	$79.3_{0.6}$
TAPT	0.2K	500	80KB	$79.8_{1.4}$
50NN-TAPT	1.1K	24K	3MB	$80.8_{0.6}$
150NN-TAPT	3.2K	66K	8MB	$81.2_{0.8}$
500NN-TAPT	9.0K	185K	24MB	$81.7_{0.4}$
Curated-TAPT	8.8K	180K	27MB	$\mathbf{83.4}_{0.3}$
DAPT	12.5K	25M	47GB	$82.5_{0.5}$
DAPT + TAPT	12.6K	25M	47GB	$83.0_{0.3}$

Table 9: Computational requirements for adapting to the RCT-500 task, comparing DAPT (§3) and the various TAPT modifications described in §4 and §5.

Gururangan, Suchin, et al. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." ACL 2020.

Domain/Task-adaptive pre-training

Pretraining	Steps	Docs.	Storage	F_1
RoBERTa	-	-	-	$79.3_{0.6}$
TAPT	0.2K	500	80KB	$79.8_{1.4}$
50NN-TAPT	1.1K	24K	3MB	$80.8_{0.6}$
150NN-TAPT	3.2K	66K	8MB	$81.2_{0.8}$
500NN-TAPT	9.0K	185K	24MB	$81.7_{0.4}$
Curated-TAPT	8.8K	180K	27MB	$\mathbf{83.4}_{0.3}$
DAPT	12.5K	25M	47GB	$82.5_{0.5}$
DAPT + TAPT	12.6K	25M	47GB	$83.0_{0.3}$

Similar to SentAugment

Table 9: Computational requirements for adapting to the RCT-500 task, comparing DAPT (§3) and the various TAPT modifications described in §4 and §5.

Gururangan, Suchin, et al. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." ACL 2020.

Domain/Task-adaptive pre-training

Pretraining	Steps	Docs.	Storage	F_1
RoBERTa	-	-	-	$79.3_{0.6}$
TAPT	0.2K	500	80KB	$79.8_{1.4}$
50NN-TAPT	1.1K	24K	3MB	$80.8_{0.6}$
150NN-TAPT	3.2K	66K	8MB	$81.2_{0.8}$
500NN-TAPT	9.0K	185K	24MB	$81.7_{0.4}$
Curated-TAPT	8.8K	180K	27MB	$\mathbf{83.4}_{0.3}$
DAPT	12.5K	25M	47GB	$82.5_{0.5}$
DAPT + TAPT	12.6K	25M	47GB	$83.0_{0.3}$

Table 9: Computational requirements for adapting to the RCT-500 task, comparing DAPT (§3) and the various TAPT modifications described in §4 and §5.

"Oracle" unlabeled data works best

Gururangan, Suchin, et al. "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks." ACL 2020.

Pattern-exploiting training

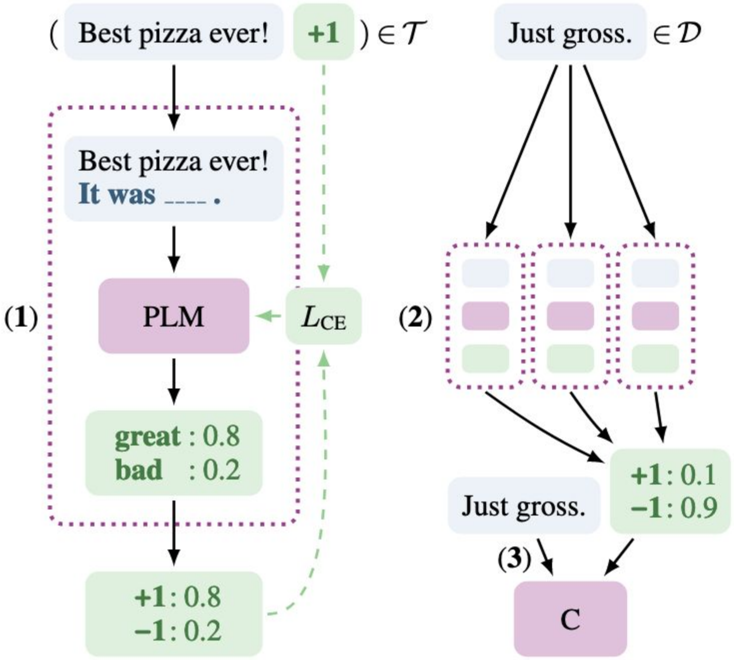


Figure 1: PET for sentiment classification. **(1)** A number of patterns encoding some form of task description are created to convert training examples to cloze questions; for each pattern, a pretrained language model is finetuned. **(2)** The ensemble of trained models annotates unlabeled data. **(3)** A classifier is trained on the resulting soft-labeled dataset.

Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." EACL 2021.

Pattern-exploiting training

Ex.	Method	Yelp	AG's	Yahoo	MNLI
$\|\mathcal{T}\| = 10$	UDA	27.3	72.6	36.7	34.7
	MixText	20.4	81.1	20.6	32.9
	PET	48.8	84.1	59.0	39.5
	iPET	**52.9**	**87.5**	**67.0**	**42.1**
$\|\mathcal{T}\| = 50$	UDA	46.6	83.0	60.2	40.8
	MixText	31.3	84.8	61.5	34.8
	PET	55.3	86.4	63.3	55.1
	iPET	**56.7**	**87.3**	**66.4**	**56.3**

Table 2: Comparison of PET with two state-of-the-art semi-supervised methods using RoBERTa (base)

Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." EACL 2021.

Pattern-exploiting training

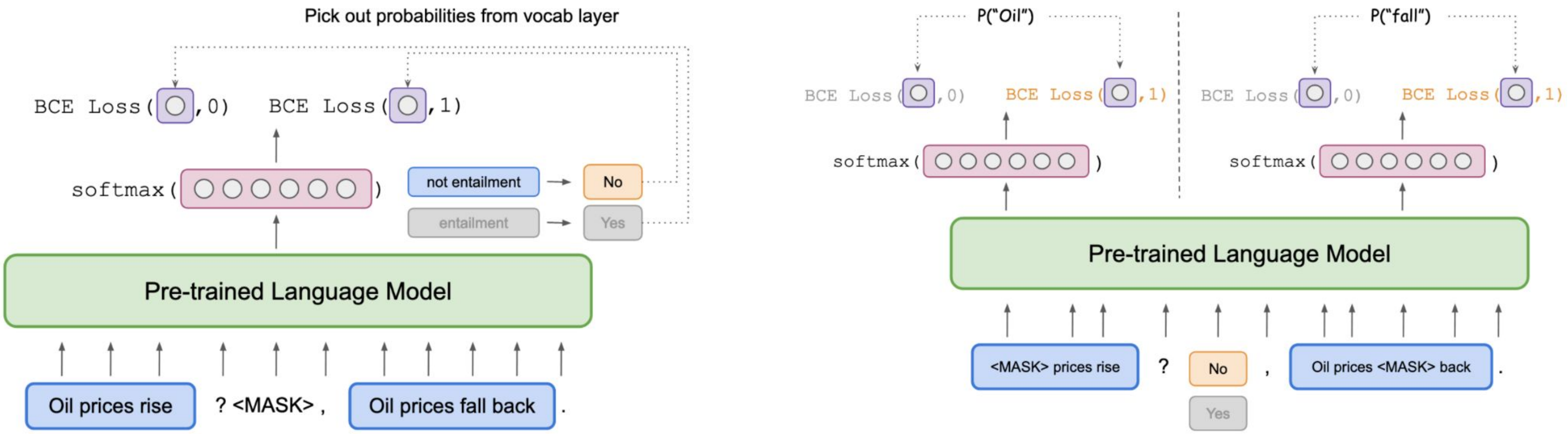
Ex.	Method	Yelp	AG's	Yahoo	MNLI
$\|\mathcal{T}\| = 10$	UDA	27.3	72.6	36.7	34.7
	MixText	20.4	81.1	20.6	32.9
	PET	48.8	84.1	59.0	39.5
	iPET	**52.9**	**87.5**	**67.0**	**42.1**
$\|\mathcal{T}\| = 50$	UDA	46.6	83.0	60.2	40.8
	MixText	31.3	84.8	61.5	34.8
	PET	55.3	86.4	63.3	55.1
	iPET	**56.7**	**87.3**	**66.4**	**56.3**

Using the pretraining format helps!

Table 2: Comparison of PET with two state-of-the-art semi-supervised methods using RoBERTa (base)

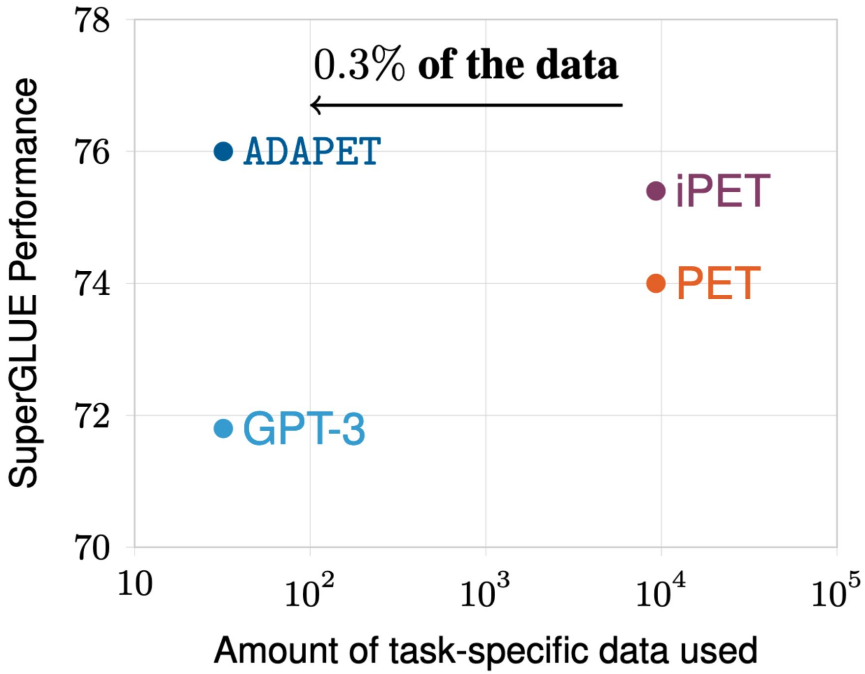
Schick, Timo, and Hinrich Schütze. "Exploiting Cloze-Questions for Few-Shot Text Classification and Natural Language Inference." EACL 2021.

Simplifying PET



Tam, Derek, et al. "Improving and Simplifying Pattern Exploiting Training." EMNLP 2021.

Is unlabeled data necessary?



Tam, Derek, et al. "Improving and Simplifying Pattern Exploiting Training." EMNLP 2021.

Outline

- [Introduction]: Overview (Colin)
- [Session 1]: Data Augmentation (Diyi)
- [Session 2]: Semi-supervised Learning (Colin)
- [Session 3]: Applications to Multilinguality (Ankur)
- [Conclusion]: Moving Forward (Diyi)

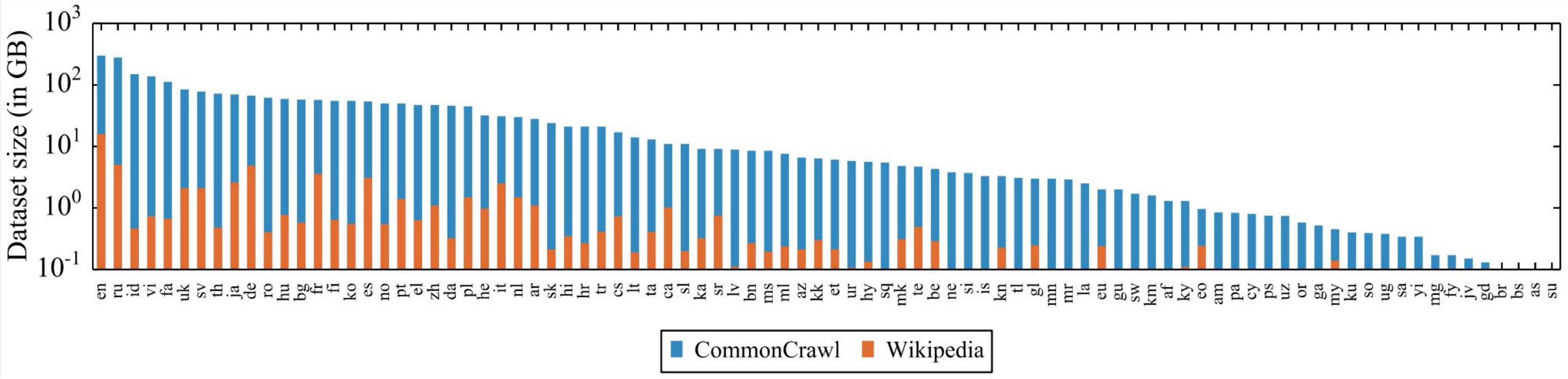
Applications to Multilinguality

- Introduction
- Multilingual Pre-training
- Back-Translation for Machine Translation
- Zero shot Translation
- Unsupervised Machine Translation

Applications to Multilinguality

- Introduction
- Multilingual Pre-training
- Back-Translation for Machine Translation
- Zero shot Translation
- Unsupervised Machine Translation

Long tail of Multilinguality



Source: Conneau et al. 2020.

Common Approaches

Most multilingual/cross-lingual approaches use one or both of the below techniques:

- *Multilingual pre-training:* Natural way to leverage high resource languages to improve low resource ones.

- *Machine translation:* Translate the training set (works much better than translating in test time)

Example Task: Multilingual Zero shot classification

XNLI [Conneau et al. 2018]:

- Initialize from multilingual pretrained model
- Fine-tune on English data.
- Evaluate on 14 other languages.

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
Machine translation baselines (TRANSLATE TRAIN)															
BiLSTM-last	71.0	66.7	67.0	65.7	65.3	65.6	65.1	61.9	63.9	63.1	61.3	65.7	61.3	55.2	55.2
BiLSTM-max	**73.7**	68.3	68.8	66.5	66.4	67.4	66.5	64.5	65.8	66.0	62.8	67.0	62.1	58.2	56.6
Machine translation baselines (TRANSLATE TEST)															
BiLSTM-last	71.0	68.3	68.7	66.9	67.3	68.1	66.2	64.9	65.8	64.3	63.2	66.5	61.8	60.1	58.1
BiLSTM-max	**73.7**	**70.4**	**70.7**	**68.7**	**69.1**	**70.4**	**67.8**	**66.3**	**66.8**	**66.5**	**64.4**	**68.3**	**64.2**	**61.8**	**59.3**
Evaluation of XNLI multilingual sentence encoders (in-domain)															
X-BiLSTM-last	71.0	65.2	67.8	66.6	66.3	65.7	63.7	64.2	62.7	65.6	62.7	63.7	62.8	54.1	56.4
X-BiLSTM-max	**73.7**	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4
Evaluation of pretrained multilingual sentence encoders (transfer learning)															
X-CBOW	64.5	60.3	60.7	61.0	60.5	60.4	57.8	58.7	57.5	58.8	56.9	58.8	56.3	50.4	52.2

Example Task: Multilingual Zero shot classification

XNLI [Conneau et al. 2018]:

- Initialize from multilingual pretrained model
- Fine-tune on English data.
- Evaluate on 14 other languages.

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
Machine translation baselines (TRANSLATE TRAIN)															
BiLSTM-last	71.0	66.7	67.0	65.7	65.3	65.6	65.1	61.9	63.9	63.1	61.3	65.7	61.3	55.2	55.2
BiLSTM-max	**73.7**	68.3	68.8	66.5	66.4	67.4	66.5	64.5	65.8	66.0	62.8	67.0	62.1	58.2	56.6
Machine translation baselines (TRANSLATE TEST)															
BiLSTM-last	71.0	68.3	68.7	66.9	67.3	68.1	66.2	64.9	65.8	64.3	63.2	66.5	61.8	60.1	58.1
BiLSTM-max	**73.7**	**70.4**	**70.7**	**68.7**	**69.1**	**70.4**	**67.8**	**66.3**	**66.8**	**66.5**	**64.4**	**68.3**	**64.2**	**61.8**	**59.3**
Evaluation of XNLI multilingual sentence encoders (in-domain)															
X-BiLSTM-last	71.0	65.2	67.8	66.6	66.3	65.7	63.7	64.2	62.7	65.6	62.7	63.7	62.8	54.1	56.4
X-BiLSTM-max	**73.7**	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4
Evaluation of pretrained multilingual sentence encoders (transfer learning)															
X-CBOW	64.5	60.3	60.7	61.0	60.5	60.4	57.8	58.7	57.5	58.8	56.9	58.8	56.3	50.4	52.2

Uses machine translation

Example Task: Multilingual Zero shot classification

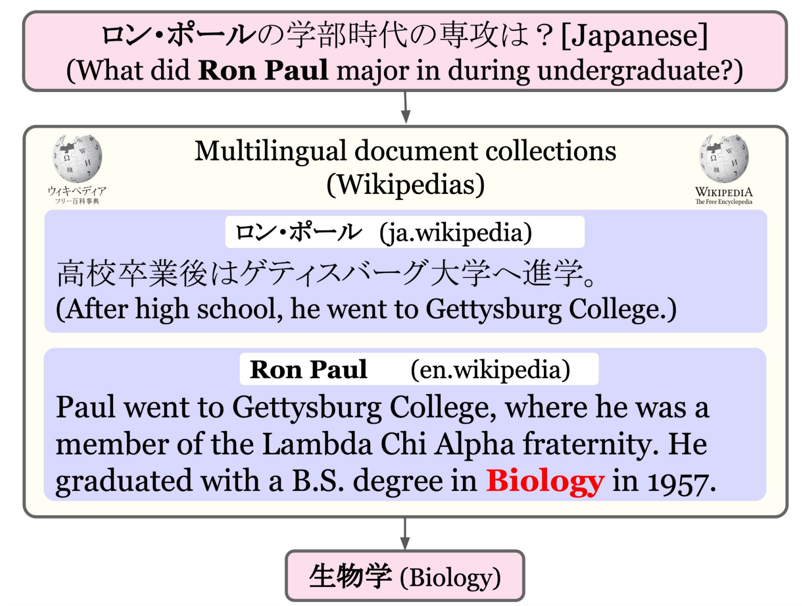
XNLI [Conneau et al. 2018]:

- Initialize from multilingual pretrained model
- Fine-tune on English data.
- Evaluate on 14 other languages.

	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur
Machine translation baselines (TRANSLATE TRAIN)															
BiLSTM-last	71.0	66.7	67.0	65.7	65.3	65.6	65.1	61.9	63.9	63.1	61.3	65.7	61.3	55.2	55.2
BiLSTM-max	**73.7**	68.3	68.8	66.5	66.4	67.4	66.5	64.5	65.8	66.0	62.8	67.0	62.1	58.2	56.6
Machine translation baselines (TRANSLATE TEST)															
BiLSTM-last	71.0	68.3	68.7	66.9	67.3	68.1	66.2	64.9	65.8	64.3	63.2	66.5	61.8	60.1	58.1
BiLSTM-max	**73.7**	**70.4**	**70.7**	**68.7**	**69.1**	**70.4**	**67.8**	**66.3**	**66.8**	**66.5**	**64.4**	**68.3**	**64.2**	**61.8**	**59.3**
Evaluation of XNLI multilingual sentence encoders (in-domain)															
X-BiLSTM-last	71.0	65.2	67.8	66.6	66.3	65.7	63.7	64.2	62.7	65.6	62.7	63.7	62.8	54.1	56.4
X-BiLSTM-max	**73.7**	67.7	68.7	67.7	68.9	67.9	65.4	64.2	64.8	66.4	64.1	65.8	64.1	55.7	58.4
Evaluation of pretrained multilingual sentence encoders (transfer learning)															
X-CBOW	64.5	60.3	60.7	61.0	60.5	60.4	57.8	58.7	57.5	58.8	56.9	58.8	56.3	50.4	52.2

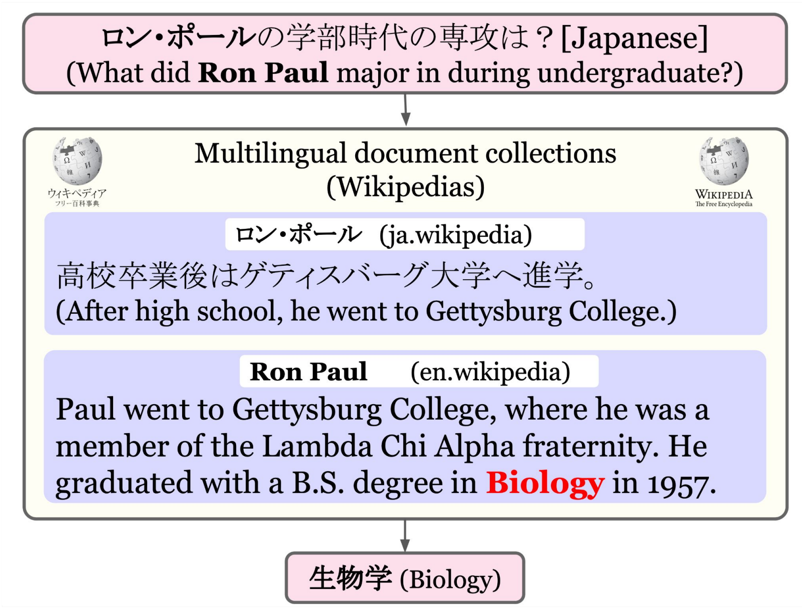
Doesn't use machine translation

Example Task: Cross-Lingual Question Answering [Asai et al. 2021, Muller et al. 2021]



	Human Translation			GMT		Our MT		*Multi.*
	DPR	PATH	BM	DPR	PATH	DPR	PATH	DPR
Ar	68.3	**70.0**	41.6	67.5	63.3	52.5	51.6	50.4
Bn	**85.6**	82.0	57.0	83.2	78.9	63.2	64.8	57.7
Fi	**73.1**	70.2	43.7	68.1	64.1	65.9	59.5	58.9
Ja	**68.9**	63.0	38.8	60.1	52.3	52.1	41.7	37.3
Ko	**70.9**	63.6	43.8	66.3	54.0	46.5	37.6	42.8
Ru	**65.2**	63.7	35.2	60.4	56.5	47.3	38.1	44.0
Te	**72.2**	64.1	44.6	65.0	62.5	22.7	18.1	44.9
Av.	**72.1**	68.1	43.5	67.2	61.7	50.0	44.5	48.0

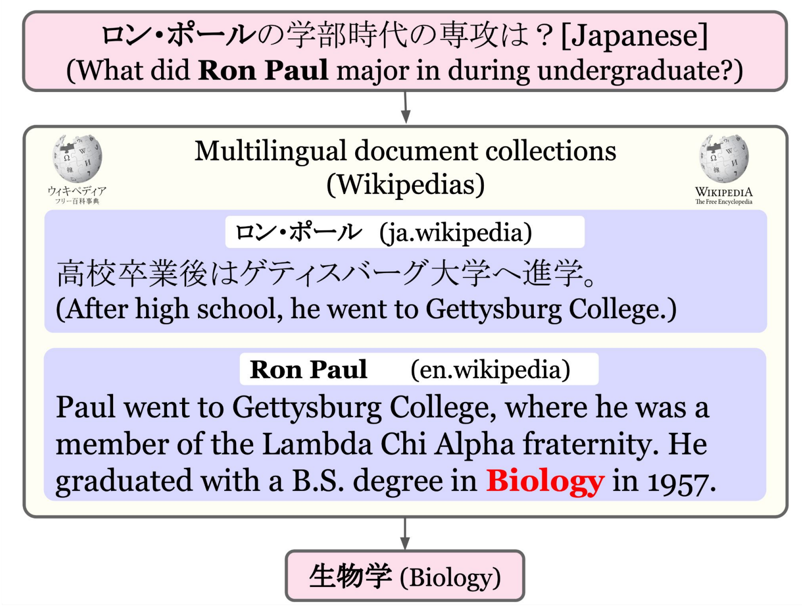
Example Task: Cross-Lingual Question Answering [Asai et al. 2021, Muller et al. 2021]



	Human Translation			GMT		Our MT		*Multi.*
	DPR	PATH	BM	DPR	PATH	DPR	PATH	DPR
Ar	68.3	**70.0**	41.6	67.5	63.3	52.5	51.6	50.4
Bn	**85.6**	82.0	57.0	83.2	78.9	63.2	64.8	57.7
Fi	**73.1**	70.2	43.7	68.1	64.1	65.9	59.5	58.9
Ja	**68.9**	63.0	38.8	60.1	52.3	52.1	41.7	37.3
Ko	**70.9**	63.6	43.8	66.3	54.0	46.5	37.6	42.8
Ru	**65.2**	63.7	35.2	60.4	56.5	47.3	38.1	44.0
Te	**72.2**	64.1	44.6	65.0	62.5	22.7	18.1	44.9
Av.	**72.1**	68.1	43.5	67.2	61.7	50.0	44.5	48.0

Uses machine translation

Example Task: Cross-Lingual Question Answering [Asai et al. 2021, Muller et al. 2021]



	Human Translation			GMT		Our MT		*Multi.*
	DPR	PATH	BM	DPR	PATH	DPR	PATH	DPR
Ar	68.3	**70.0**	41.6	67.5	63.3	52.5	51.6	50.4
Bn	**85.6**	82.0	57.0	83.2	78.9	63.2	64.8	57.7
Fi	**73.1**	70.2	43.7	68.1	64.1	65.9	59.5	58.9
Ja	**68.9**	63.0	38.8	60.1	52.3	52.1	41.7	37.3
Ko	**70.9**	63.6	43.8	66.3	54.0	46.5	37.6	42.8
Ru	**65.2**	63.7	35.2	60.4	56.5	47.3	38.1	44.0
Te	**72.2**	64.1	44.6	65.0	62.5	22.7	18.1	44.9
Av.	**72.1**	68.1	43.5	67.2	61.7	50.0	44.5	48.0

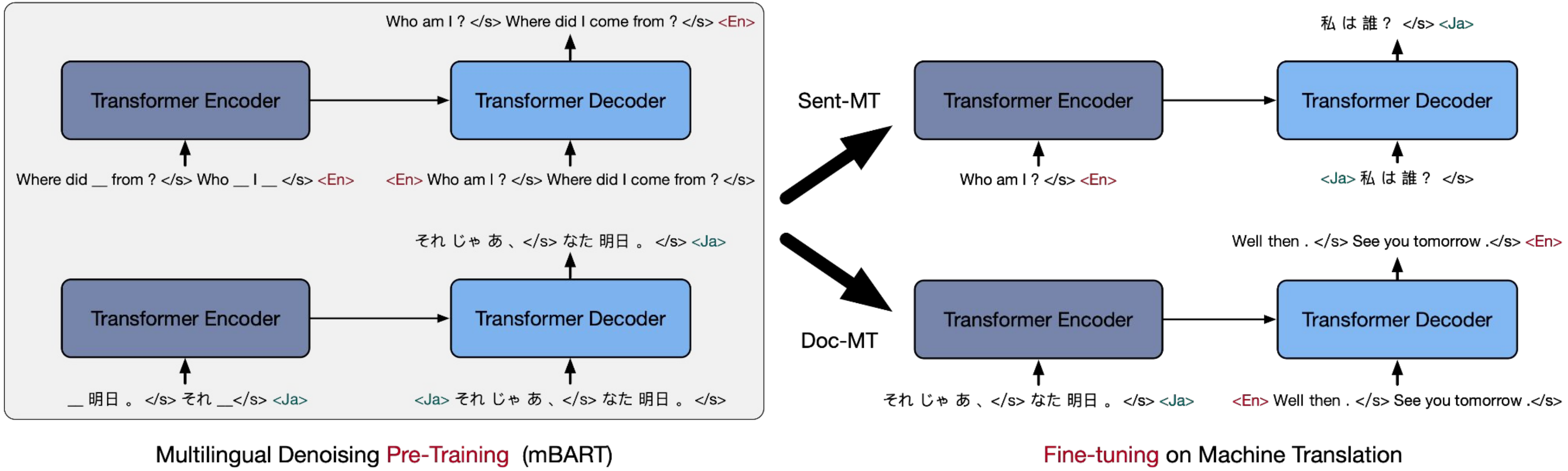
Doesn't use machine translation

Applications to Multilinguality

- Introduction
- Multilingual Pre-training
- Back-Translation for Machine Translation
- Zero shot Translation
- Unsupervised Machine Translation

Multilingual Pretraining [Conneau et al. 2020, Liu et al. 2020, Xue et al. 2020, Chung et al. 2021]

Use span-denoising objectives on monolingual data from various languages.



Source: Liu et al. 2020

Large Gains in Zero-shot classification

Setup:

- Initialize from multilingual pretrained model
- Fine-tune on English data.
- Evaluate on 14 other languages.

Results on XNLI [Conneau et al. 2020]

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
Lample and Conneau (2019)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
Huang et al. (2019)	Wiki+MT	N	15	85.1	79.0	79.4	77.8	77.2	77.2	76.3	72.8	73.5	76.4	73.6	76.2	69.4	69.7	66.7	75.4
Devlin et al. (2018)	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
Lample and Conneau (2019)	Wiki	N	100	83.7	76.2	76.6	73.7	72.4	73.0	72.1	68.1	68.4	72.0	68.2	71.5	64.5	58.0	62.4	71.3
Lample and Conneau (2019)	Wiki	1	100	83.2	76.7	77.7	74.0	72.7	74.1	72.7	68.7	68.6	72.9	68.9	72.5	65.6	58.2	62.4	70.7
XLM-R$_{\text{Base}}$	CC	1	100	85.8	79.7	80.7	78.7	77.5	79.6	78.1	74.2	73.8	76.5	74.6	76.7	72.4	66.5	68.3	76.2
XLM-R	CC	1	100	**89.1**	**84.1**	**85.1**	**83.9**	**82.9**	**84.0**	**81.2**	**79.6**	**79.8**	**80.8**	**78.1**	**80.2**	**76.9**	**73.9**	**73.8**	**80.9**

Enables zero-shot evaluation metrics for generation [Sellam et al. 2020]

Setup:

- Initialize from multilingual pretrained model
- Fine-tune on WMT ratings data for X language pairs
- Evaluate on Y other language pairs for which ratings data has not been seen.

	en-cs	en-de	en-fi	*en-gu*	*en-kk*	*en-lt*	en-ru	en-zh	de-cs	*de-fr*	fr-de	**avg**
YISI1	0.475	0.351	0.537	0.551	0.546	0.470	0.585	0.355	0.376	0.349	0.310	0.446
YISI1-SRL	-	0.368	-	-	-	-	-	0.361	-	-	0.299	-
ESIM	-	0.329	0.511	-	0.510	0.428	0.572	0.339	0.331	0.290	0.289	-
BERTSCORE	0.485	0.345	0.524	0.558	0.533	0.463	0.580	0.347	0.352	0.325	0.274	0.435
PRISM	0.582	**0.426**	0.591	0.313	0.531	0.558	0.584	0.376	0.458	**0.453**	0.426	0.482
BLEURT Configurations												
BERT-CHINESE-L2	-	-	-	-	-	-	-	0.356	-	-	-	-
MBERT	0.506	0.364	0.551	0.550	0.529	0.516	**0.592**	**0.381**	0.385	0.388	0.291	0.459
MBERT-WMT	**0.603**	0.422	**0.615**	**0.577**	**0.558**	**0.584**	0.492	0.337	**0.461**	0.449	**0.427**	**0.502**

zero shot languages

Shortcomings

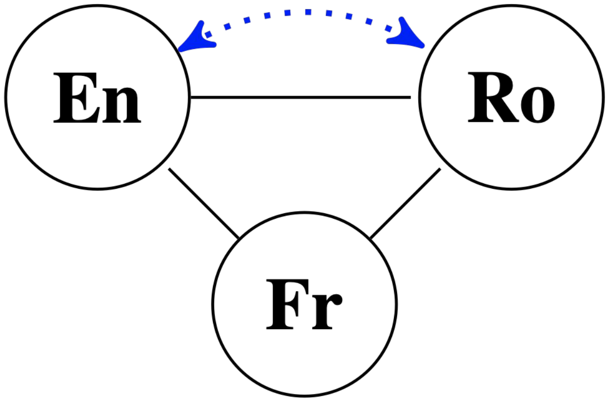
While pre-training enables a lot of amazing things it does not explicitly force similar words/phrases in different languages to have similar representations.

Thus it may not sufficient for certain challenging applications e.g. translating into low resource languages

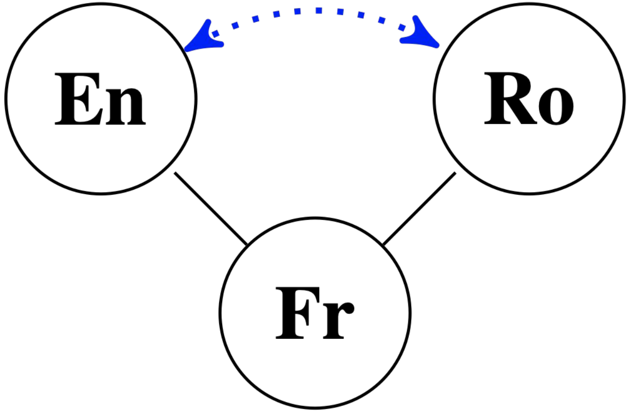
Applications to Multilinguality

- Introduction
- Multilingual Pre-training
- Back-Translation for Machine Translation
- Zero shot Translation
- Unsupervised Machine Translation

Neural Machine Translation Setups



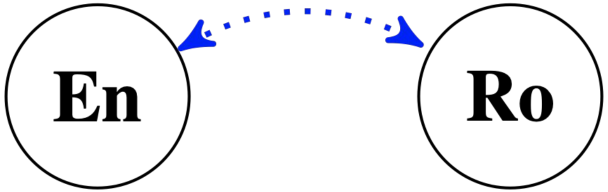
Supervised (Multilingual) Translation
[Johnson et al. 2016,
Firat et al. 2016]



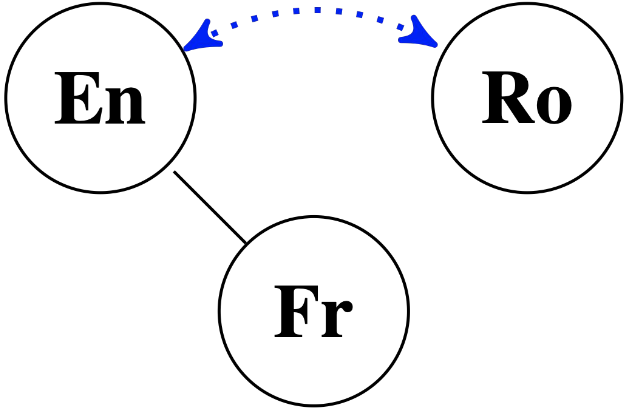
Zero shot translation
[Johnson et al. 2016, Chen et al. 2017,
Cheng et al. 2017, Al-Shedivat and Parikh 2019]

Solid lines indicate presence of parallel data

Neural Machine Translation Setups



Unsupervised translation [Ravi and Knight 2011, Lample et al. 2018, Artexe et al. 2018]



Multilingual Unsupervised Translation [Siddhant et al. 2020, Garcia et al. 2020, Li et al. 2020, Wang et al. 2021, Garcia et al. 2021]

Solid lines indicate presence of parallel data

Back-Translation (for Supervised MT) [Sennrich et al. 2015]

Let *x* be a sentence in the source language and *X* the set of all source sentences.

Let *y* be a sentence in the target language and *Y* the set of all target sentences.

Forward (supervised) translation model: $p_\theta(y|x)$

Backward (supervised) translation model: $p_\phi(x|y)$

Back-Translation

Back-translation is a form of data augmentation where

- *f* generates synthetic data for *g*
- *g* generates synthetic data for *f*

This can significantly increase the performance of models especially in the case where either the source or target is low resource.

Back-Translation: A form of data augmentation

f generates synthetic data for *g:* Given source sentences $x^{(1)}, ..., x^{(n)}$

$$\hat{y}^{(i)} = \text{argmax}_y p_\theta(y|x^{(i)}) \quad \forall 1 \leq i \leq n$$

Synthetic dataset:

$$(\hat{y}^{(1)}, x^{(1)}), ..., (\hat{y}^{(n)}, x^{(n)})$$

Continue training *g* on synthetic dataset using maximum likelihood

$$\phi \leftarrow \text{argmax}_\phi \sum_{i=1}^{n} \log p_\phi(x^{(i)}|\hat{y}^{(i)})$$

Back-Translation: A form of data augmentation

g generates synthetic data for *f*: Given source sentences $y^{(1)}, ..., y^{(m)}$

$$\hat{x}^{(j)} = \text{argmax}_y p_\phi(x|y^{(j)}) \quad \forall 1 \leq j \leq m$$

Synthetic dataset:

$$(\hat{x}^{(1)}, y^{(1)}), ..., (\hat{x}^{(m)}, y^{(m)})$$

Continue training *f* on synthetic dataset using maximum likelihood

$$\theta \leftarrow \text{argmax}_\theta \sum_{j=1}^{m} \log p_\theta(y^{(j)}|\hat{x}^{(j)})$$

When does Back-translation help?

Let X be Romanian and Y be English.

The performance of the forward (Ro -> En) model is likely to be better so generating synthetic data with the forward model can yield high quality training data for the backward (En -> Ro) model

Thus back-translation can greatly help translating into low resource languages.

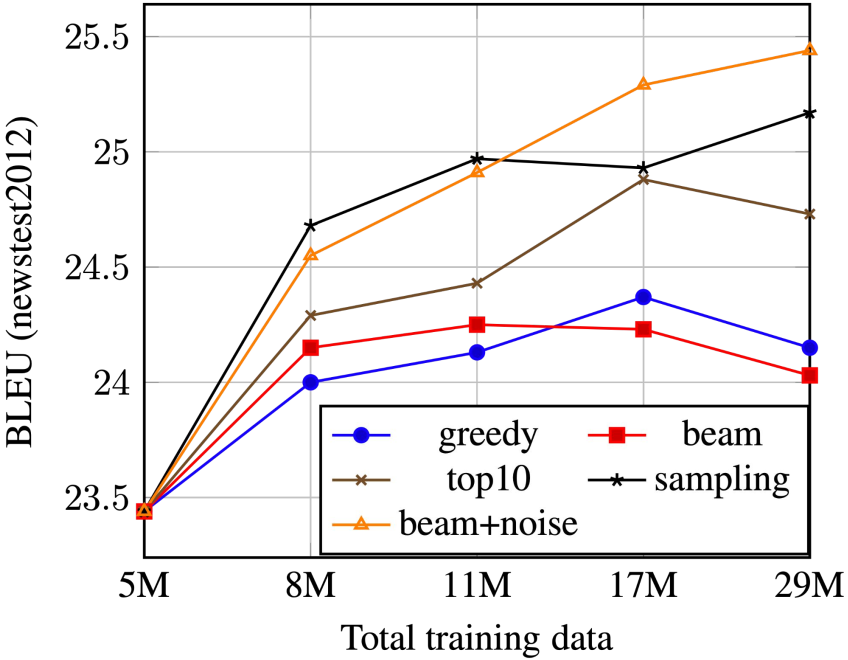
Understanding Back-Translation at Scale [Edunov et al. 2018]

Experimented with different decoding strategies of generating the synthetic data.

source	Diese gegenstzlichen Auffassungen von Fairness liegen nicht nur der politischen Debatte zugrunde.
reference	These competing principles of fairness underlie not only the political debate.
beam	These conflicting interpretations of fairness are not solely based on the political debate.
sample	*Mr President*, these contradictory interpretations of fairness are not based solely on the political debate.
top10	Those conflicting interpretations of fairness are not solely at the heart of the political debate.
beam+noise	conflicting BLANK interpretations BLANK are of not BLANK based on the political debate.

Understanding Back-Translation at Scale [Edunov et al. 2018]

Consistent improvements across tasks.

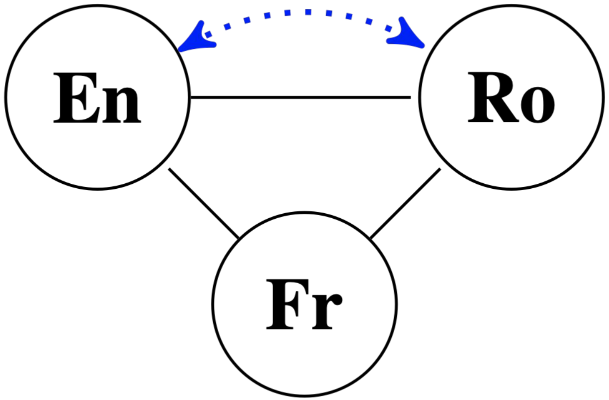


	En–De	En–Fr
a. Gehring et al. (2017)	25.2	40.5
b. Vaswani et al. (2017)	28.4	41.0
c. Ahmed et al. (2017)	28.9	41.4
d. Shaw et al. (2018)	29.2	41.5
DeepL	33.3	**45.9**
Our result	**35.0**	45.6
detok. sacreBLEU[3]	*33.8*	*43.8*

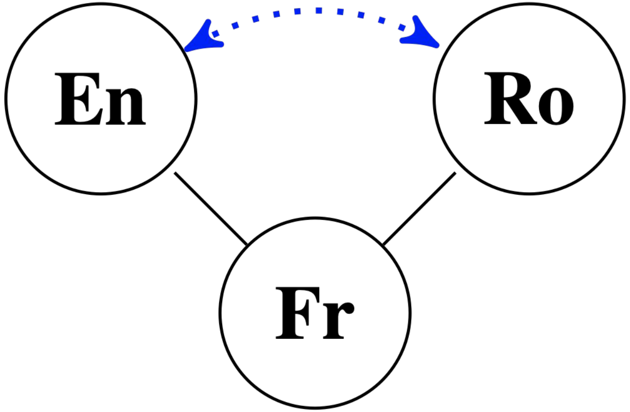
Applications to Multilinguality

- Introduction
- Multilingual Pre-training
- Back-Translation for Machine Translation
- Zero shot Translation
- Unsupervised Machine Translation

Neural Machine Translation Setups



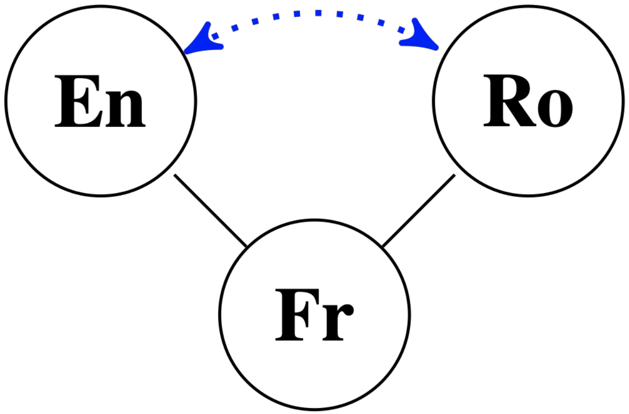
Supervised (Multilingual) Translation
[Johnson et al. 2016,
Firat et al. 2016]



Zero shot translation
[Johnson et al. 2016, Chen et al. 2017,
Cheng et al. 2017, Al-Shedivat and Parikh 2019]

Solid lines indicate presence of parallel data

Synthetic Data Generation (Distillation) for Zero-Shot Translation [Chen et al. 2017]



- Train supervised (*En*, *Fr*) model f and (*Fr*, *Ro*) model g
- Use g to label (*En*, *Fr*) data to generate synthetic (*En*, *Ro*) data
- Train (*Er*, *Ro*) model

Encoder Consistency for Zero Shot Translation [Arivazhagan et al. 2019]

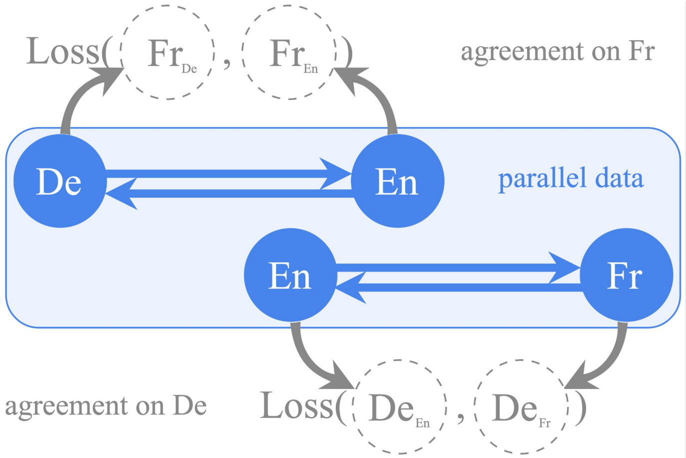
$$\mathrm{Loss}_{En,Fr} = -\log P(x_{Fr}|x_{En}) - \log P(x_{En}|x_{Fr}) \\ - \underline{\mathrm{similarity}}(\mathrm{Enc}(x_{Fr}), \mathrm{Enc}(x_{En}))$$

Similarity function like cosine similarity
on pooled encoder states

Regularize encoder output to be language-invariant

Decoder Consistency for Zero Shot Translation

[Al-Shedivat and Parikh 2019]



$$\mathrm{Loss}_{En,Fr} = -\log P(x_{Fr}|x_{En}) - \log P(x_{En}|x_{Fr}) - \log \sum_z P(z|x_{En})P(z|x_{Fr})$$

- Model sees (En, De) and (En, Fr) supervised pairs in training.

- For each (En, Fr) example also try to translate to a third language e.g.
  - De_{En}: the predicted *De* translation from *En*
  - De_{Fr}: the predicted *De* translation from *Fr*

Have regularizer that enforces agreement between De_{En} and De_{Fr}

(analogously for each (En, De) example)

Zero Shot Results [Al-Shedivat and Parikh 2019]

	Previous work		Our baselines		
	Soft‡	Distill†	Basic	Pivot	Agree
En → Es	—	—	34.69	34.69	33.80
En → De	—	—	23.06	23.06	22.44
En → Fr	31.40	—	33.87	33.87	32.55
Es → En	31.96	—	34.77	34.77	34.53
De → En	26.55	—	29.06	29.06	29.07
Fr → En	—	—	33.67	33.67	33.30
Supervised (avg.)	—	—	31.52	31.52	30.95
Es → De	—	—	18.23	20.14	**20.70**
De → Es	—	—	20.28	**26.50**	22.45
Es → Fr	30.57	**33.86**	27.99	32.56	30.94
Fr → Es	—	—	27.12	**32.96**	29.91
De → Fr	23.79	**27.03**	21.36	25.67	24.45
Fr → De	—	—	18.57	**19.86**	19.15
Zero-shot (avg.)	—	—	22.25	26.28	24.60

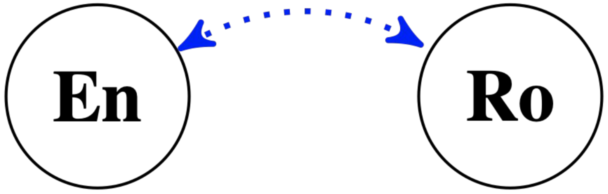
† Soft pivoting (Cheng et al., 2017). ‡ Distillation (Chen et al., 2017).

- Pivoting (i.e. translating twice - first to English and then out English) is a strong baseline.
- Decoder agreement works well for a multilingual model, however, distillation works better.

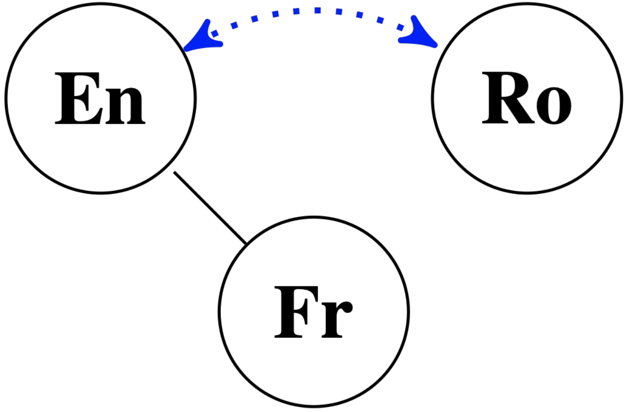
Applications to Multilinguality

- Introduction
- Multilingual Pre-training
- Back-Translation for Machine Translation
- Zero shot Translation
- Unsupervised Machine Translation

Neural Machine Translation Setups



Unsupervised translation [Ravi and Knight 2011, Lample et al. 2018, Artexe et al. 2018]



Multilingual Unsupervised Translation [Siddhant et al. 2020, Garcia et al. 2020, Li et al. 2020, Wang et al. 2021, Garcia et al. 2021]

Solid lines indicate presence of parallel data

Unsupervised Translation [Ravi and Knight 2011, Lample et al. 2018, Artexe et al. 2018]

Given monolingual data in two languages learn a mapping between them. Training examples are unaligned.

Spanish

Ciro II el Grande

Ciro II el Grande (*circa* 600/575 – 530 a. C.) fue un rey aqueménida de Persia (*circa* 559-530 a. C.) y Imperio aqueménida (en persa antiguo: Haxā Imperio persa, luego de vencer a Astiages, últ (550 a. C.) y extendió su dominio por la mese y gran parte de Mesopotamia. Sus conquistas sobre Media, Lidia y Babilonia, desde el mar M hasta la cordillera del Hindu Kush, con lo que imperio conocido hasta ese momento. Este du doscientos años, hasta su conquista final por

Juana de Arco

Para otros usos de este término, véase Juana de Arco (desambiguación).

Juana de Arco (en francés: *Jeanne d'Arc*), [b] también conocida como la **Doncella de Orleans** (en francés: *La Pucelle d'Orléans*; Domrémy, h. 1412-Ruan, 30 de mayo de 1431),[c] fue una joven campesina que es considerada una heroína de Francia por su papel durante la fase final de la Guerra de los Cien Años. Juana afirmó haber tenido visiones del Arcángel Miguel, de Santa Margarita y de Catalina de Alejandría, quienes le dieron instrucciones para que ... inglesa en el ... todavía no había ... egrante de un ... rque el asedio fue ... permitieron que ... ento tan esperado ... nal.

Atenas

Para otros usos de este término, véase Atenas (desambiguación).

Atenas (griego antiguo: Ἀθῆναι, romanización: *Athênai*, griego moderno: Αθήνα, romanización: *Athína*) es la capital de Grecia y actualmente la ciudad más grande, importante y poblada del país. La población del municipio de Atenas era de 664 046 (en 2011), pero su área metropolitana es mucho mayor y comprende una población de 3,8 millones (en 2011). Es el centro principal de la vida económica, cultural y política griega.

German

Napoleon Bonaparte

Napoleon ist eine Weiterleitung auf diesen Artikel. Weitere Bedeutungen sind aufgeführt.

Napoleon Bonaparte, als Kaiser **Napo** *Bonaparte* bzw. *Napoléon Ier*; * 15. Aug **Napoleone Buonaparte**[1]; † 5. Mai 18 im Südatlantik), war ein französischer G Kaiser der Franzosen.

Aus korsischer Familie stammend, stieg

Berlin

Der Titel dieses Artikels ist mehrdeutig. Weitere Bedeutungen sind unter

Berlin ([bɛʁˈliːn]) ist die Hauptstadt und ein Land der Bundesrepublik Deutschland.[14] Die Stadt ist mit rund 3,7 Millionen Einwohnern die bevölkerungsreichste und mit 892 Quadratkilometern die flächengrößte

Aztekenreich

Das **Aztekenreich** entstand aus dem Aztekischen Dreibund der drei Stadtstaaten Tenochtitlan, Texcoco und Tlacopan im heutigen Mexiko, welcher seine Wurzeln auf das Jahr 1428 zurückführt. Diese drei Stadtstaaten unter Führung von Tenochtitlan (gegründet 1325) beherrschten das Gebiet im und um das Tal von Mexiko und errichteten ein bedeutendes Reich, welches bis zu seiner Eroberung durch spanische Konquistadoren und ihrer einheimischen Verbündeten unter Hernán Cortés 1521 existierte. Mit einer geschätzten Bevölkerung von 150.000 bis 200.000 Menschen zählte Tenochtitlan, die de facto Hauptstadt der Azteken, um das Jahr 1500 zu den größten Städten der Welt.[1] Die Bevölkerung des gesamten Reiches wurde auf 5 bis 6 Millionen geschätzt.[2]

Unsupervised Neural Machine Translation [UNMT] [Lample et al. 2018, Artexe et al. 2018, Lample et al. 2018, Song et al. 2019]

Step 1: Train a pretrained language model based on monolingual data in the source and target languages (with span denoising)

<es> Ciro II el Grande (circa 600/575 – 530 a. C.) fue un rey aqueménida de Persia (circa 559-530 a. C.) y el fundador del Imperio aqueménida ….

<de> Napoleon Bonaparte, als Kaiser Napoleon I. (französisch Napoléon Bonaparte bzw. Napoléon Ier; * 15. August 1769 in Ajaccio auf Korsika als Napoleone Buonaparte[1]; † 5. Mai 1821 in Longwood House auf St. Helena im Südatlantik), war ein französischer General, revolutionärer Diktator und Kaiser der Franzosen.

<es> Juana de Arco (en francés: Jeanne d'Arc), btambién conocida como la Doncella de Orleans (en francés: La Pucelle d'Orléans; Domrémy, h. 1412-Ruan, 30 de mayo de 1431) …..

<de> Das Aztekenreich entstand aus dem Aztekischen Dreibund der drei Stadtstaaten Tenochtitlan, Texcoco und Tlacopan im heutigen Mexiko, welcher seine Wurzeln auf das Jahr 1428 zurückführt.

Unsupervised Neural Machine Translation [UNMT]

[Lample et al. 2018, Artexe et al. 2018, Lample et al. 2018, Song et al. 2019]

Step 2: Use online back-translation:

SG = *stop gradient i.e. since the gradient won't pass through the argmax*

$$\hat{z}_{\mathrm{de}} = \mathrm{SG}(\mathrm{argmax}_z p_\theta(z|x_{es}))$$

- Compute the most likely translation to *de* (using the <de> tag so that the decoded output is in the right language).

$$\mathrm{loss}_{\mathrm{es}} = -\log p_\theta(x_{\mathrm{es}}|\hat{z}_{\mathrm{de}})$$

- Maximizing the likelihood of the original *es* sentence under the backward translation model.

Unsupervised Neural Machine Translation [UNMT]

[Lample et al. 2018, Artexe et al. 2018, Lample et al. 2018, Song et al. 2019]

Step 2: Use online back-translation:

Do the same for the other direction

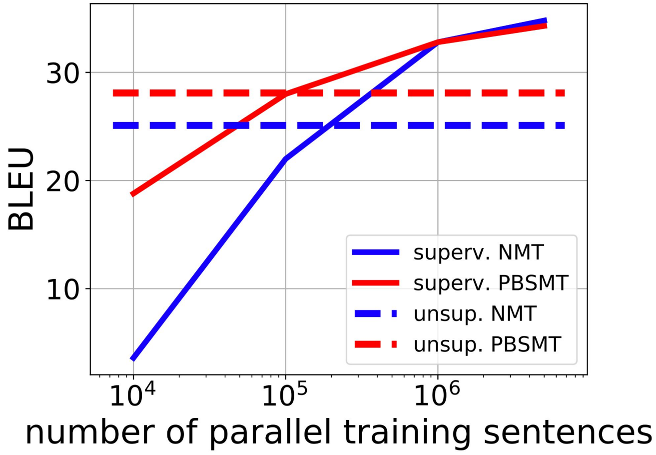
$$\hat{z}_{\text{es}} = \text{SG}(\text{argmax}_z p_\theta(z|x_{de}))$$

$$\text{loss}_{\text{de}} = -\log p_\theta(x_{\text{de}}|\hat{z}_{\text{es}})$$

- Compute the most likely translation to *es* (using the <es> tag so that the decoded output is in the right language).
- Maximizing the likelihood of the original *de* sentence under the backward translation model.

Results

Lample et al. 2018



Song et al. 2019

Method	Setting	en - fr	fr - en	en - de	de - en	en - ro	ro - en
Artetxe et al. (2017)	2-layer RNN	15.13	15.56	6.89	10.16	-	-
Lample et al. (2017)	3-layer RNN	15.05	14.31	9.75	13.33	-	-
Yang et al. (2018)	4-layer Transformer	16.97	15.58	10.86	14.62	-	-
Lample et al. (2018)	4-layer Transformer	25.14	24.18	17.16	21.00	21.18	19.44
XLM (Lample & Conneau, 2019)	6-layer Transformer	33.40	33.30	27.00	34.30	33.30	31.80
MASS	6-layer Transformer	**37.50**	**34.90**	**28.30**	**35.20**	**35.20**	**33.10**

But what about a real use case? [Guzmán et al. 2019, Marchisio et al. 2020, Kim et al. 2020]

The results are much worse when running on actual low resource languages.

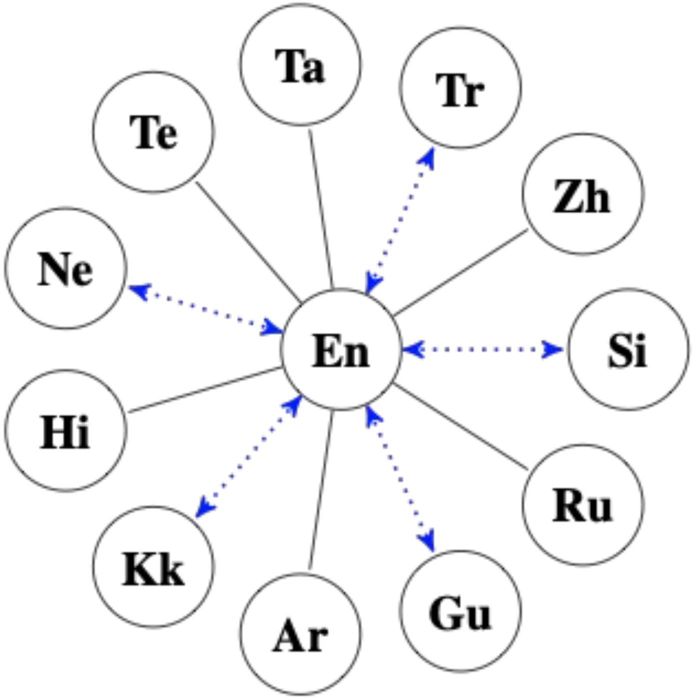
	BLEU [%]									
Approach	**de-en**	**en-de**	**ru-en**	**en-ru**	**zh-en**	**en-zh**	**kk-en**	**en-kk**	**gu-en**	**en-gu**
Supervised	39.5	39.1	29.1	24.7	26.2	39.6	10.3	2.4	9.9	3.5
Semi-supervised	43.6	41.0	30.8	28.8	25.9	42.7	12.5	3.1	14.2	4.0
Unsupervised	23.8	20.2	12.0	9.4	1.5	2.5	2.0	0.8	0.6	0.6

Multilingual Unsupervised Neural Machine Translation (M-UNMT)

[Siddhant et al. 2020, Garcia et al. 2020, Li et al. 2020, Wang et al. 2021, Garcia et al. 2021]

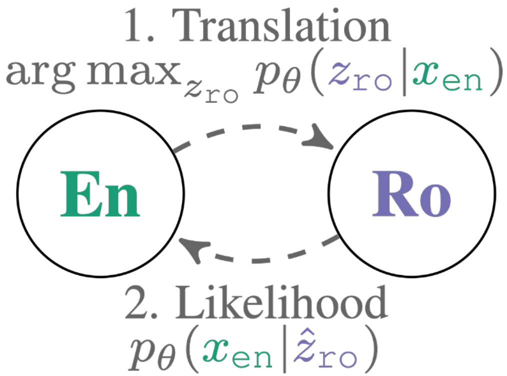
Leverage parallel data among auxiliary high resource pairs (solid lines)

Note that the target low resource languages are not associated with any parallel data

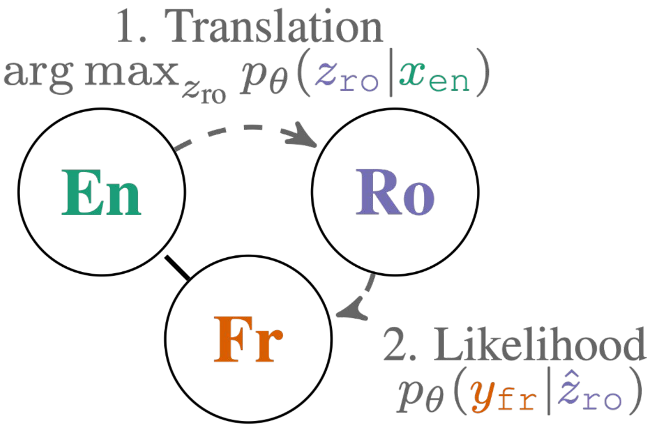


Cross-translation [Ren et al. 2018, Garcia et al. 2020, Li et al. 2020, Wang et al. 2021]

When more than one language is present, we can have a variant of back-translation called *cross-translation*



(a) Back-translation



(b) Cross-translation

Results [Garcia et al. 2021]

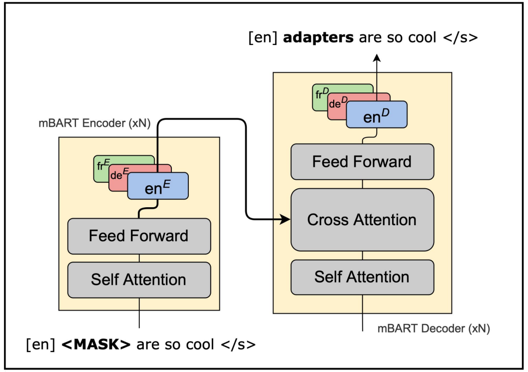
Multilinguality allows to achieve SOTA over existing unsupervised methods and close gap to supervised systems.

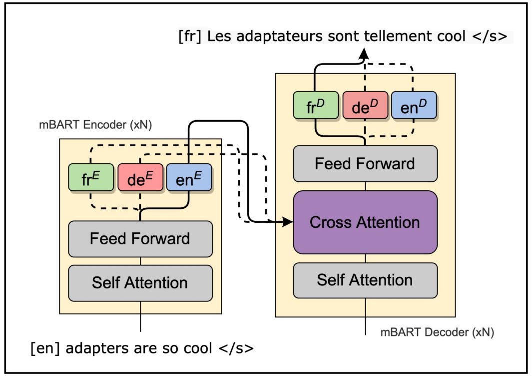
	Model	*FLoRes devtest* Ne ↔ En		*FLoRes devtest* Si ↔ En	
No parallel data	Guzmán et al. (2019)	0.1	0.5	0.1	0.1
No parallel data with{Ne,Si}	Liu et al. (2020)	-	17.9	-	9.0
	Guzmán et al. (2019)	8.3	18.3	0.1	0.1
	Stage 1 (Ours)	3.3	18.3	1.4	11.5
	Stage 2 (Ours)	8.6	20.8	7.7	15.7
	Stage 3 (Ours)	**8.9**	**21.7**	**7.9**	**16.2**
With parallel data for{Ne,Si}	*Mult. MT Baseline* (Ours)	8.6	20.1	7.6	15.3
	Liu et al. (2020)	9.6	21.3	9.3	20.2
	Guzmán et al. (2019)	8.8	21.5	6.5	15.1

M-UNMT without Back-Translation [Üstün et al. 2021]

- Using back-translation is expensive. Instead use adapter modules [Houlsby et al. 2019]

- Procedure:
  - Initialize with pretrained multilingual encoder-decoder (mBART)
  - Add adapter modules and pretrain only those on monolingual data for all languages.
  - Freeze adapter modules and fine-tune cross attention of encoder-decoder on parallel data





M-UNMT without Back-Translation [Üstün et al. 2021]

No back-translation

		$en \rightarrow zz$											
		es	nl	hr	uk	sv	lt	id	fi	et	ur	kk	AVG-11
(1)	BILINGUAL	**40.3**	**32.8**	**27.6**	**19.9**	**31.5**	13.2	21.4	9.5	6.8	2.4	0.4	**19.5**
(2)	MBART-FT	1.3	1.9	1.8	0.8	1.6	0.9	1.6	1.4	0.7	0.6	0.4	1.3
	TASK ADAPTERS	2.0	2.0	2.1	1.0	1.5	0.9	0.8	1.6	1.1	0.9	0.5	1.4
	DENOIS. ADAPTERS	28.4	21.6	19.0	12.2	22.9	11.0	23.8	10.1	12.7	9.6	3.8	15.9
(3)	MBART-FT (+BT)	30.9	22.0	20.0	14.2	22.7	13.7	20.2	9.4	14.1	5.7	3.5	16.3
	TASK ADAPTERS (+BT)	31.5	22.4	21.9	15.7	25.3	14.6	22.9	10.1	15.2	9.4	4.2	17.6
	DENOIS. ADAPT. (+BT)	32.2	22.9	23.1	15.4	27.1	**16.3**	**24.4**	**11.7**	**17.1**	**11.7**	**4.9**	18.9

back-translation

Outline

- [Introduction]: Overview (Colin)
- [Session 1]: Data Augmentation (Diyi)
- [Session 2]: Semi-supervised Learning (Colin)
- [Session 3]: Applications to Multilinguality (Ankur)
- [Conclusion]: Moving Forward (Diyi)

Discussion, Challenges, and Future Directions

Regarding Data Augmentation

- ❑ Theoretical Guarantees
- ❑ Data Distribution Shift
- ❑ Automatic Data Augmentation
- ❑ Selecting Labeled Data

Discussion, Challenges, and Future Directions

Regarding Semi-Supervised Learning

- ❑ Learning with noisy labels
- ❑ Large amount of unlabeled data
- ❑ Context specific consistency training
- ❑ Knowledge enhanced semi-supervised learning

Readings and References

- Edunov, Sergey, Myle Ott, Michael Auli, and David Grangier. "Understanding back-translation at scale." *arXiv preprint arXiv:1808.09381* (2018).
- Lample, Guillaume, and Alexis Conneau. "Cross-lingual language model pretraining." *arXiv preprint arXiv:1901.07291* (2019).
- Chen, Jiaao, Zichao Yang, and Diyi Yang. "Mixtext: Linguistically-informed interpolation of hidden space for semi-supervised text classification." *arXiv preprint arXiv:2004.12239* (2020).
- Morris, John X., Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. "Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp." *arXiv preprint arXiv:2005.05909* (2020).
- Chau, Ethan C., Lucy H. Lin, and Noah A. Smith. "Parsing with multilingual BERT, a small corpus, and a small treebank." *arXiv preprint arXiv:2009.14124* (2020).
- Du, Jingfei, Edouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Ves Stoyanov, and Alexis Conneau. "Self-training improves pre-training for natural language understanding." *arXiv preprint arXiv:2010.02194* (2020).
- Chen, Jiaao, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. "An empirical survey of data augmentation for limited data learning in nlp." *arXiv preprint arXiv:2106.07499* (2021).

ACL 2022
22ND – 27TH MAY | 60TH MEETING | DUBLIN

Data Augmentation and Semi-Supervised Learning for Natural Language Processing

Github: https://github.com/diyiy/ACL2022_Limited_Data_Learning_Tutorial

Questions: diyi.yang@cc.gatech.edu, aparikh@google.com, craffel@gmail.com