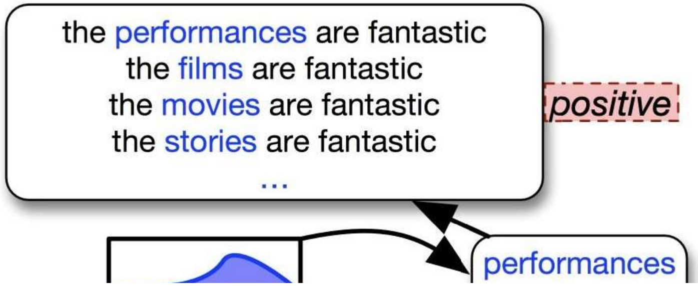
Data Augmentation for NLP

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



	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 a ugmented by replacing only *actors* with words predicted based on the context (Kobayashi, 2018)

Soft contextual data augmentation

(Gao et al., 2019)

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

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

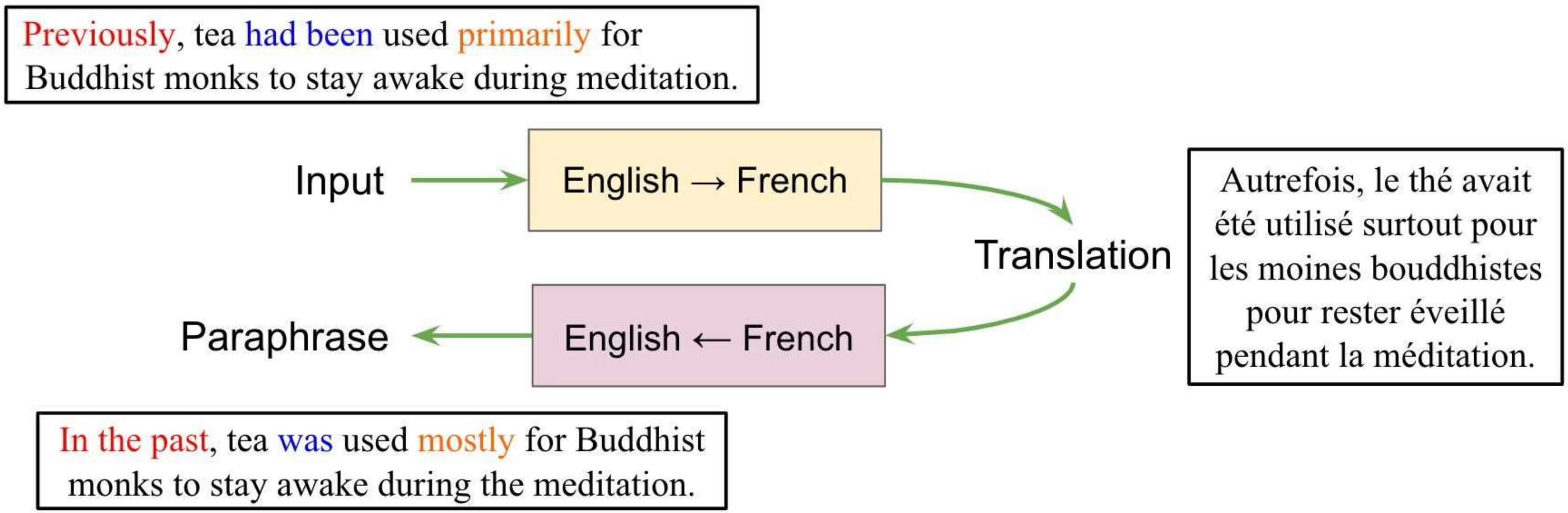


Image credit to https://github.com/vietai/dab

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 .

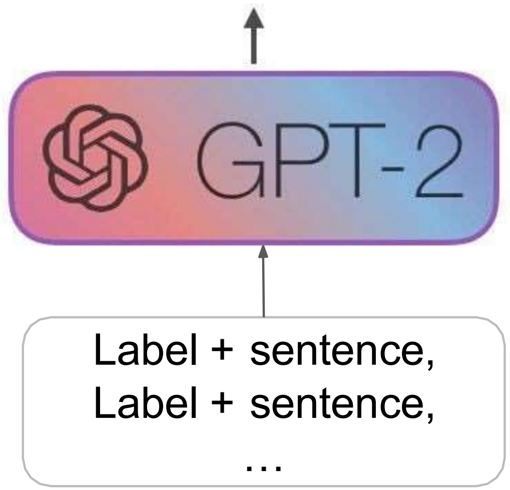
syntactically controlled paraphrase generation (Iyyer et al., 2018)

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



White-box Attack

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

(Ebrahimi et al., 2018)

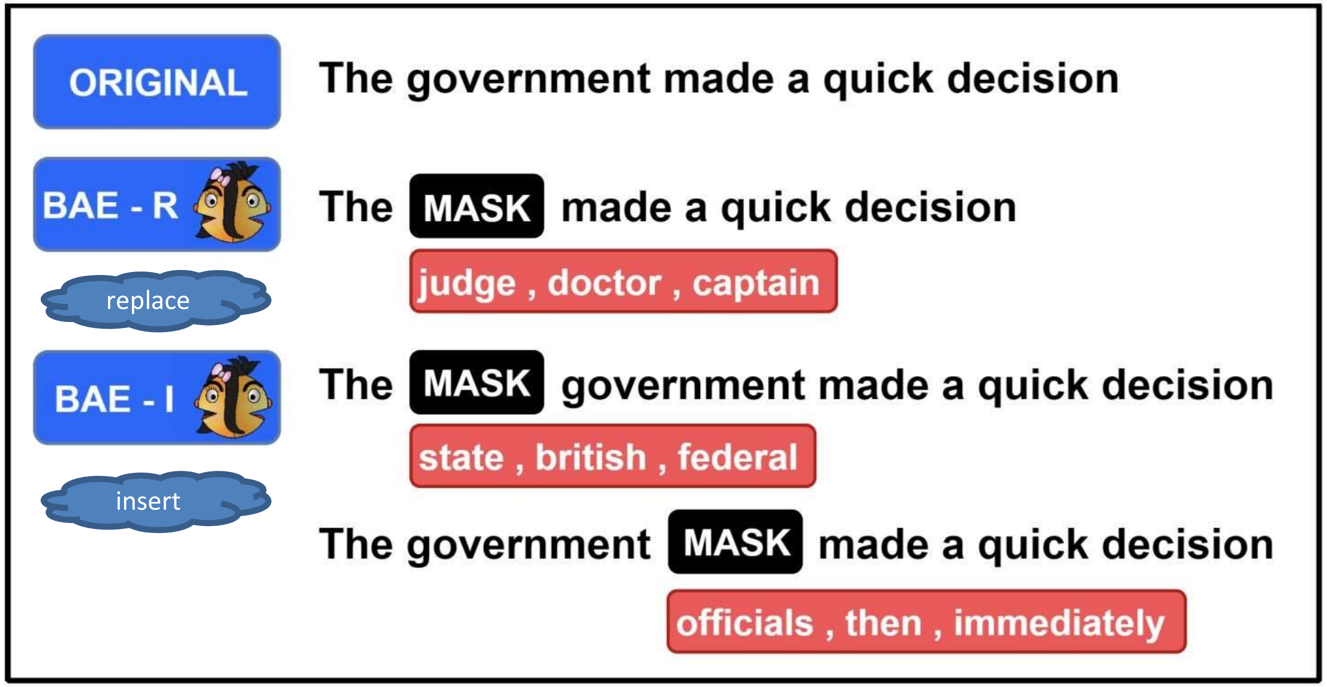
$$\max \nabla_x J(\mathbf{x}, \mathbf{y})^T \cdot \vec{v}_{ijb} = \max_{ijb} \frac{\partial J}{\partial x_{ij}}^{(b)} - \frac{\partial J}{\partial x_{ij}}^{(a)}$$

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)

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

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

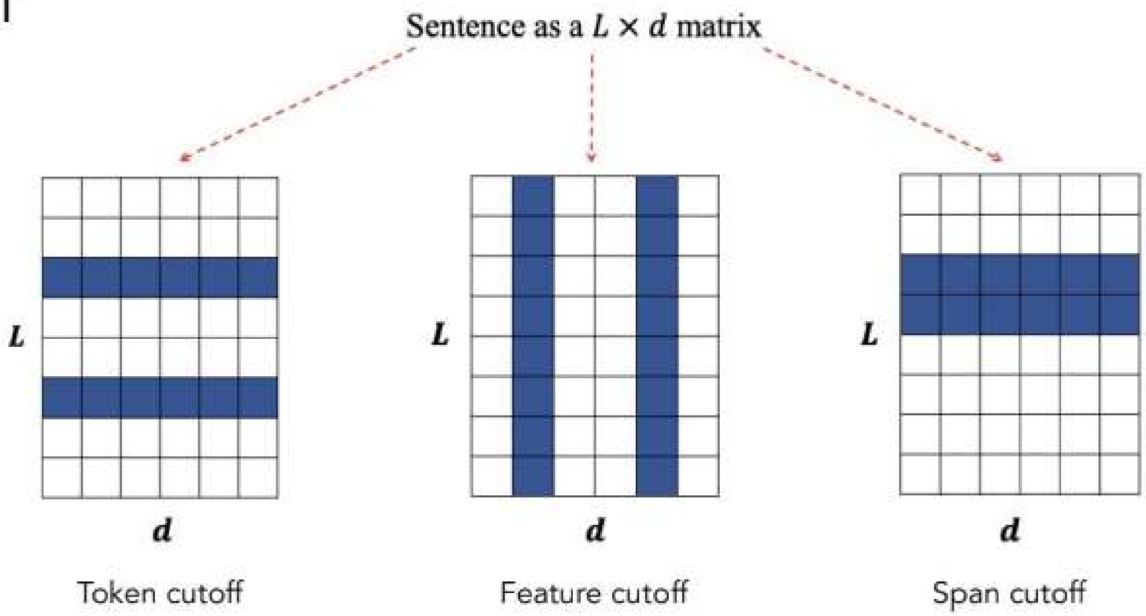
label for x' (e.g., sentiment, named entity)

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

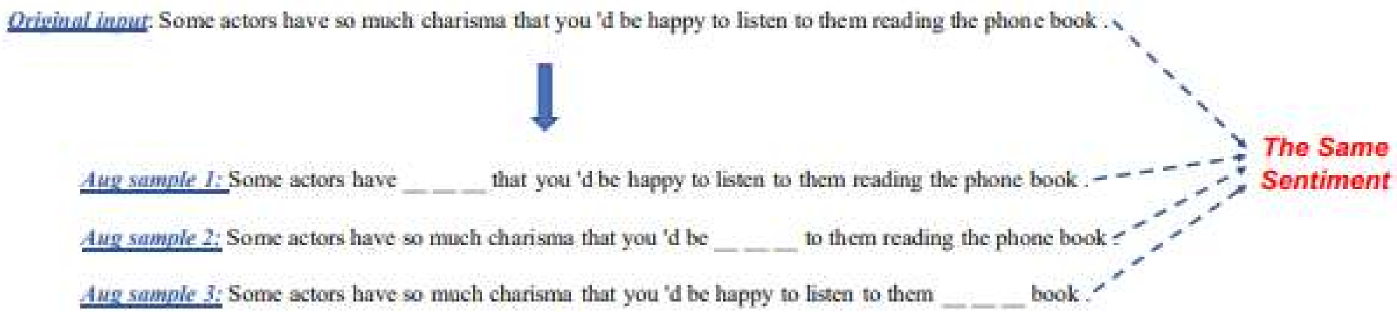
a given data point (e.g., sentence, token, subtree)

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

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).



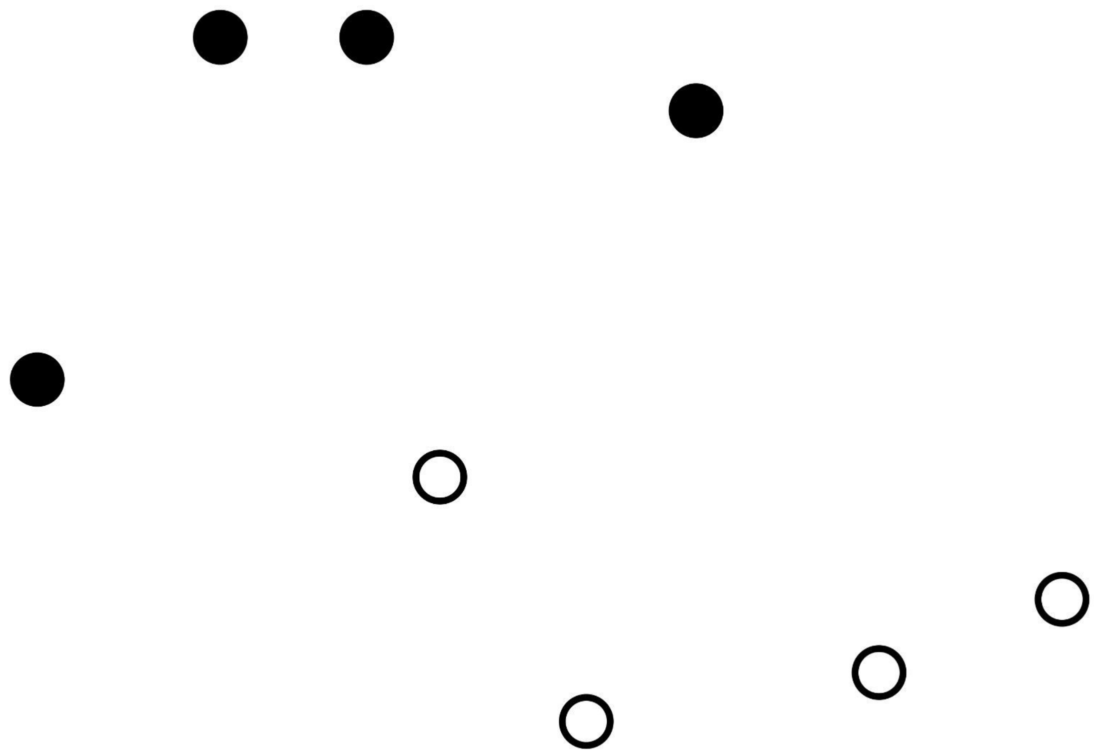
	Methods	Types	Inference			Paraphrase		Single Sentence	
			MNLI	QNLI	RTE	QQP	MRPC	SST-2	CoLA
Supervised	None	-	35.2(0.7)	51.8(7.0)	49.8(3.1)	63.9(9.1)	61.8(21.2)	60.5(13.1)	12.9(6.32)
	SR	Token	35.1(2.3)	51.4(7.2)	51.5(3.4)	61.3(9.7)	59.7(26.3)	62.1(17.4)	7.2(11.6)
	LM	Token	35.3(0.8)	51.0(8.0)	49.0(1.4)	62.4(11)	61.0(24.3)	62.8(9.8)	6.8(15.8)
	RI	Token	34.9(2.6)	51.5(8.4)	**51.5(1.4)**	60.6(10.9)	60.6(25.0)	63.3(12.2)	7.8(7.42)
	RD	Token	**35.5(2.1)**	51.1(8.4)	50.9(2.4)	62.4(11.3)	61.2(22.0)	59.7(18.4)	7.1(16.6)
	RS	Token	35.1(1.1)	51.5(7.0)	50.9(5.0)	**62.6(6.7)**	**63.2(22.5)**	61.2(10.8)	5.2(17.0)
	WR	Token	34.5(2.6)	**52.0(3.8)**	50.0(0.9)	60.6(10.2)	61.0(25.3)	61.8(12.5)	7.0(10.6)
	RT	Sentence	35.3(0.5)	51.1(9.6)	50.8(4.4)	60.5(17.8)	61.8(23.7)	62.0(1.99)	8.37(8.35)
	ADV	Hidden	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	Hidden	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 so metimes hurt performances.
- Token-level augmentations work well in general for supervised learni ng, especially with limited labeled data

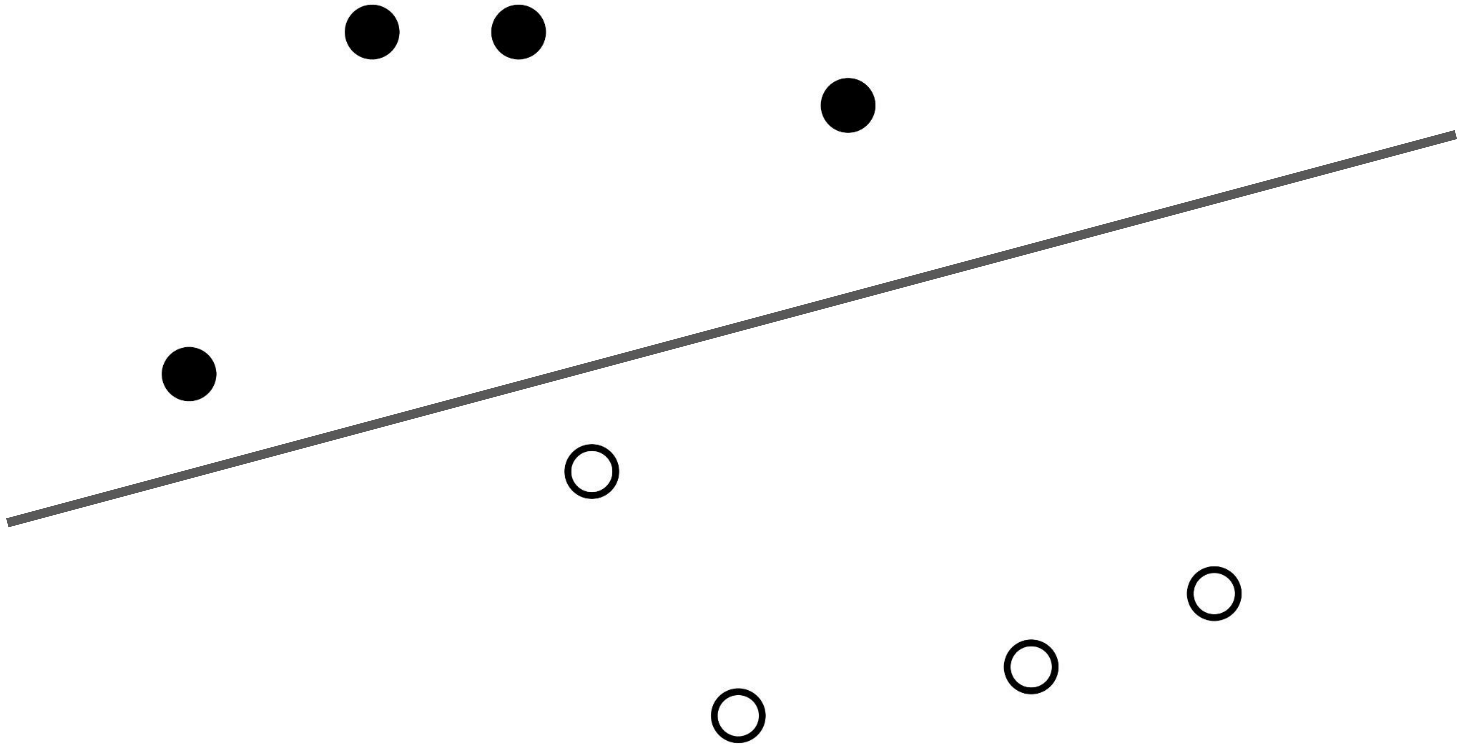
Semi-Supervised Learning

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

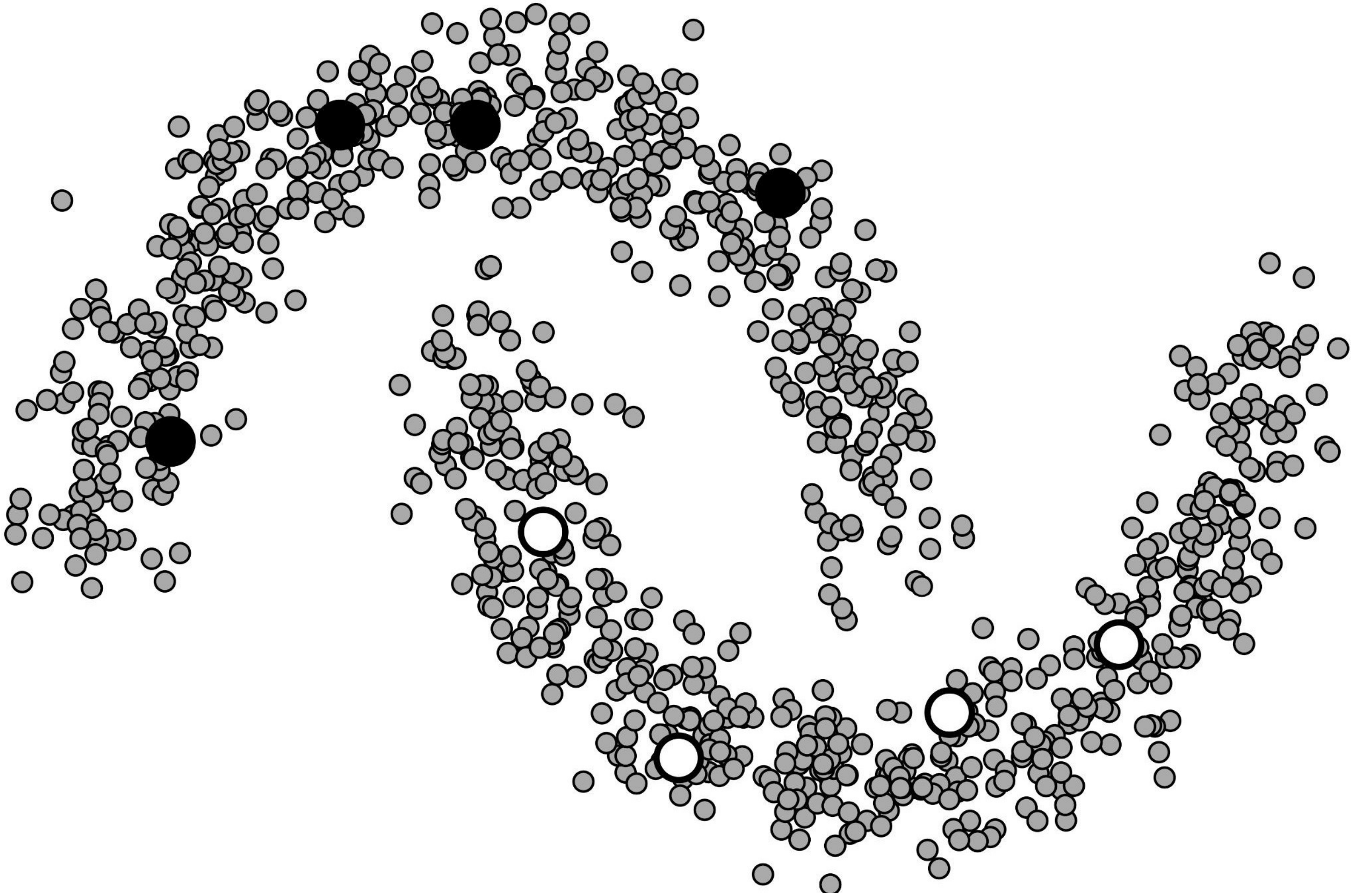
Semi-Supervised Learning



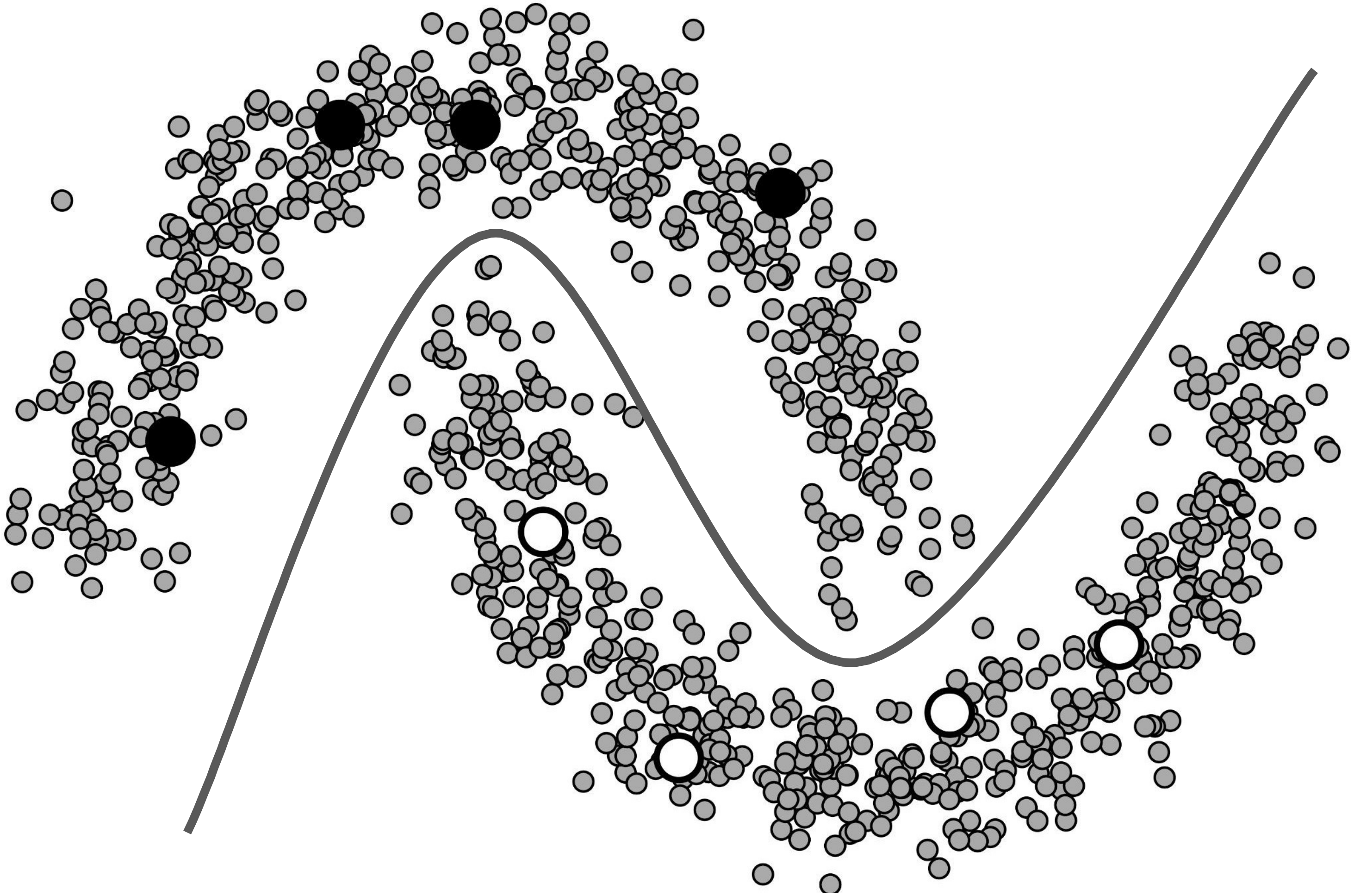
Semi-Supervised Learning



Semi-Supervised Learning



Semi-Supervised Learning



Supervised Learning

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

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

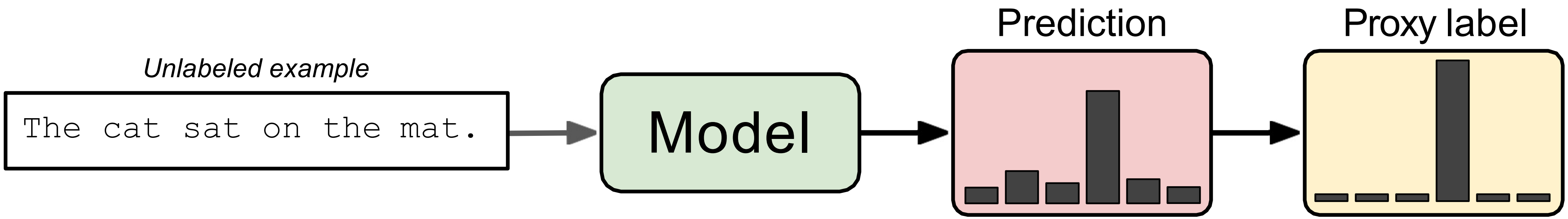
Use a proxy-label/pseudo-label/label guess

$$x \sim p(x)$$

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

P-Labeller

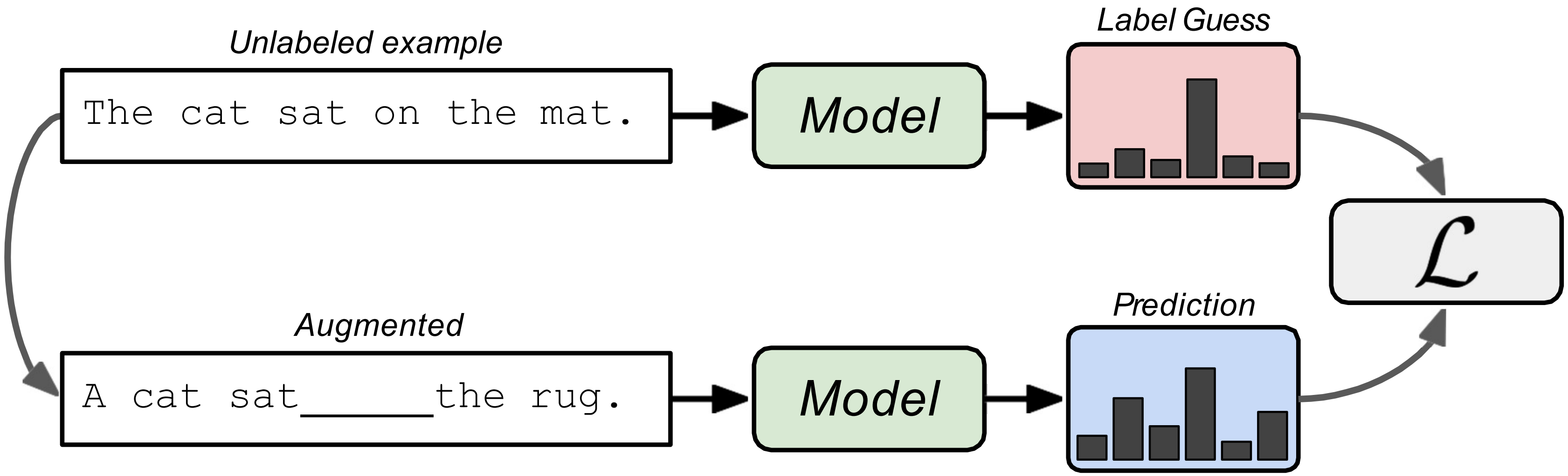
Self-training



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

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

Consistency regularization



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

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

LM as pseudo-labeller

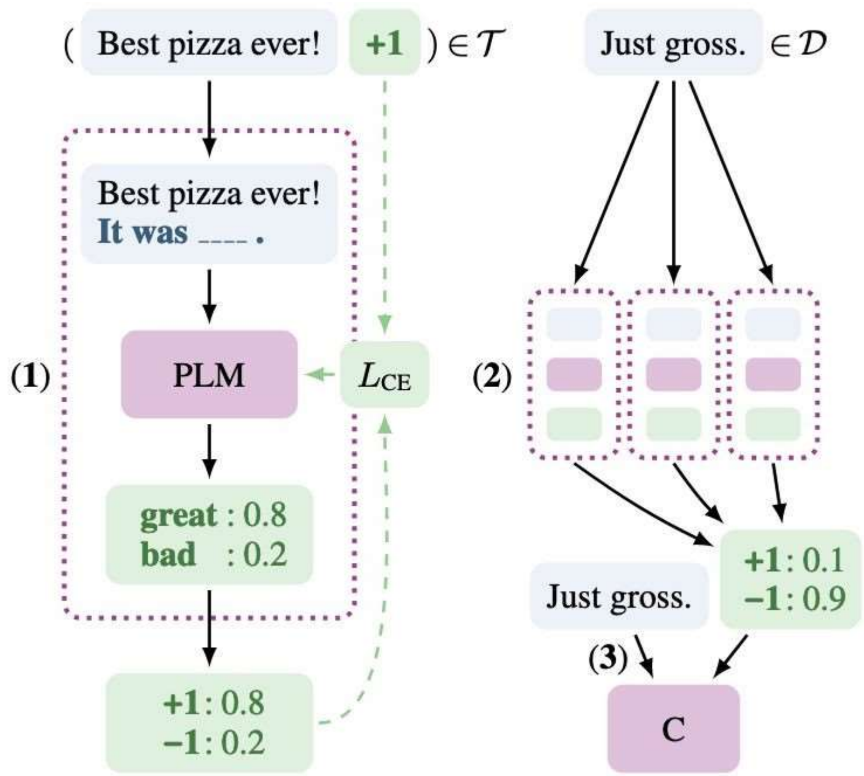


Figure 1: PET for sentiment classification. **(1)** A number of patterns encoding some form of task description are created to convert training examples to cloze questions; for each pattern, a pretrained language model is finetuned. **(2)** The ensemble of trained models annotates unlabeled data. **(3)**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

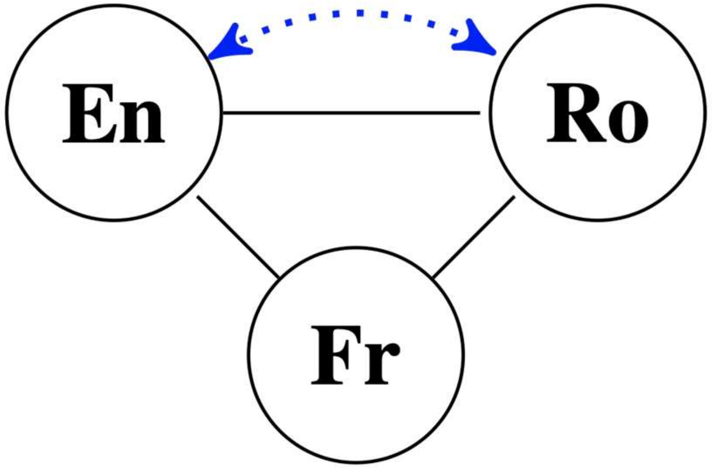
Pattern-exploiting training

Ex.	Method	Yelp	AG's	Yahoo	MNLI
$\lvert\mathcal{T}\rvert = 10$	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**
$\lvert\mathcal{T}\rvert = 50$	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**

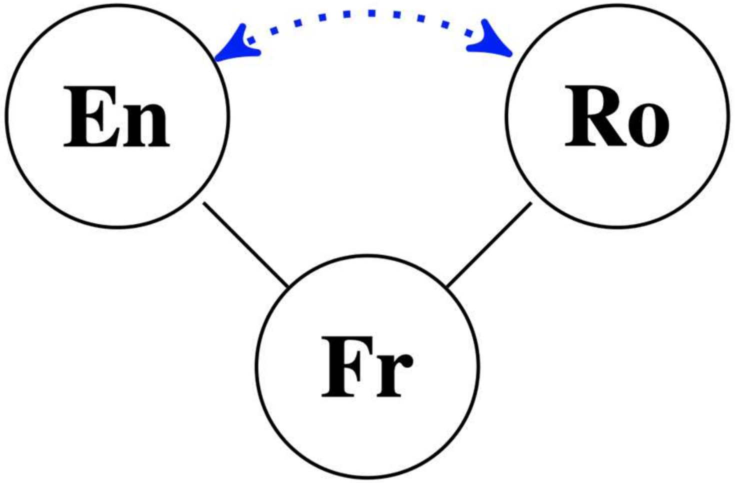
Table 2: Comparison of PET with two state-of-the-art semi-supervised methods using RoBERTa (base)

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

Preview: DataAug for Multilinguality



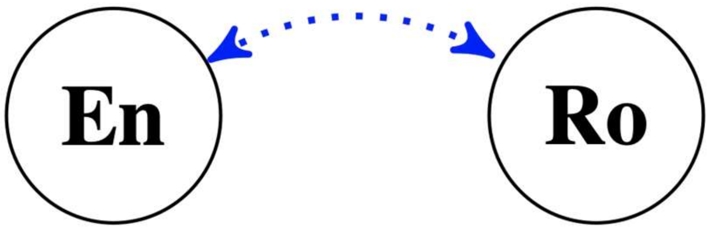
Supervised (Multilingual) Translation [Johnson et al. 2016, Firat et al. 2016]



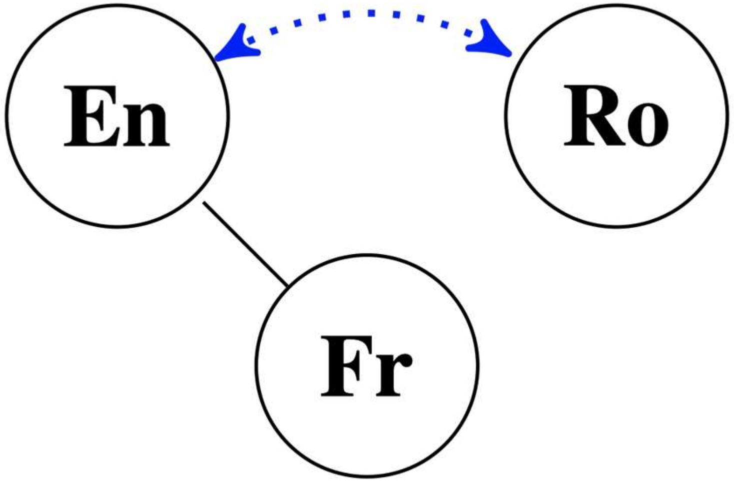
Zero shot translation [Johnson et al. 2016, Chen et al. 2017, Cheng et al. 2017, Al-Shedivat and Parikh 2019]

Solid lines indicate presence of parallel data

DataAug for Multilinguality



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