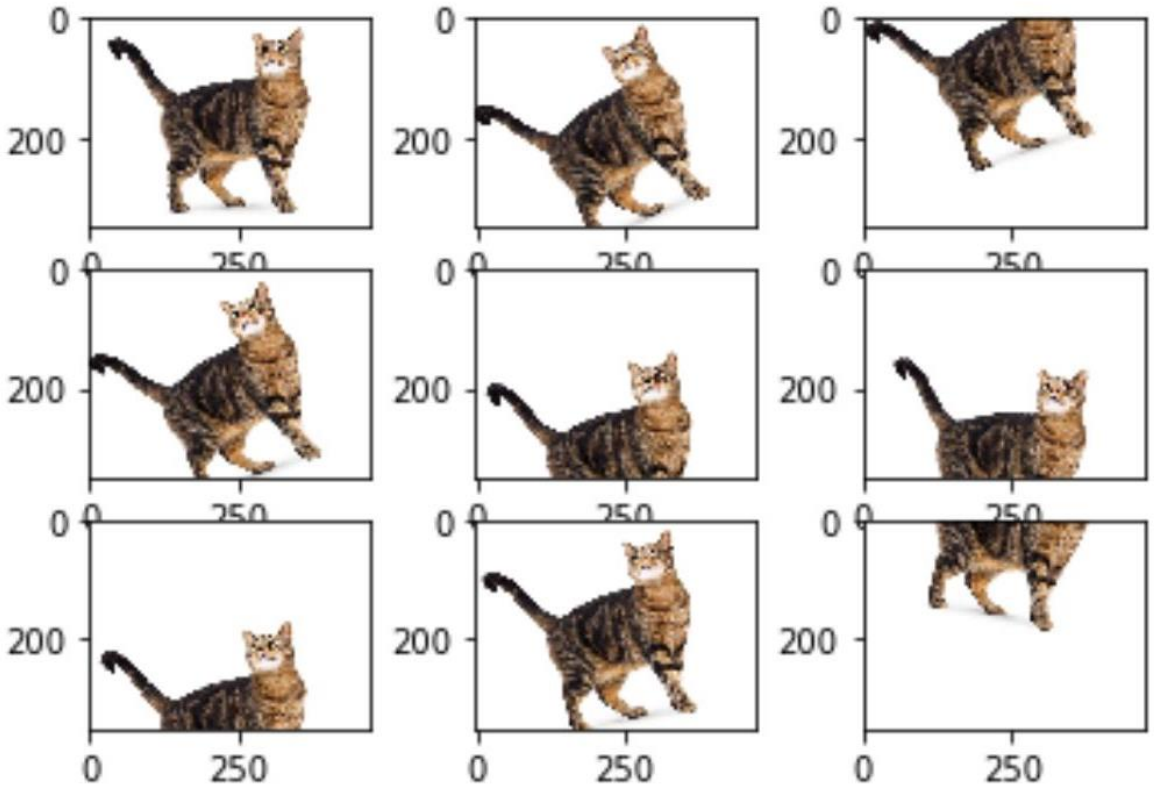
Data Augmentation for NLP

Dominik Lewy

Information Sensitivity: General\External

What data augmentation is?

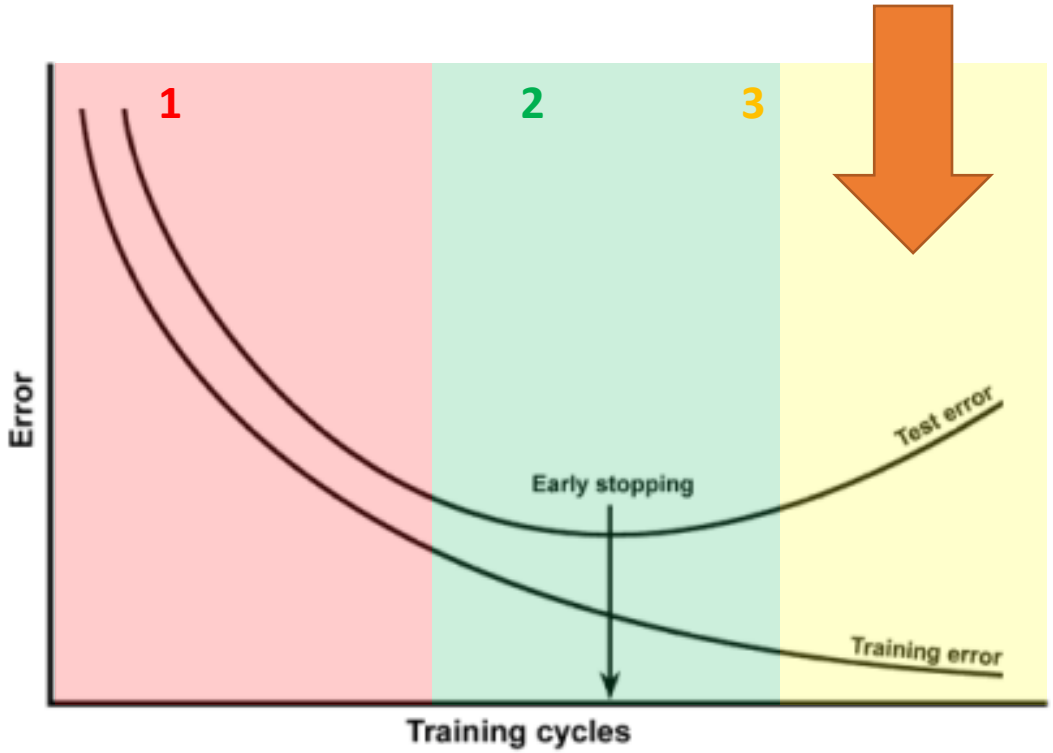
Data augmentation (DA) refers to strategies for increasing the diversity of training examples without explicitly collecting new data.



Information Sensitivity: General\External

Why do we use it?

Train/Test error curves



Train Error	Test Error	Strategy
High	High	Increase the capacity/complexity of the model
Low	Optimal	OK
Low	Much worse than train	Data augmentation, Regularization (weight decay, dropout)

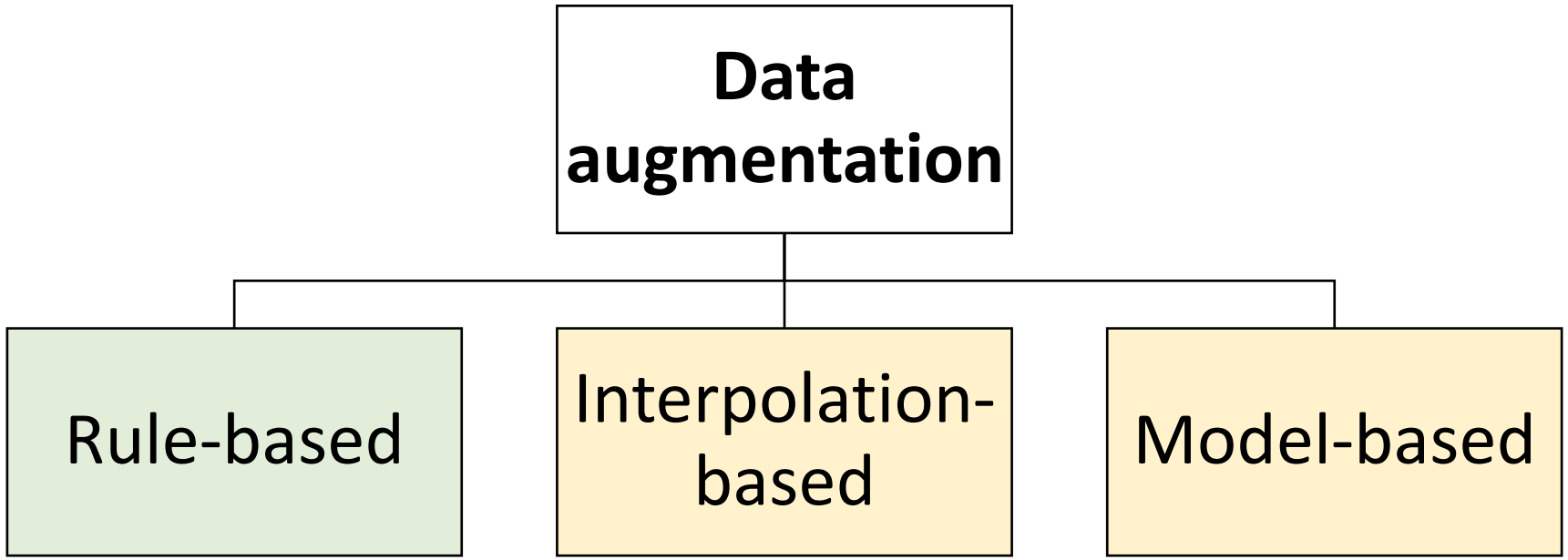
When discussing generalization and overfitting three scenarios arise:

1. Model trained excludes the true data-generating process – this corresponds to underfitting and induces bias
2. Model matched the true data-generating process
3. Model included the data-generating process but also many other possible generating processes – this corresponds to overfitting, in such scenario variance rather than bias dominates the error

The goal of regularization is to take the model from third scenario to the second.

Information Sensitivity: General\External

How can we categorize data augmentation methods in NLP space?



Copy is modified

Synthetic sample is created

Information Sensitivity: General\External

Example of a rule-based method

EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks

Jason Wei[1,2] **Kai Zou**[3]
[1]Protago Labs Research, Tysons Corner, Virginia, USA
[2]Department of Computer Science, Dartmouth College
[3]Department of Mathematics and Statistics, Georgetown University
jason.20@dartmouth.edu kz56@georgetown.edu

https://arxiv.org/pdf/1901.11196.pdf

Information Sensitivity: General\External

EDA – Easy Data Augmentation

Operation	Sentence
None	A sad, superior human comedy played out on the back roads of life.
SR	A ***lamentable***, superior human comedy played out on the ***backward*** road of life.
RI	A sad, superior human comedy played out on ***funniness*** the back roads of life.
RS	A sad, superior human comedy played out on ***roads*** back ***the*** of life.
RD	A sad, superior human out on the roads of life.

Table 1: Sentences generated using EDA. SR: synonym replacement. RI: random insertion. RS: random swap. RD: random deletion.

Information Sensitivity: General\External

https://arxiv.org/pdf/1901.11196.pdf

EDA – Easy Data Augmentation

Data sets include:

- SST-2 – Stanford Sentiment Treebank
- CR – customer reviews
- SUBJ – subjective/objective dataset
- TREC – question type dataset
- PC – Pro-Con dataset

	Training Set Size			
Model	500	2,000	5,000	full set
RNN	75.3	83.7	86.1	87.4
+EDA	79.1	84.4	87.3	88.3
CNN	78.6	85.6	87.7	88.3
+EDA	80.7	86.4	88.3	88.8
Average	76.9	84.6	86.9	87.8
+EDA	79.9	85.4	87.8	**88.6**

Table 2: Average performances (%) across five text classification tasks for models with and without EDA on different training set sizes.

Architectures include:

- RNN – LSTM-RNN
- CNN

https://arxiv.org/pdf/1901.11196.pdf

Information Sensitivity: General\External

EDA – Easy Data Augmentation

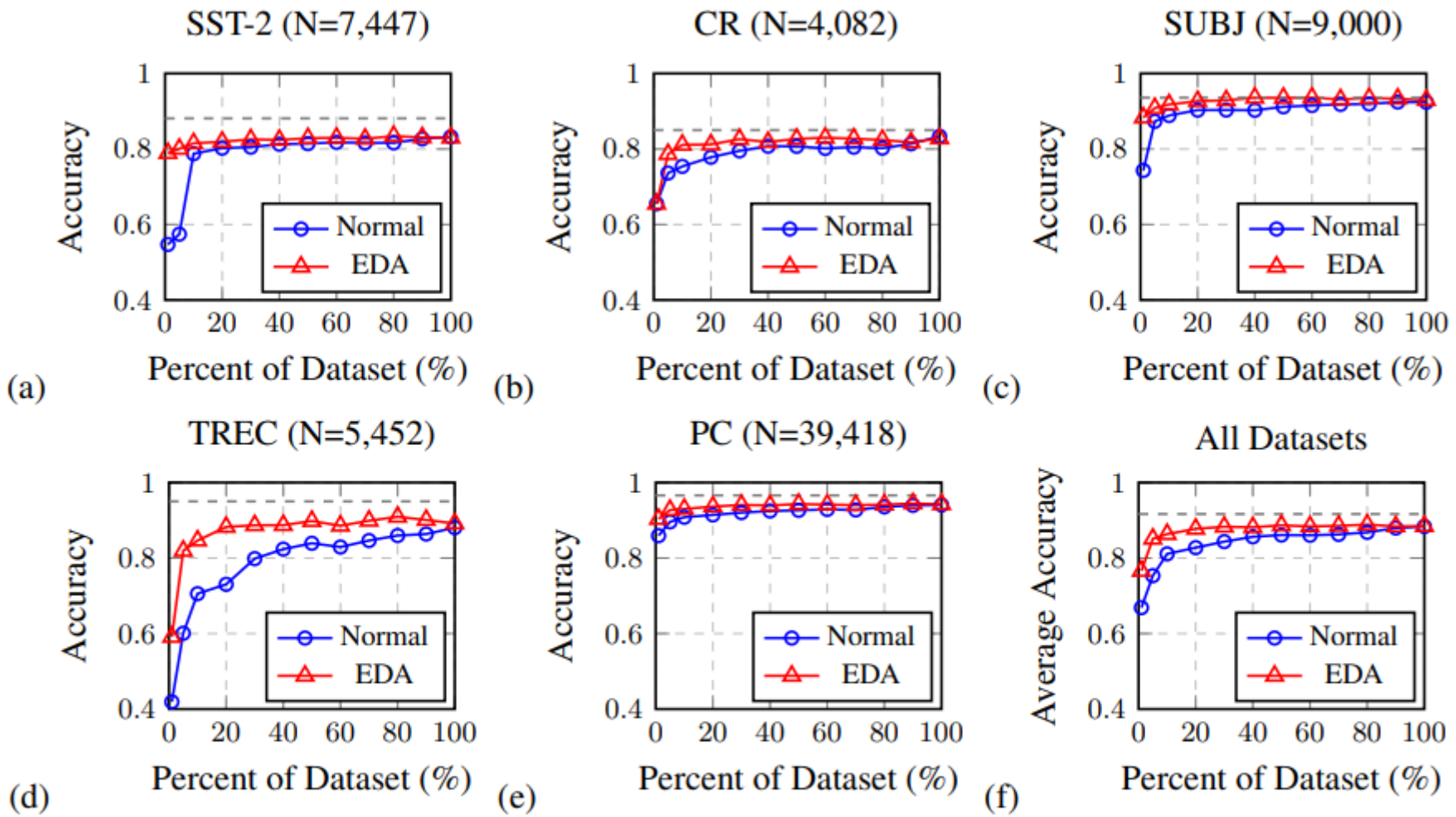


Figure 1: Performance on benchmark text classification tasks with and without EDA, for various dataset sizes used for training. For reference, the dotted grey line indicates best performances from Kim (2014) for SST-2, CR, SUBJ, and TREC, and Ganapathibhotla (2008) for PC.

https://arxiv.org/pdf/1901.11196.pdf

Information Sensitivity: General\External

EDA – Easy Data Augmentation

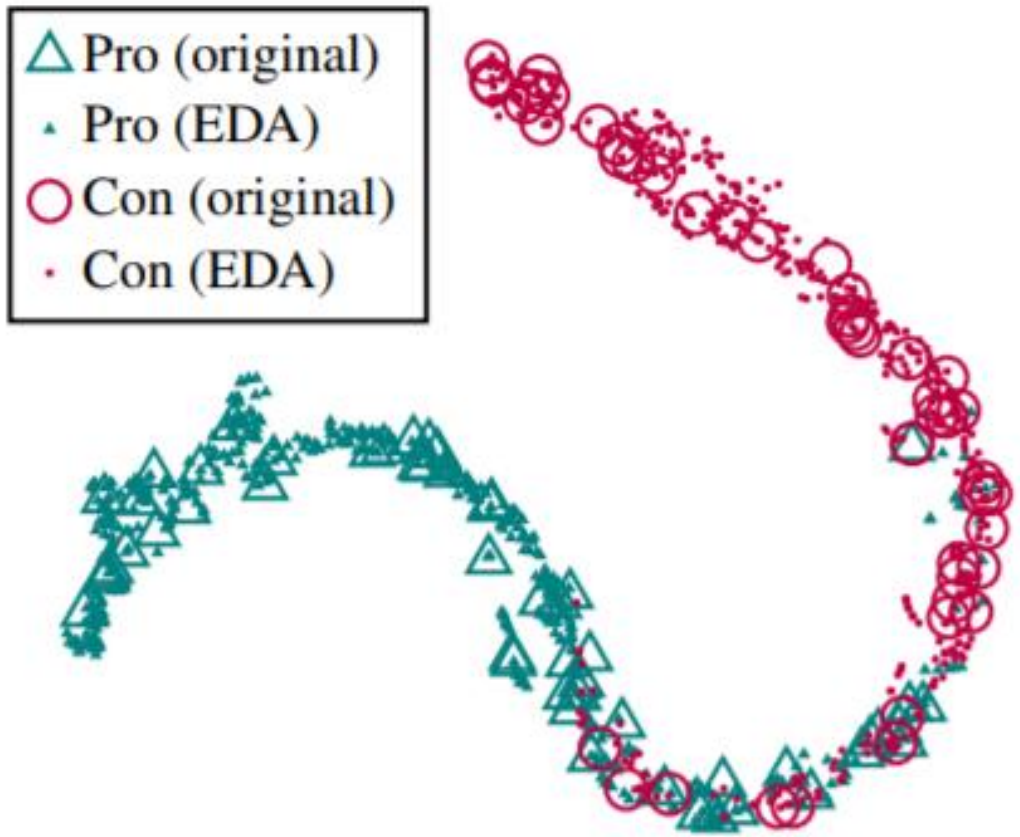


Figure 2: Latent space visualization of original and augmented sentences in the Pro-Con dataset. Augmented sentences (small triangles and circles) closely surround original sentences (big triangles and circles) of the same color, suggesting that augmented sentences maintianed their true class labels.

https://arxiv.org/pdf/1901.11196.pdf

Information Sensitivity: General\External

EDA – Easy Data Augmentation

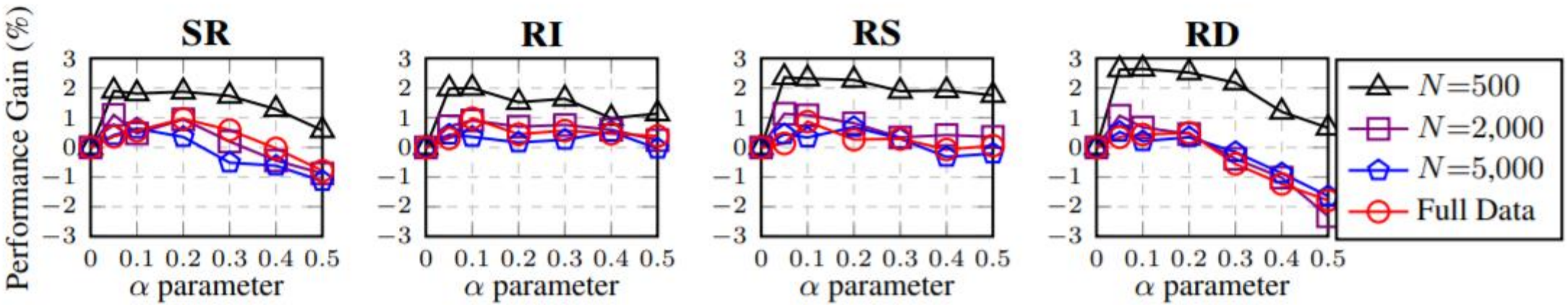


Figure 3: Average performance gain of EDA operations over five text classification tasks for different training set sizes. The α parameter roughly means "percent of words in sentence changed by each augmentation." SR: synonym replacement. RI: random insertion. RS: random swap. RD: random deletion.

Information Sensitivity: General\External

https://arxiv.org/pdf/1901.11196.pdf

Example of an interpolation-based method

Leveraging BERT with Mixup for Sentence Classification (Student Abstract)

Amit Jindal,[1] Dwaraknath Gnaneshwar,[1] Ramit Sawhney,[2] Rajiv Ratn Shah[3]
[1]Manipal Institute of Technology
{amitj646, dwarakasharma} @gmail.com
[2]Netaji Subhas Institute of Technology
ramits.co@nsit.net.in
[3]Indraprastha Institute of Information Technology, Delhi
rajivratn@iiitd.ac.in

Information Sensitivity: General\External

https://ojs.aaai.org/index.php/AAAI/article/view/7186

Data Augmentation via Mixing Images

Image 1 -
label: dog

Image 2 -
label: cow

Mixup -
label: (dog:0.3, cow:0.7)

SamplePairing -
label: dog

CutMix -
label: (dog:0.7, cow:0.3)

SmoothMix -
label: (dog:0.3, cow:0.7)

AttentiveMix -
label: (dog:0.14, cow:0.86)

RandomSquare -
label: (dog:0.42, cow:0.58)

Information Sensitivity: General\External

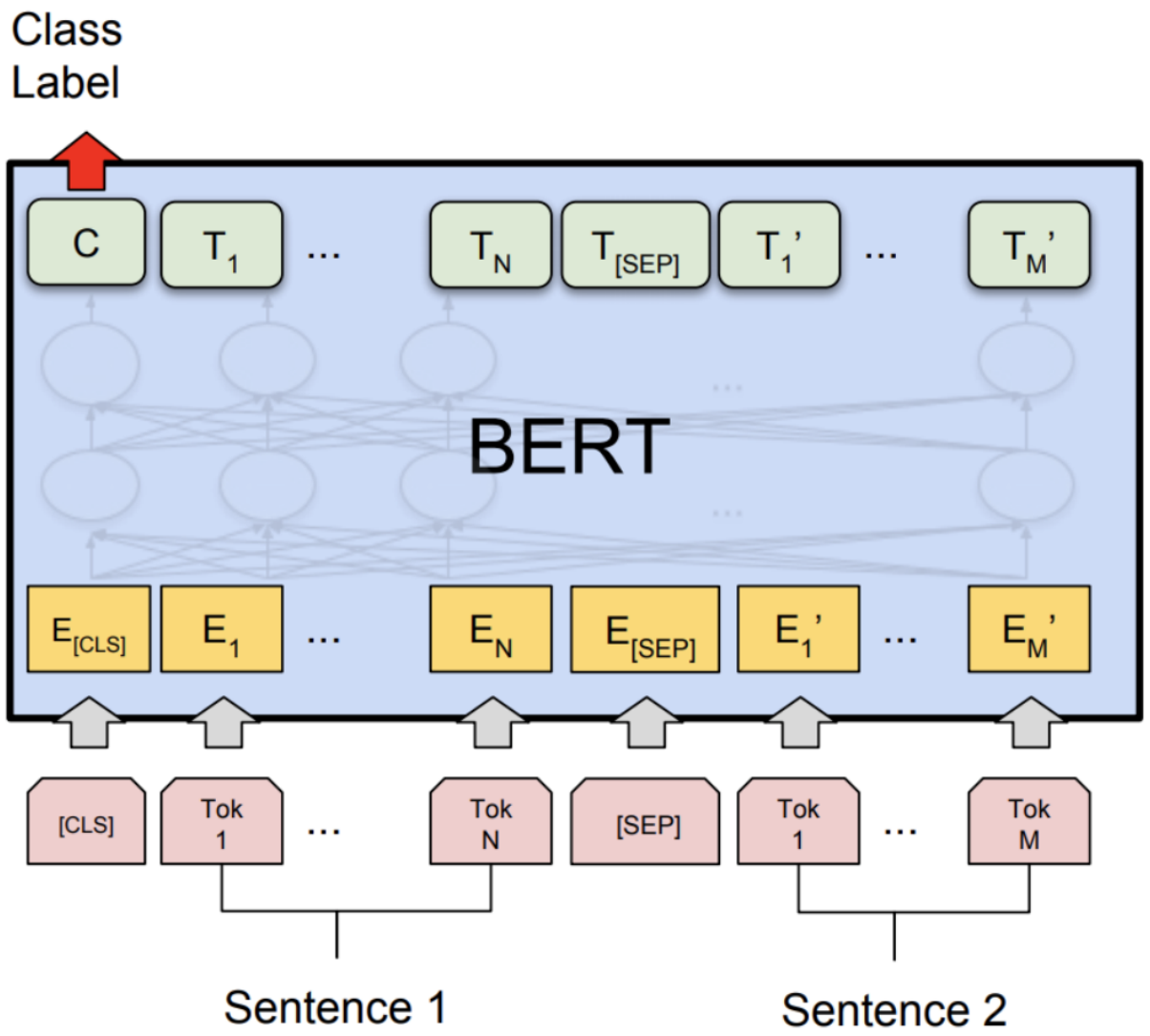
Mixup for Sentence Classification

Sentence A: *The quick brown fox jumps over the lazy dog.*

Sentence B: *Pack my box with five dozen liquor jugs.*

Information Sensitivity: General\External

Mixup for Sentence Classification



Information Sensitivity: General\External

Mixup for Sentence Classification

Experiments error rates					
Model	**IMDB**	**SST-1**	**MR**	**TREC**	**SUBJ**
CNN	7.67	55.0	21.61	8.8	7.59
LSTM	12.85	53.6	21.67	13.5	8.60
BERT	6.46	46.75	**10.83**	2.62	2.20
BERT + Input Mixup	6.17	45.55	12.81	2.41	1.505
BERT + Manifold Mixup	**6.02**	**44.20**	11.94	**2.20**	**1.501**

Table 1: Test error (%) of the testing methods using BERT. Best results highlighted in Bold.

Data sets include:

- IMDB – binary sentiment classification dataset
- MR – movie review data set with binary sentiment
- SST-1 – Stanford Sentiment Treebank
- TREC – question type dataset
- SUBJ – subjective/objective dataset

Information Sensitivity: General\External

https://ojs.aaai.org/index.php/AAAI/article/view/7186

Mixup for Sentence Classification

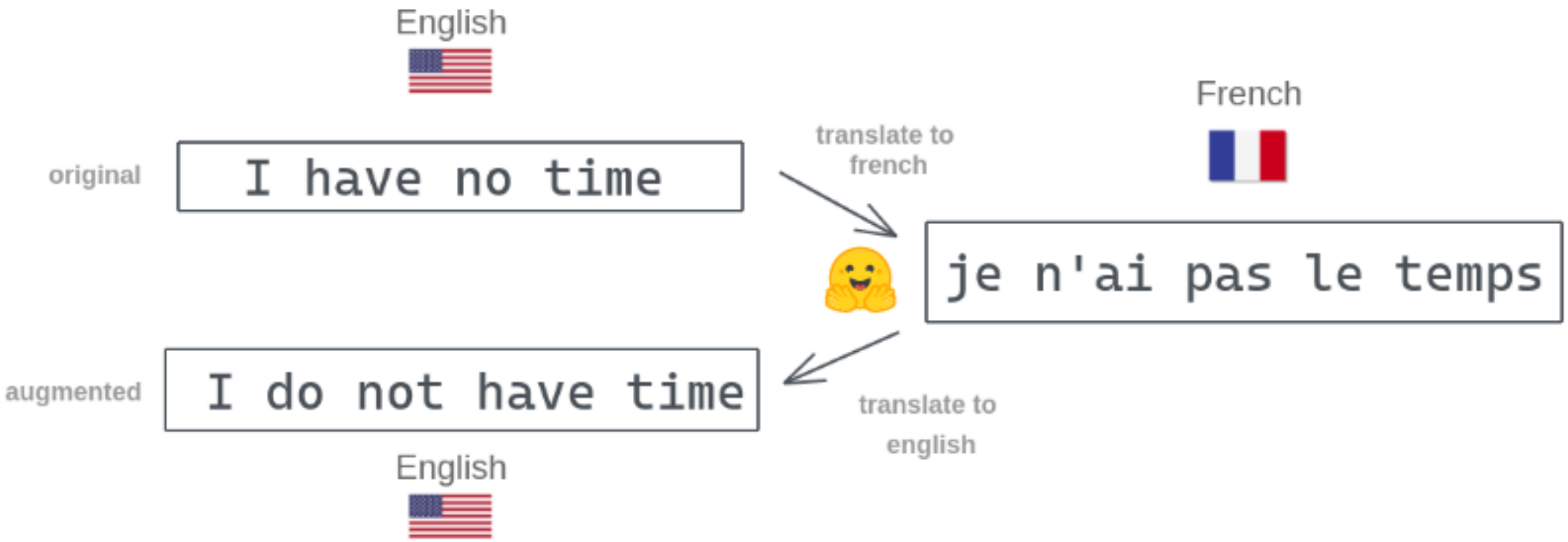
Table 2: Test error (%) Manifold Mixup for different sets of eligible layers S on IMDB

S	IMDB
{0}	6.17
{0, 1}	6.10
{0, 1, 2}	6.27
{0, 1, 2, 3}	6.21
{0, 1, 2, 3, 4}	6.15
{0, 1, 2, 3, 4, 5}	6.22
{0, 1, 2, 3, 4, 5, 6}	6.09
{0, 1, 2, 3, 4, 5, 6, 7 }	**6.02**
{0, 1, 2, 3, 4, 5, 6, 7, 8}	6.23
{0, 1, 2, 3, 4, 5, 6, 7, 8, 9}	6.25
{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10}	6.28
{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}	6.40

Information Sensitivity: General\External

https://ojs.aaai.org/index.php/AAAI/article/view/7186

Example of a model-based method



Information Sensitivity: General\External

Where is augmentation mostly used now?

- Low-Resource Languages
- Mitigating Bias
- Fixing Class Imbalance
- Few-Shot Learning

What is the array of tasks in NLP to be considered?

- Sentence Classification
- Summarization
- Question Answering
- Sequence Tagging
- Parsing
- Grammatical Error Correction
- Neural Machine Translation
- Data to text (NLG)
- ….

Information Sensitivity: General\External

Sequence Tagging – rule-based equivalent

An Analysis of Simple Data Augmentation for Named Entity Recognition

Xiang Dai[1,2,3] **Heike Adel**[1]
[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]University of Sydney, Sydney, Australia
[3]CSIRO Data61, Sydney, Australia
dai.dai@csiro.au heike.adel@de.bosch.com

Information Sensitivity: General\External

https://aclanthology.org/2020.coling-main.343.pdf

Sequence Tagging – rule-based equivalent

	Instance											
None	She	did	not	complain	of	headache	or	any	other	neurological	symptoms	.
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O
LwTR	L.	One	not	complain	of	headache	he	any	interatrial	neurological	current	.
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O
SR	She	did	non	complain	of	headache	or	whatsoever	former	neurologic	symptom	.
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O
MR	She	did	not	complain	of	neuropathic	pain	syndrome	or	acute	pulmonary	disease	.
	O	O	O	O	O	B-problem	I-problem	I-problem	O	B-problem	I-problem	I-problem	O
SiS	not	complain	She	did	of	headache	or	neurological	any	symptoms	other	.	
	O	O	O	O	O	B-problem	O	B-problem	I-problem	I-problem	I-problem	O	

Table 1: Original training instance and different types of augmented instances. We highlight changes using blue color. Note that LwTR (Label-wise token replacement) and SiS (Shuffle within segments) change token sequence only, whereas SR (Synonym replacement) and MR (Mention replacement) may also change the label sequence.

Information Sensitivity: General\External

https://aclanthology.org/2020.coling-main.343.pdf

Sequence Tagging – interpolation-based equivalent

SeqMix: Augmenting Active Sequence Labeling via Sequence Mixup

Rongzhi Zhang
Georgia Tech
rongzhi.zhang@gatech.edu

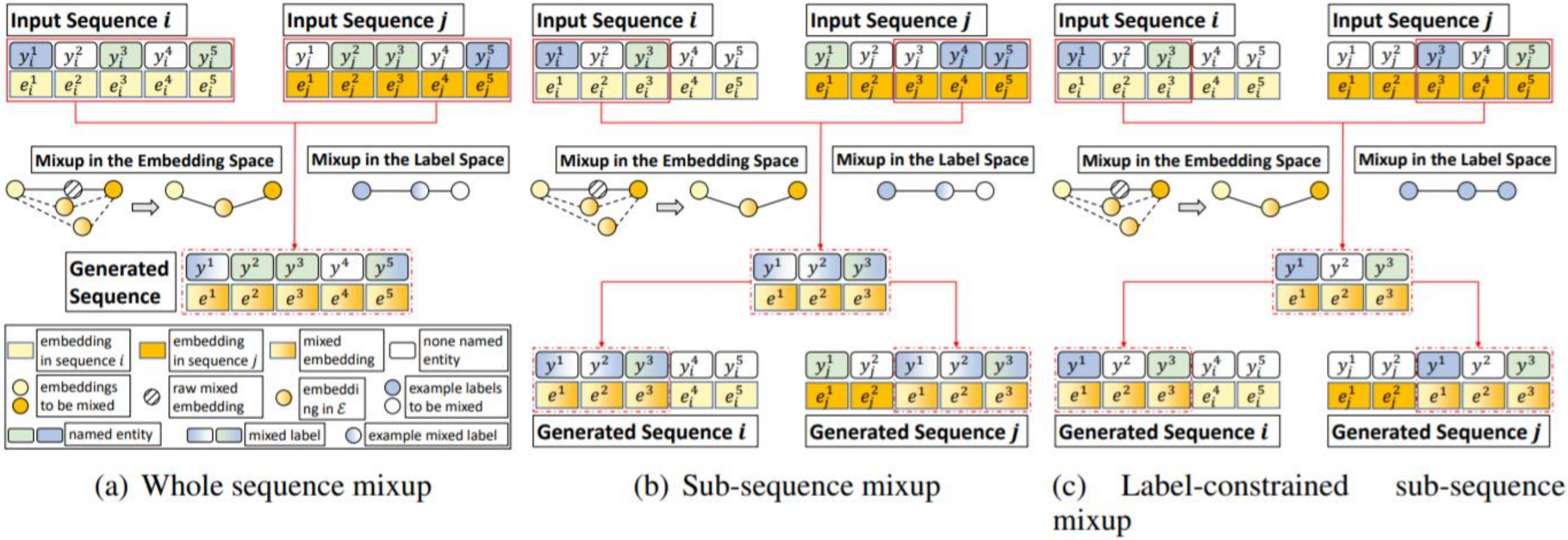
Yue Yu
Georgia Tech
yueyu@gatech.edu

Chao Zhang
Georgia Tech
chaozhang@gatech.edu

Information Sensitivity: General\External

http://chaozhang.org/papers/2020-emnlp-seqmix.pdf

Sequence Tagging – interpolation-based equivalent



(a) Whole sequence mixup (b) Sub-sequence mixup (c) Label-constrained sub-sequence mixup

http://chaozhang.org/papers/2020-emnlp-seqmix.pdf

Information Sensitivity: General\External

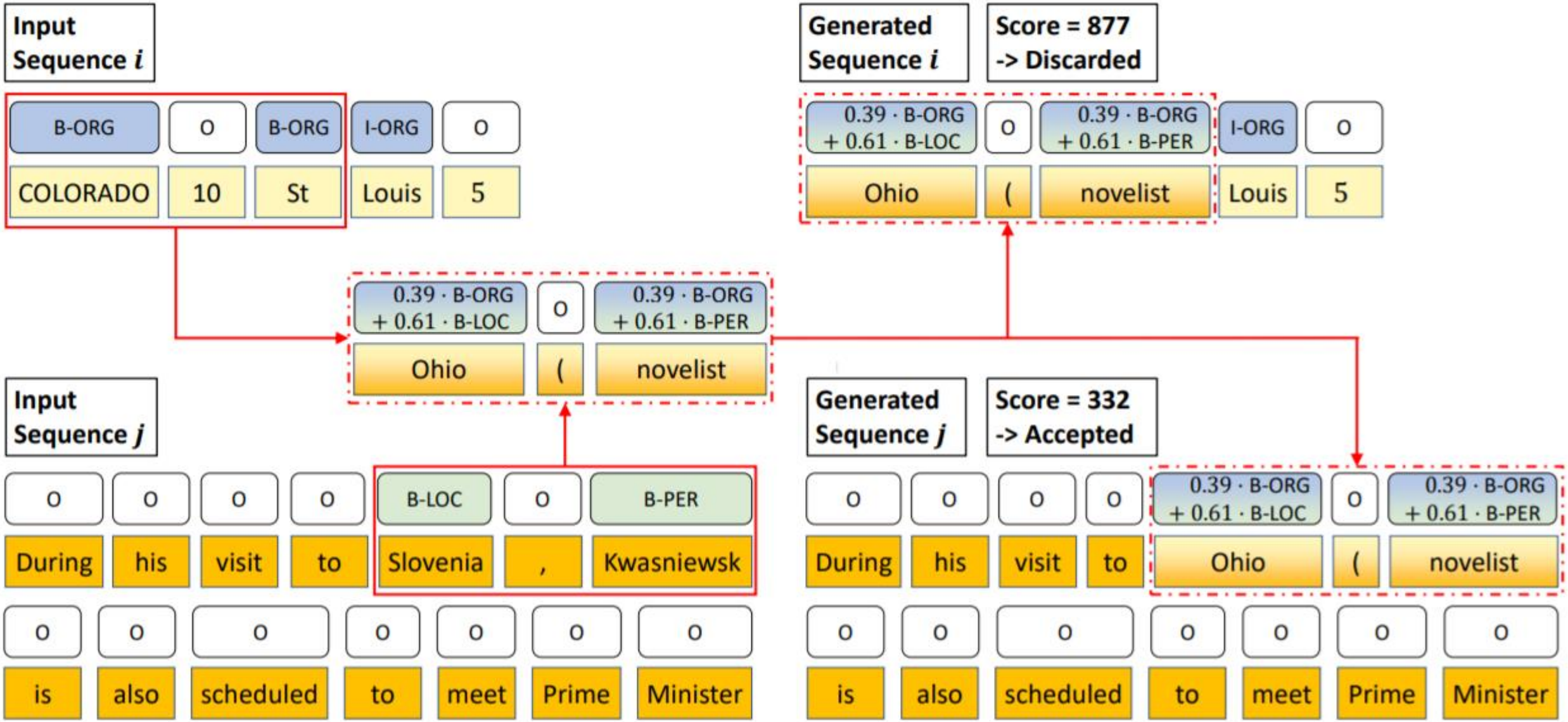


Figure 5: A generation case of sub-sequence mixup.

Other methods

DA Method	Ext.Know	Pretrained	Preprocess	Level	Task-Agnostic
SYNONYM REPLACEMENT (Zhang et al., 2015)	✓	×	tok	Input	✓
RANDOM DELETION (Wei and Zou, 2019)	×	×	tok	Input	✓
RANDOM SWAP (Wei and Zou, 2019)	×	×	tok	Input	✓
BACKTRANSLATION (Sennrich et al., 2016)	×	✓	Depends	Input	✓
SCPN (Wieting and Gimpel, 2017)	×	✓	const	Input	✓
SEMANTIC TEXT EXCHANGE (Feng et al., 2019)	×	✓	const	Input	✓
CONTEXTUALAUG (Kobayashi, 2018)	×	✓	-	Input	✓
LAMBADA (Anaby-Tavor et al., 2020)	×	✓	-	Input	×
GECA (Andreas, 2020)	×	×	tok	Input	×
SEQMIXUP (Guo et al., 2020)	×	×	tok	Input	×
SWITCHOUT (Wang et al., 2018b)	×	×	tok	Input	×
EMIX (Jindal et al., 2020a)	×	×	-	Emb/Hidden	✓
SPEECHMIX (Jindal et al., 2020b)	×	×	-	Emb/Hidden	Speech/Audio
MIXTEXT (Chen et al., 2020c)	×	×	-	Emb/Hidden	✓
SIGNEDGRAPH (Chen et al., 2020b)	×	×	-	Input	×
DTREEMORPH (Şahin and Steedman, 2018)	×	×	dep	Input	✓
Sub^2 (Shi et al., 2021)	×	×	dep	Input	Substructural
DAGA (Ding et al., 2020)	×	×	tok	Input+Label	×
WN-HYPERS (Feng et al., 2020)	✓	×	const+KWE	Input	✓
SYNTHETIC NOISE (Feng et al., 2020)	×	×	tok	Input	✓
UEDIN-MS (DA part) (Grundkiewicz et al., 2019)	✓	×	tok	Input	✓
NONCE (Gulordava et al., 2018)	✓	×	const	Input	✓
XLDA (Singh et al., 2019)	×	✓	Depends	Input	✓
SEQMIX (Zhang et al., 2020)	×	✓	tok	Input+Label	×
SLOT-SUB-LM (Louvan and Magnini, 2020)	×	✓	tok	Input	✓
UBT & TBT (Vaibhav et al., 2019)	×	✓	Depends	Input	✓
SOFT CONTEXTUAL DA (Gao et al., 2019)	×	✓	tok	Emb/Hidden	✓
DATA DIVERSIFICATION (Nguyen et al., 2020)	×	✓	Depends	Input	✓
DIPS (Kumar et al., 2019a)	×	✓	tok	Input	✓
AUGMENTED SBERT (Thakur et al., 2021)	×	✓	-	Input+Label	Sentence Pairs

Information Sensitivity: General\External

https://arxiv.org/pdf/2105.03075.pdf

The end.
Thank you!

Information Sensitivity: General\External

Discussion

Information Sensitivity: General\External