香港中文大學
The Chinese University of Hong Kong

Effective Training with Data Engineering for Language Understanding and Generation

Wenxiang Jiao

Ph.D. Oral Defense

Supervisor: Prof. Irwin King & Prof. Michael R. Lyu

Committee: Prof. Kevin Yip

Prof. Xunying Liu

Prof. Shou-dou Lin

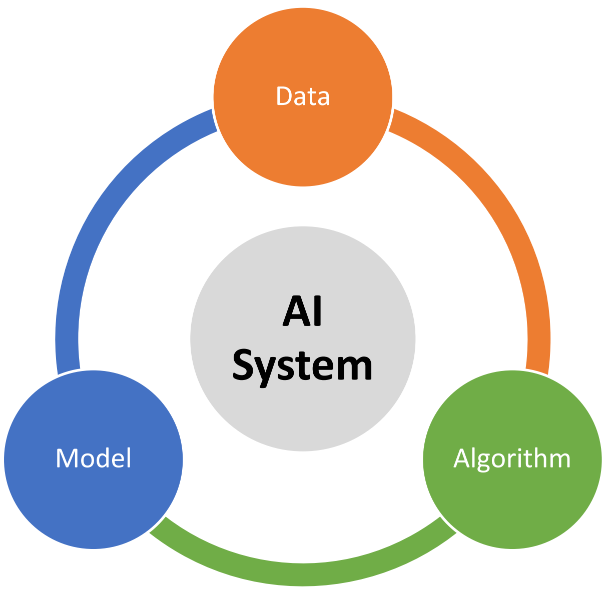
Outline

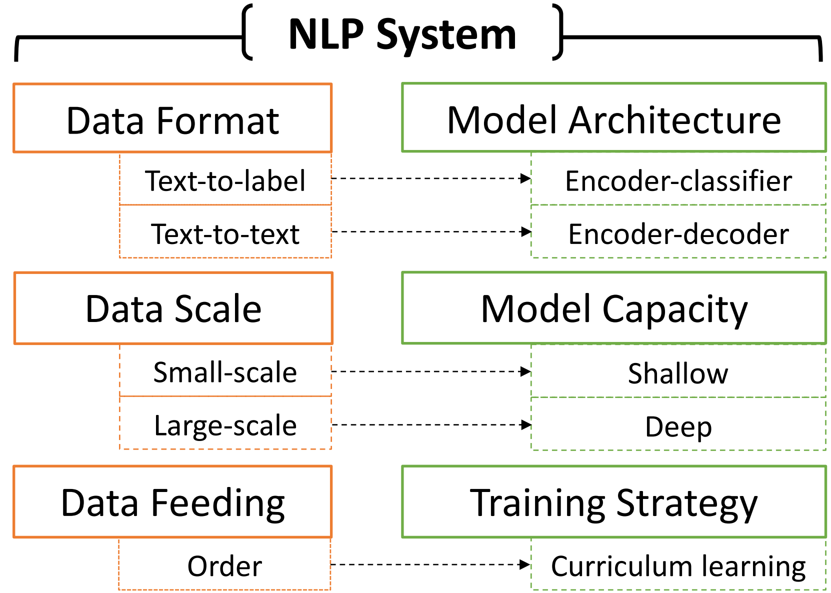
- ❑ Introduction
- ❑ Context Enhancement with Intra-Sample Structure Mining
- ❑ Self-Supervised Learning with Intra-Sample Structure Mining
- ❑ Data Rejuvenation with Inter-Sample Quality Mining
- ❑ Self-Training Sampling with Inter-Sample Quality Mining
- ❑ Conclusion

Outline

Importance of Data in Artificial Intelligence

- Data is the foundation of natural language processing (NLP) systems





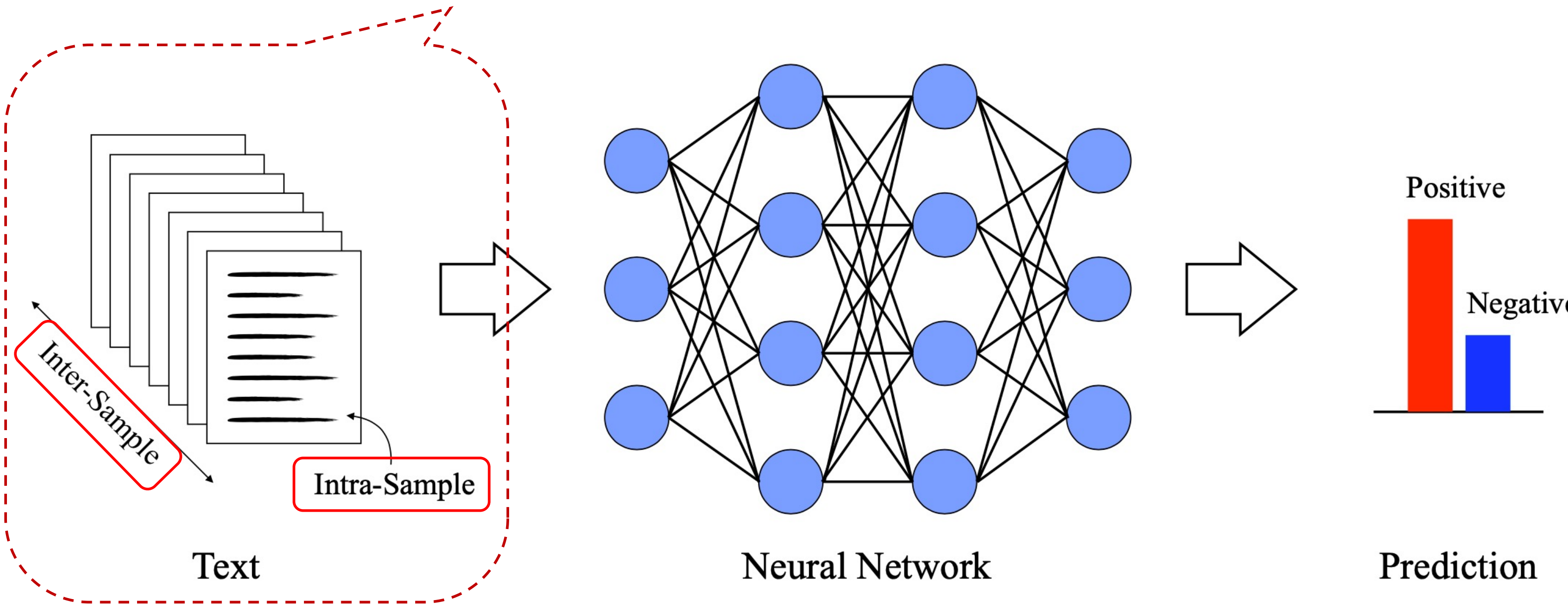
AI System = Code + Data

(model/algorithm)

Training NLP models more effectively on data is critical!

How to Exploit Training Data

- Two dimensions
  - Intra-sample structure: structure information shared by text samples
  - Inter-sample quality: differentiate text samples by their quality



A general diagram for a text classification task

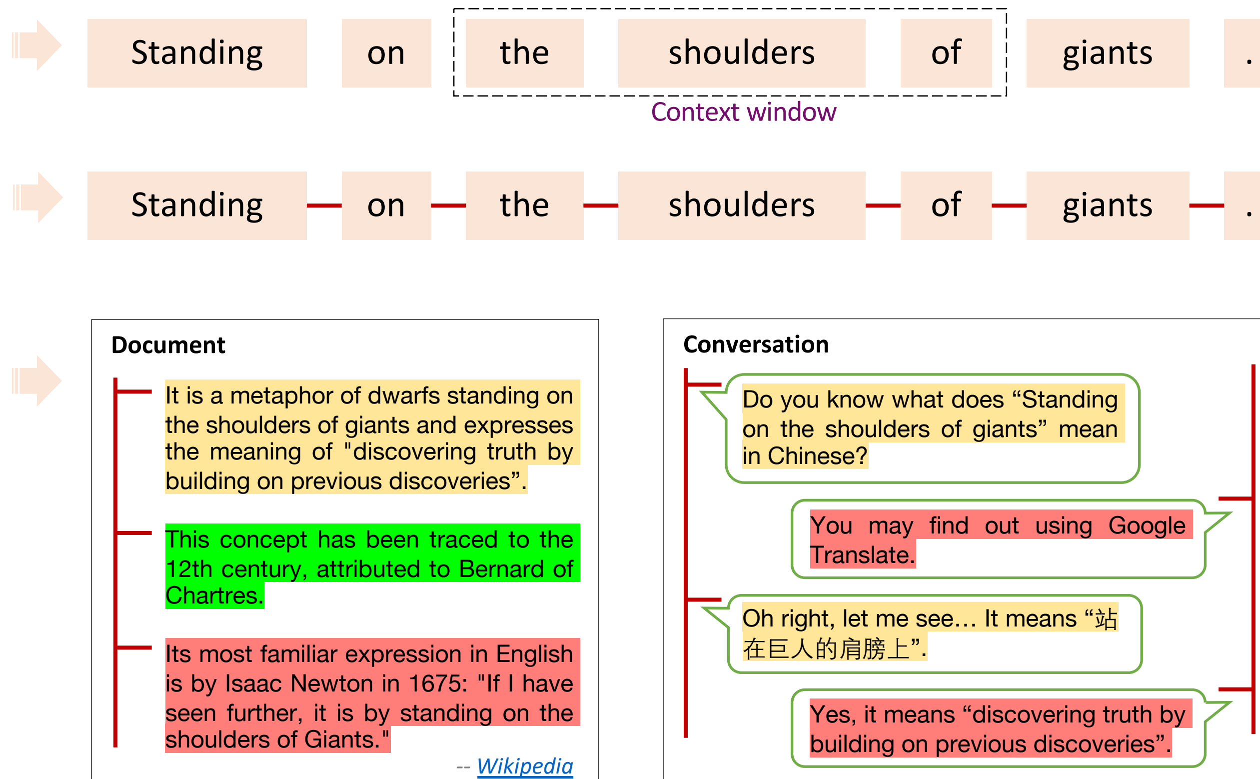
Advances in Intra-Sample Structure Exploitation

- Intra-sample structure provides signals for representation learning, i.e., the context information in texts

Local context
- Word2Vec, GloVe

Sentential context
- CNN, RNN, SAN
- ELMo, GPT, BERT

Hierarchical context
- HAN, cLSTM, HiGRU
- Pre-CODE, TL-ERC

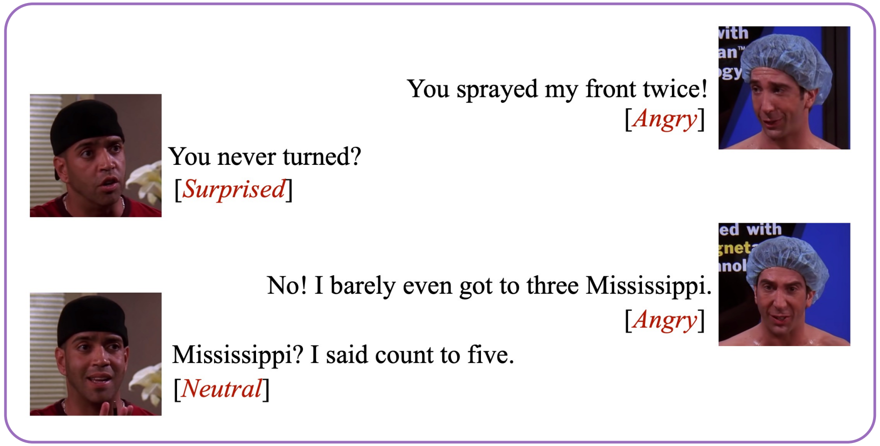


Challenges in Intra-Sample Structure Exploitation

Challenges

- Most studies have been conducted on sentential context with limited structure information
- Documents or conversations are more practical scenarios
- Early studies on hierarchical context do not fully utilize the structure information
- Few studies on self-supervised learning from hierarchical context

Emotion Recognition in Conversations (ERC)



Reasons

- Recognizing emotions in conversations is novel
- Rich structure information as learning signals
- Small scale datasets may gain more
- Unlabeled conversation data becomes available

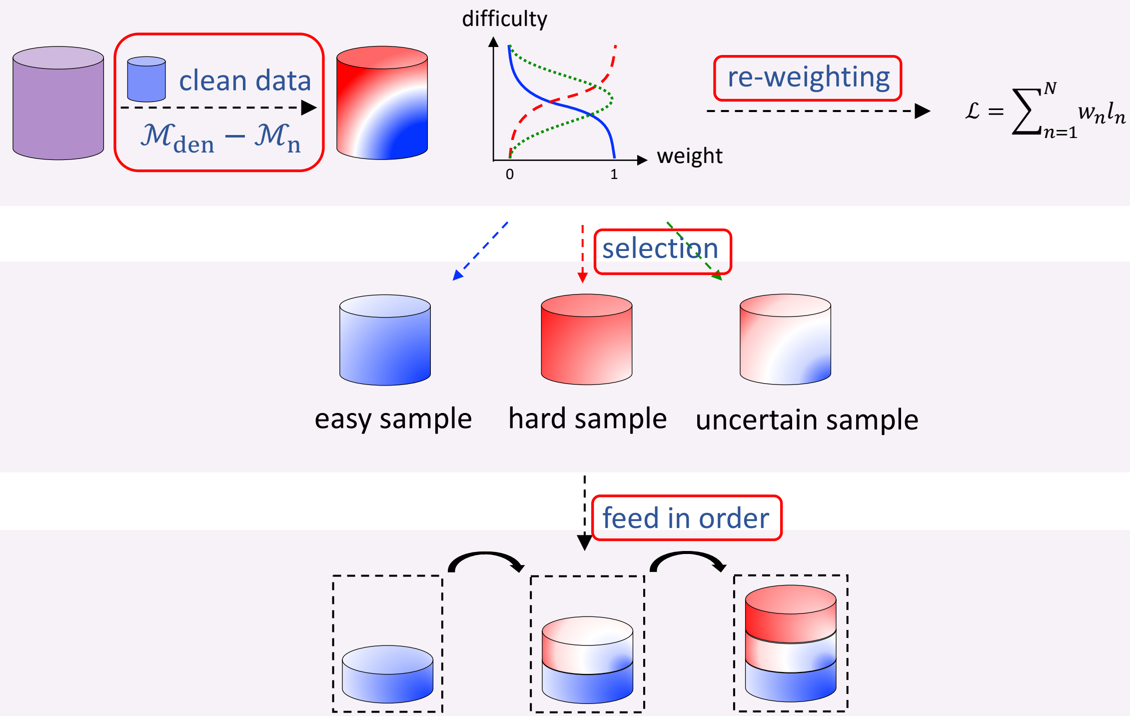
Advances in Inter-Sample Quality Exploitation

- Inter-sample quality should be considered when facing large-scale datasets, which may contain noises or mistakes

Dynamic weighting
- Self-paced
- Hard sample
- Active bias

Data selection
- Language model
- Difficult words
- Uncertainty

Curriculum learning
- Linguistic properties
- Embedding norm

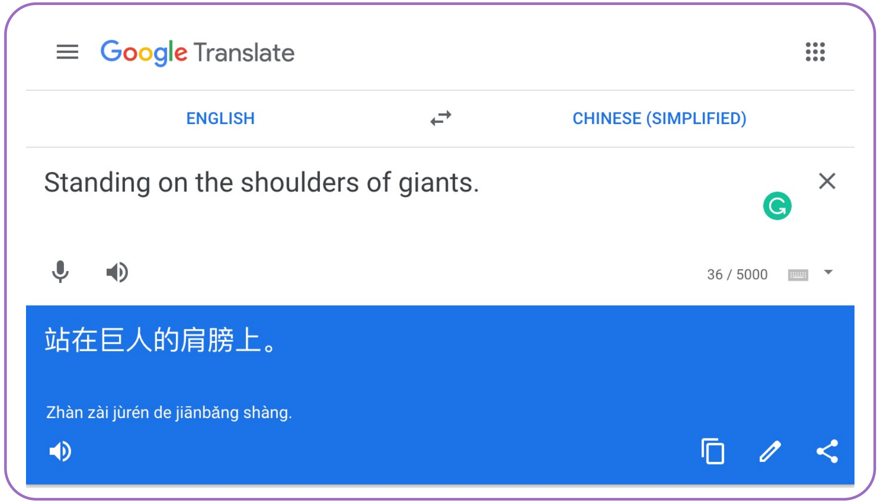


Challenges in Inter-Sample Quality Exploitation

Challenges

- Dynamic weighting and curriculum learning require the modification of training strategies
- Data selection is easy to implement but there is a lack of understanding on the unpreferred data
- Few studies on how to re-use the unpreferred data
- Data selection for efficient and effective data augmentation is still under-explored

Neural Machine Translation (NMT)

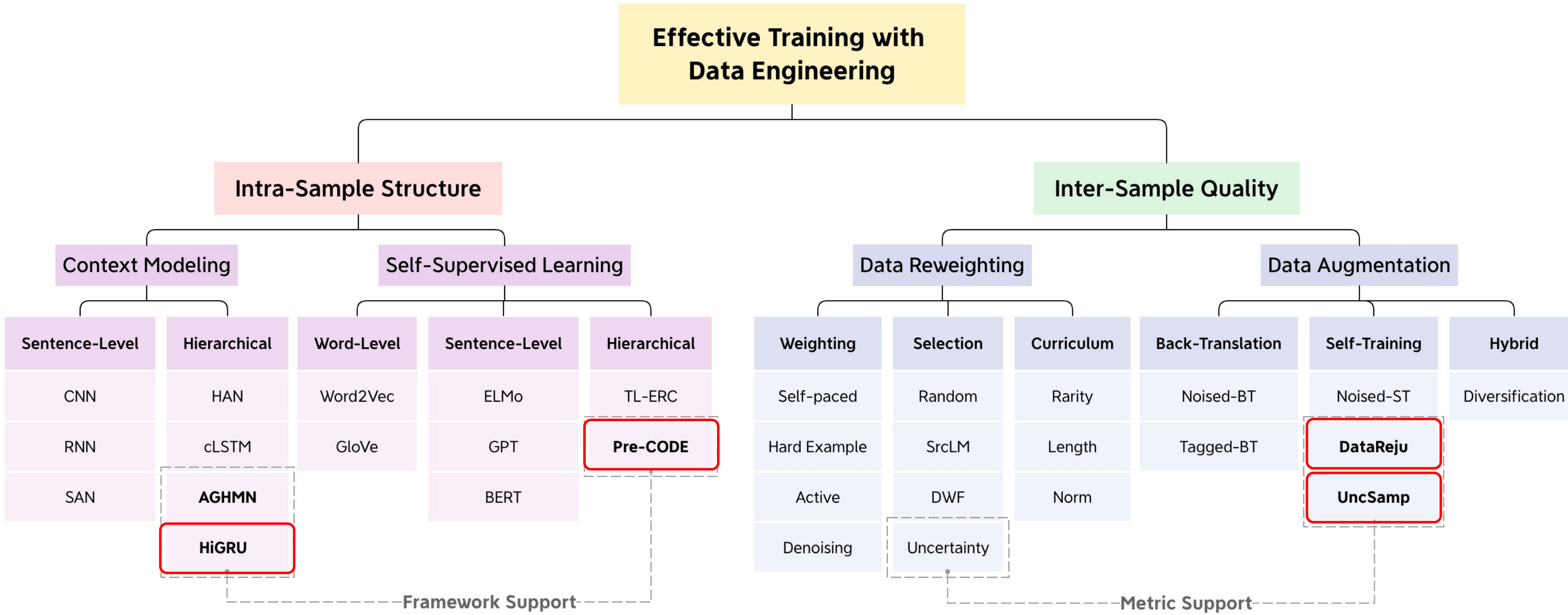


Reasons

- Classic NLG task with complete evaluation criteria
- Large-scale datasets by automatic annotation
- Data selection and augmentation are active areas

Overall Taxonomy

- Our contributions



Ch3: Context enhancement (HiGRU) [NAACL'19]

Ch4: Self-supervised learning (Pre-CODE) [EMNLP'20]

Ch5: Data rejuvenation (DataReju) [EMNLP'20]

Ch6: Self-training sampling (UncSamp) [ACL'21]

Outline

Motivation

- Context is important for capturing accurate meaning
  - The same word or sentence may express different emotions in different contexts

Speaker	Utterance	Emotion
Rachel	Oh okay, I'll fix that to. What's her email address?	Neutral
Ross	Rachel!	Anger
Rachel	All right, I promise. I'll fix this. I swear. I'll-I'll- I'll-I'll talk to her.	Non-neutral
Ross	**Okay!**	**Anger**
Rachel	**Okay.**	**Neutral**
Nurse	This room's available.	Neutral
Rachel	**Okay!**	**Joy**
Rachel	Okay wait!	Non-neutral
Rachel	You listen to me!	Anger

The word "okay" exhibits different emotions in the American television sitcom, Friends.

- Previous cLSTM on hierarchical context
  - Sentential context is not well captured
  - Long-range context is ignored
  - Not an end-to-end model

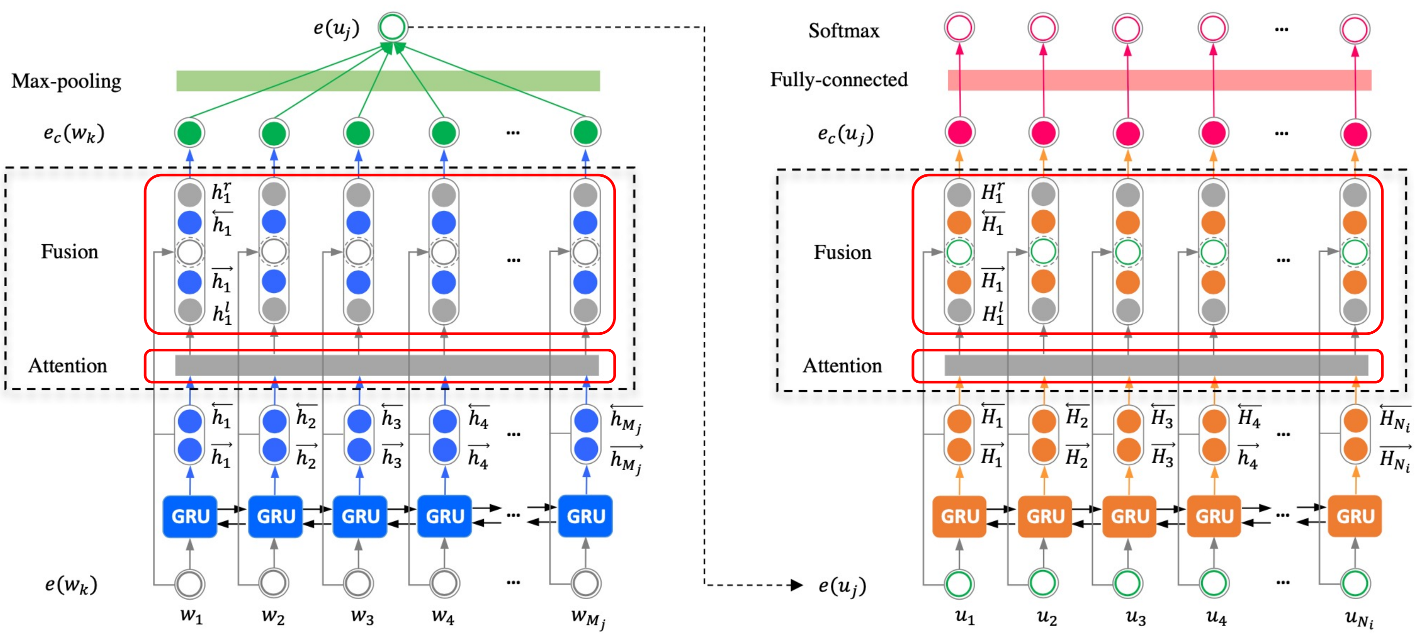
Motivation

- Research problem
  - Context enhancement: exploit the hierarchical structure of conversations to capture various contexts so as to improve the ERC task

- Main findings
  - Proposed a hierarchical gated recurrent unit (HiGRU) to exploit both the context of words and the context of utterances
  - Promoted HiGRU to two progressive variants, HiGRU-f and HiGRU-sf, to effectively incorporate the individual word- and utterance-level information and the long-range contextual information, respectively
  - Achieved consistent improvements over SOTA methods on three ERC datasets, namely, IEMOCAP, Friends and EmotionPush

Approach: Hierarchical Gated Recurrent Unit

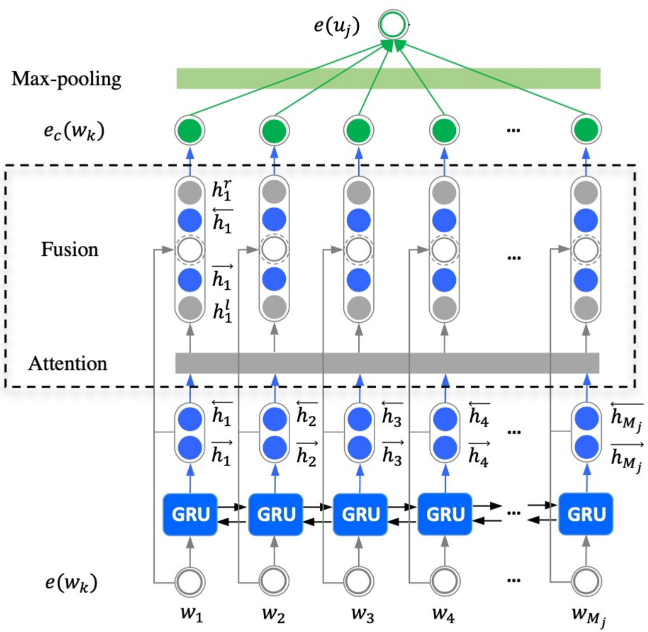
- Overall framework: HiGRU $\xrightarrow{+\text{ Fusion}}$ $\boxed{\text{HiGRU-f}}$ $\xrightarrow{+\text{ Attention}}$ $\boxed{\text{HiGRU-sf}}$
  - Word-level context: a bidirectional gated recurrent unit (GRU)
  - Utterance-level context: another bidirectional GRU



Individual Utterance Embedding **Contextual Utterance Embedding**

Approach: Hierarchical Gated Recurrent Unit

- Word-level context



$$e(\mathbf{x}_j) = \text{maxpool}\left(\{e_c(w_k)\}_{k=1}^{M_j}\right)$$

$$e_c(w_k) = \tanh(W_w \cdot hs + b_w)$$

feature fusion $[h_k^l; \overrightarrow{h_k}; e(w_k); \overleftarrow{h_k}; h_k^r]$

self-attention

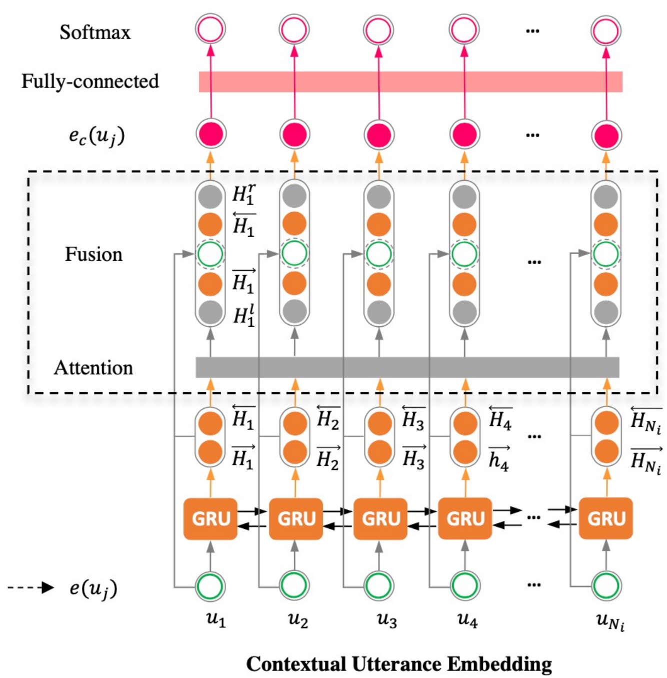
$$hs = [\overrightarrow{h_k}; \overleftarrow{h_k}]$$

$$\overrightarrow{h_k} = (e(w_k), \overrightarrow{h_{k-1}})$$

$$\overleftarrow{h_k} = (e(w_k), \overleftarrow{h_{k+1}})$$

Approach: Hierarchical Gated Recurrent Unit

- Utterance-level context



$$\hat{y}_j = \mathrm{softmax}(W_{fc} \cdot e_c(\mathbf{x}_j) + b_{fc})$$

$$e_c(\mathbf{x}_j) = \tanh(W_u \cdot Hs + b_u)$$

feature fusion $[H_j^l; \overrightarrow{H_j}; e(\mathbf{x}_j); \overleftarrow{H_j}; H_j^r]$

self-attention

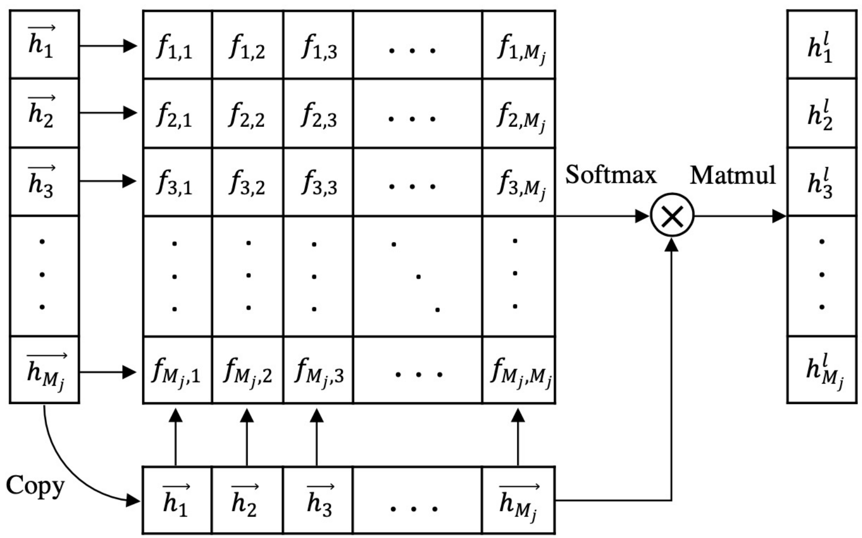
$$Hs = [\overrightarrow{H_j}; \overleftarrow{H_j}]$$

$$\overrightarrow{H_j} = (e(\mathbf{x}_j), \overrightarrow{H_{j-1}})$$
$$\overleftarrow{H_j} = (e(\mathbf{x}_j), \overleftarrow{H_{j+1}})$$

Approach: Hierarchical Gated Recurrent Unit

- Self-attention mechanism
  - Enable the capturing of long-range context

$$f(\overrightarrow{h_k}, \overrightarrow{h_p}) = \begin{cases} \overrightarrow{h_k}^\top \overrightarrow{h_p}, & \text{if } k, p \leq M_j \\ -\infty, & \text{otherwise} \end{cases}$$



Summarize from all positions:

$$h_k^l = \sum_{p=1}^{M_j} a_{kp} \overrightarrow{h_p}$$

$$a_{kp} = \frac{\exp(f(\overrightarrow{h_k}, \overrightarrow{h_p}))}{\sum_{p'=1}^{M_j} \exp\left(f(\overrightarrow{h_k}, \overrightarrow{h_{p'}})\right)}$$

Experiments: Setup

- Datasets
  - IEMOCAP, Friends, and EmotionPush

Dataset	Emotion				
	Ang	Hap/Joy	Sad	Neu	Others
IEMOCAP	1,090	1,627	1,077	1,704	0
Friends	759	1,710	498	6,530	5,006
EmotionPush	140	2,100	514	9,855	2,133

Emotion distributions of the three datasets

- Evaluation metrics
  - Weighted accuracy (WA), unweighted accuracy (UA)
- Compared baselines
  - CNN-DCNN (SocialNLP@ACL'18), SA-BiLSTM (SocialNLP@ACL'18)
  - bcLSTM (ACL'17), CMN (NAACL'18)

Experiments: Main Results

- IEMOCAP
  - The bidirectional GRU is more effective in capturing word-level context than CNNs
  - The long-range context captured by self-attention brings additional benefits

Model (Feat)	Ang	Hap	Sad	Neu	WA	UWA
bcLSTM [4] (T)	76.07	78.97	76.23	67.44	73.6	74.6
(T+V+A)	77.98	79.31	78.30	69.92	76.1	76.3
CMN [5] (T)	-	-	-	-	74.1	-
(T+V+A)	**89.88**	81.75	77.73	67.32	77.6	79.1
bcLSTM_* (T)	75.29	79.40	78.07	76.53	77.7(1.1)	77.3(1.4)
bcGRU (T)	77.20	80.99	76.26	72.50	76.9(1.6)	76.7(1.3)
HiGRU (T)	75.41	**91.64**	79.79	70.74	80.6(0.5)	79.4(0.5)
HiGRU-f (T)	76.69	88.91	80.25	75.92	81.5(0.7)	80.4(0.5)
HiGRU-sf (T)	74.78	89.65	**80.50**	**77.58**	**82.1**(0.4)	**80.6**(0.2)

Results on the IEMOCAP dataset

Experiments: Main Results

- Friends, EmotionPush
  - Combining the two training sets brings opposite effects to respective testing sets
  - EmotionPush is more imbalanced and can be alleviated slightly with Friends

Model	Train	Friends (F)						EmotionPush (E)					
		Ang	Joy	Sad	Neu	WA	UWA	Ang	Joy	Sad	Neu	WA	UWA
SA-BiLSTM [81]	F+E	49.1	68.8	30.6	**90.1**	-	59.6	24.3	70.5	31.0	**94.2**	-	55.0
CNN-DCNN [78]	F+E	55.3	71.1	55.3	68.3	-	62.5	45.9	76.0	51.7	76.3	-	62.5
bcLSTM$_*$	F	64.7	69.6	48.0	75.6	72.4(4.2)	64.4(1.6)	32.9	69.9	47.1	78.0	74.7(4.4)	57.0(2.1)
bcGRU	F	69.5	65.4	52.9	74.7	71.7(4.7)	65.6(1.2)	33.7	71.1	57.2	76.1	73.9(2.9)	59.5(1.8)
bcLSTM$_*$	F+E	54.5	75.6	43.4	73.0	70.5(4.5)	61.6(1.6)	52.4	79.1	54.7	73.3	73.4(3.8)	64.9(2.1)
bcGRU	F+E	59.0	78.6	42.3	71.4	70.2(5.1)	62.8(1.4)	49.4	74.8	61.9	72.4	72.1(4.3)	64.6(1.8)
HiGRU	F	66.9	73.0	51.8	77.2	**74.4**(1.7)	67.2(0.6)	55.6	78.1	57.4	73.8	73.8(2.0)	66.3(1.7)
HiGRU-f	F	69.1	72.1	**60.4**	72.1	71.3(2.9)	68.4(1.0)	55.9	78.9	60.4	72.4	73.0(2.2)	66.9(1.2)
HiGRU-sf	F	**70.7**	70.9	57.7	76.2	74.0(1.4)	**68.9**(1.5)	57.5	78.4	64.1	72.5	73.0(1.6)	68.1(1.2)
HiGRU	F+E	55.4	81.2	51.4	64.4	65.8(4.2)	63.1(1.5)	50.8	76.9	69.0	75.7	75.3(1.7)	68.1(1.2)
HiGRU-f	F+E	54.9	78.3	55.5	68.7	68.5(3.0)	64.3(1.2)	**58.3**	79.1	**69.6**	70.0	71.5(2.5)	69.2(0.9)
HiGRU-sf	F+E	56.8	**81.4**	52.2	68.7	69.0(2.0)	64.8(1.3)	57.8	**79.3**	66.3	77.4	**77.1**(1.0)	**70.2**(1.1)

-4.1

+2.1

Results on the Friends and EmotionPush datasets

Experiments: Analysis

- Successful cases

- The ground-truth label seems inappropriate given this parting situation
- HiGRU-sf captures the melancholic atmosphere

Speaker	Utterance	Truth	bcGRU	HiGRU-sf
Scene 1				
Phoebe	Okay. Oh but don't tell them Monica's pregnant because they frown on that.	Neu	Neu	Neu
Rachel	Okay.	Neu	Neu	Neu
Phoebe	Okay.	Neu	Neu	Neu
Scene 2				
Phoebe	Yeah! Sure! Yep! Oh, y'know what? If I heard a shot right now, I'd throw my body on you.	Joy	Ang	Joy
Gary	Oh yeah? Well maybe you and I should take a walk through a bad neighborhood.	Other	/	/
Phoebe	Okay!	Joy	Ang	Joy
Gary	All right.	Neu	Neu	Neu

Speaker	Utterance	Truth	bcGRU	HiGRU-sf
Scene 3				
Female	Can I send you, like videos and stuff? What about when they start walking.	Other	/	/
Male	Yeah yeah yeah.	Sad	Hap	Sad
Male	You you record every second. You record every second because I want to see it all. Okay?	Hap	Hap	Sad
Male	If I don't get to see it now, I get to see it later at least, you know? You've got to keep it all for me; all right?	Other	/	/
Female	Okay.	Sad	Neu	Sad

"Okay" expresses distinct emotions in three different scenes

Summary

- Proposed a hierarchical gated recurrent unit (HiGRU) to exploit both the context of words and the context of utterances
- Promoted HiGRU to two progressive variants, HiGRU-f and HiGRU-sf, to effectively incorporate the individual word- and utterance-level information and the long-range contextual information, respectively
- Achieved consistent improvements over SOTA methods on three ERC datasets, namely, IEMOCAP, Friends and EmotionPush

Outline

Motivation

- Hierarchical text format contains rich information but also makes it more difficult for label annotation
  - Subtle differences between emotions
  - Effect of context

  => Data scarcity issue

Model	Conversation			Utterance		
	Train	Val	Test	Train	Val	Test
IEMOCAP	96	24	31	3,569	721	1,208
Friends	720	80	200	10,561	1,178	2,764
EmotionPush	720	80	200	10,733	1,202	2,807
EmoryNLP	713	99	85	9,934	1,344	1,328
MOSEI*	2,250	300	676	16,331	1,871	4,662

Statistics of labeled datasets for ERC

Unlabeled conversation data has become massively available, e.g., the subtitles of movies and TV shows in OpenSubtitles

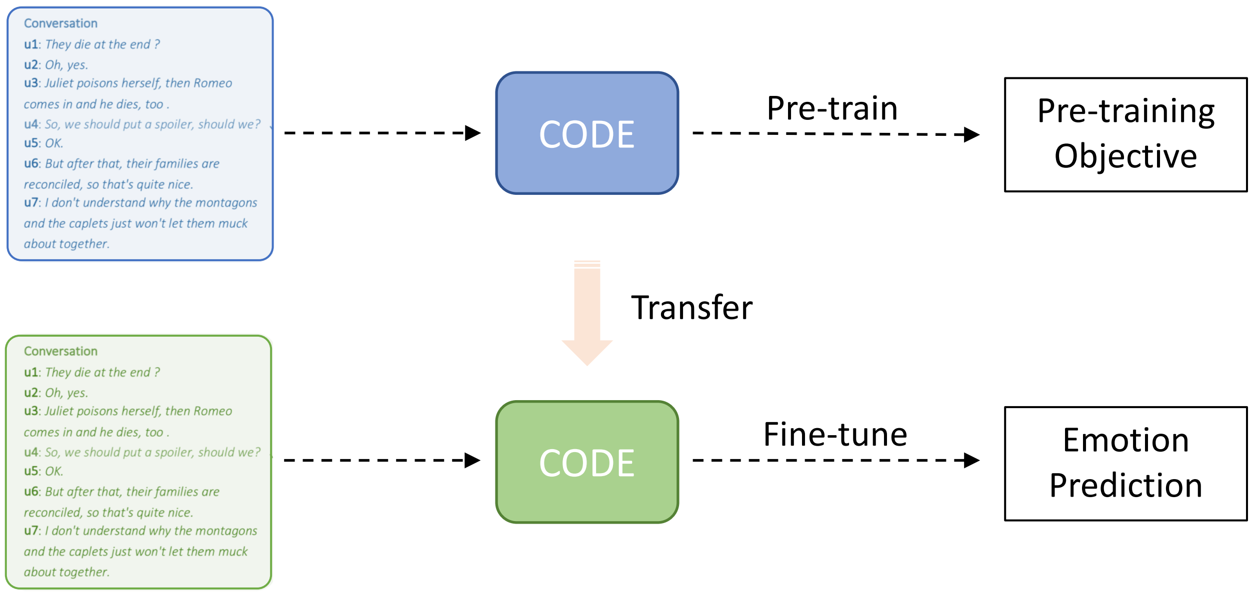
Motivation

- Research problem
  - Data scarcity: leverage unlabeled conversation data in a self-supervised fashion by exploiting the hierarchical structure of conversations

- Main findings
  - Proposed a conversation completion task to pre-train a context-dependent encoder (Pre-CODE) to learn from unlabeled conversation data
  - Fine-tuned the Pre-CODE on the datasets of ERC and achieved significant improvements of the performance over the baselines
  - Demonstrated that both utterance and conversation encoders are well pre-trained and pre-training particularly benefits the prediction of minority classes

Approach: Pre-Training Fine-tuning Paradigm

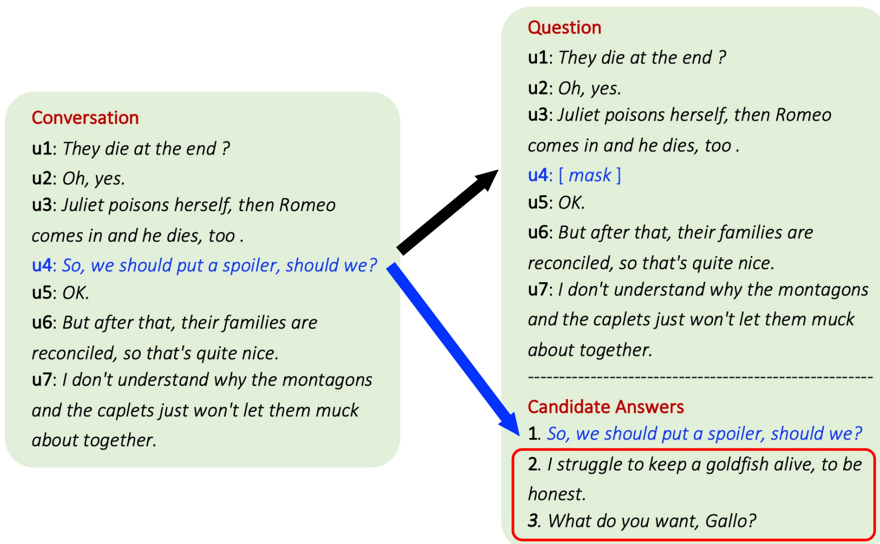
- Basic pipeline
  - Model each conversation with a context-dependent encoder (CODE)
  - Pre-train on the proposed conversation completion (ConvCom) task
  - Fine-tune the pre-trained model on the labeled datasets



Approach: Pre-Training Task

- Conversation completion task

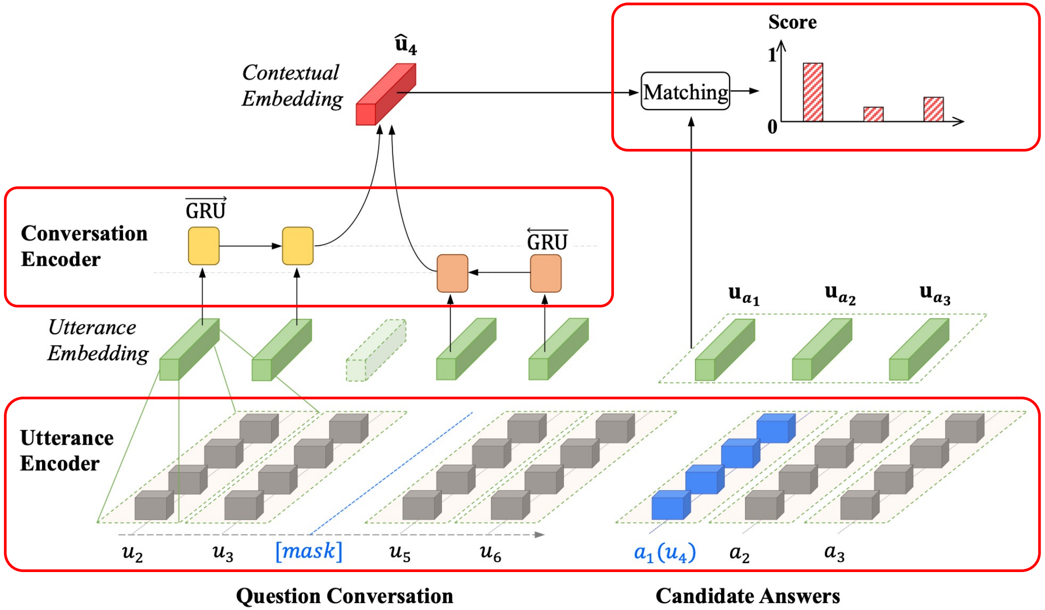
Task definition. Given a conversation, $U = \{u_1, u_2, \ldots, u_L\}$, we mask a target utterance u_l as $U \backslash u_l = \{\ldots, u_{l-1}, [mask], u_{l+1}, \ldots\}$ to create a question, and try to retrieve the correct utterance u_l from the whole training corpus



An example in the conversation completion task

Approach: Architecture

- Context-dependent encoder: a vanilla HiGRU model
  - Fuse the representations of nearby utterances and apply text matching to find the most likely candidate



Noise contrastive estimation: $\mathcal{F} = -\sum_j \left[\log \sigma(\hat{\mathbf{u}}_j^\top \mathbf{u}_{a_1}) + \sum_{n=2}^{N} \log \sigma(-\hat{\mathbf{u}}_j^\top \mathbf{u}_{a_n}) \right]$

Experiments: Data Preparation

- OpenSubtitles 2016 English
  - Remove the first and last 10 utterances for each episode
  - Split the conversations into shorter pieces with 5 to 100 utterances
  - Remove the short conversations such that over half of the utterances contain less than 8 words each
  - Split the data into training, validation, and testing sets by the ratio of 90:5:5

Set	Conversation	Utterance (Avg.)	Word (Avg.)
Train	$\boxed{58360}$	41.3	10.1
Val	3186	41.0	10.1
Test	3297	40.8	10.1

Statistics of the created datasets for the pre-training task

Experiments: Pre-Training Phase

- Evaluation metric
  - $\mathrm{R}_{N'@}k$: the recall of the true positives among k best-matched answers from N' available candidates

- Model scales
  - Small, medium, large: 150, 300, 450 as the hidden sizes

Model	d_u/d_c	$\mathbf{R}_5$@1	$\mathbf{R}_5$@2	$\mathbf{R}_{11}$@1	$\mathbf{R}_{11}$@2
Small	150	70.8	88.0	56.2	72.7
Medium	300	73.8	89.7	60.4	76.4
Large	450	77.2	91.3	64.2	79.1

The performance of CODE in varied capacities on the conversation completion pre-training task

CODE is indeed able to capture the structure of conversations and perform well in the conversation completion task

Experiments: Fine-Tuning Phase

- Evaluation metrics
  - F1 score (F1), weighted accuracy (WA)
- Compared baselines
  - CNN-DCNN (SocialNLP@ACL'18), SA-BiLSTM (SocialNLP@ACL'18)
  - bcLSTM (ACL'17), CMN (NAACL'18), SCNN (AAAI'18), HiGRU (NAACL'19)
  - bcLSTM*, bcGRU, CODE$_{\text{MED}}$

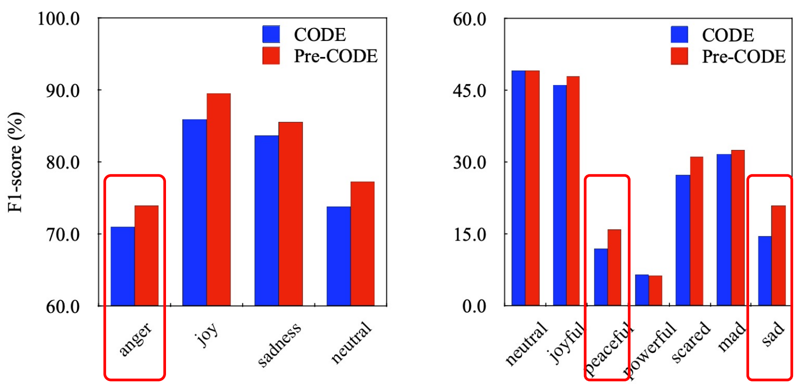
Model	Friends		EmotionPush	
	F1	WA	F1	WA
CNN-DCNN [78];	–	67.0	–	75.7
SA-BiLSTM [81]	–	79.8	–	**87.7**
HiGRU [36]	–	74.4	–	73.8
bcLSTM$_*$	63.1	79.9	60.3	84.8
bcGRU	62.4	77.6	60.5	84.6
CODE-MED	62.4	78.0	60.3	84.2
PRE-CODE	**65.9**	**81.3**	**62.6**	84.7

Model	IEMOCAP		EmoryNLP		MOSEI*	
	F1	WA	F1	WA	F1	WA
bcLSTM [4]	–	73.6	–	–	–	–
CMN [5]	–	74.1	–	–	–	–
SCNN [72]	–	–	26.9	**37.9**	–	–
HiGRU-sf [36]	–	82.1	–	–	–	–
bcLSTM$_*$	76.6	77.1	25.5	33.5	29.1	56.3
bcGRU	77.6	78.2	26.1	33.1	28.7	56.4
CODE-MED	78.6	79.6	26.7	34.7	29.7	56.6
PRE-CODE	**81.5**	**82.9**	**29.1**	36.1	**31.7**	**57.1**

Fine-tuning results on IEMOCAP, EmoryNLP, MOSEI, Friends and EmotionPush

Experiments: Analysis

- Minority classes
  - Pre-training particularly improves the prediction accuracy of minority classes



(a) IEMOCAP (b) EmoryNLP

F1-score of emotion classes

- Layer effects
  - The pre-trained utterance encoder can boost performance
  - Adding the pre-trained conversation encoder brings additional gains

Layers	IEMOCAP	Friends
Pre-CODE + Re-W	**81.6**	64.5
Pre-CODE	81.5	**65.9**
CODE + Pre-U	80.1	64.8
CODE	78.6	62.4

Ablation study on pre-trained layers

Experiments: Analysis

Speaker	Utterance	Truth	CODE	Pre-CODE
Example 1				
Joey	Come on, Lydia, you can do it.	Neu	Neu	Neu
Joey	Push!	Joy	Ang	Ang
Joey	Push 'em out, push 'em out, harder, harder.	Joy	Neu	Neu
Joey	Push 'em out, push 'em out, way out!	Joy	Ang	Joy
Joey	Let's get that ball and really move, hey, hey, ho, ho.	Joy	Neu	Joy
Joey	Let's... I was just... yeah, right.	Joy	Neu	Neu
Joey	Push!	Joy	Ang	Ang
Joey	Push!	Joy	Ang	Ang
Example 2				
Sp1	It's so hard not to cry	Sad	Ang	Sad
Sp2	What happened	Neu	Neu	Neu
Sp1	I lost another 3 set game	Sad	Neu	Sad
Sp2	It's ok person_145	Neu	Neu	Neu
Sp1	Why does it hurt so much	Sad	Neu	Sad
Sp2	Everybody loses	Neu	Neu	Neu

- Pre-trained models also make mistakes when the utterance is too short to provide information

- Pre-training performs better on minority emotion classes, e.g., Sad, here

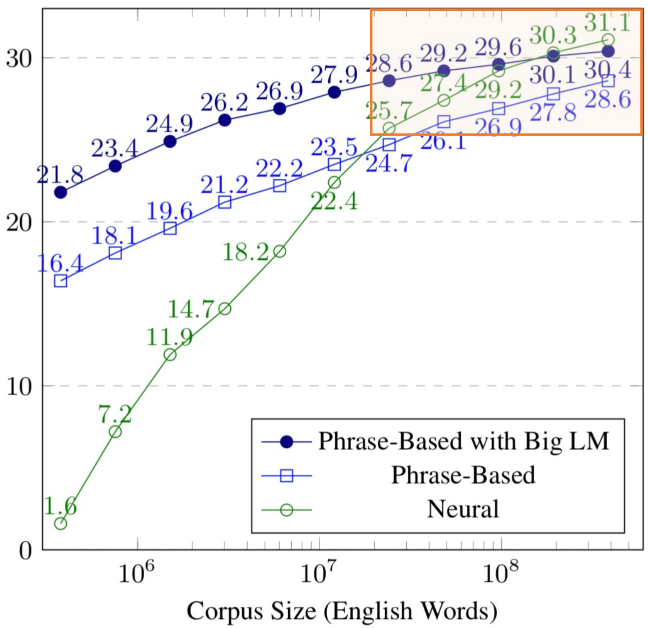
Summary

- Proposed a conversation completion task to pre-train a context-dependent encoder (Pre-CODE) to learn from unlabeled conversation data
- Fine-tuned the Pre-CODE on the datasets of ERC and achieved significant improvements of the performance over the baselines
- Demonstrated that both utterance and conversation encoders are well pre-trained and pre-training particularly benefits the prediction of minority classes

Outline

Motivation

- Data is the fuel of neural machine translation models



BLEU scores with varying amounts of training data (Koehn and Knowles 2017)

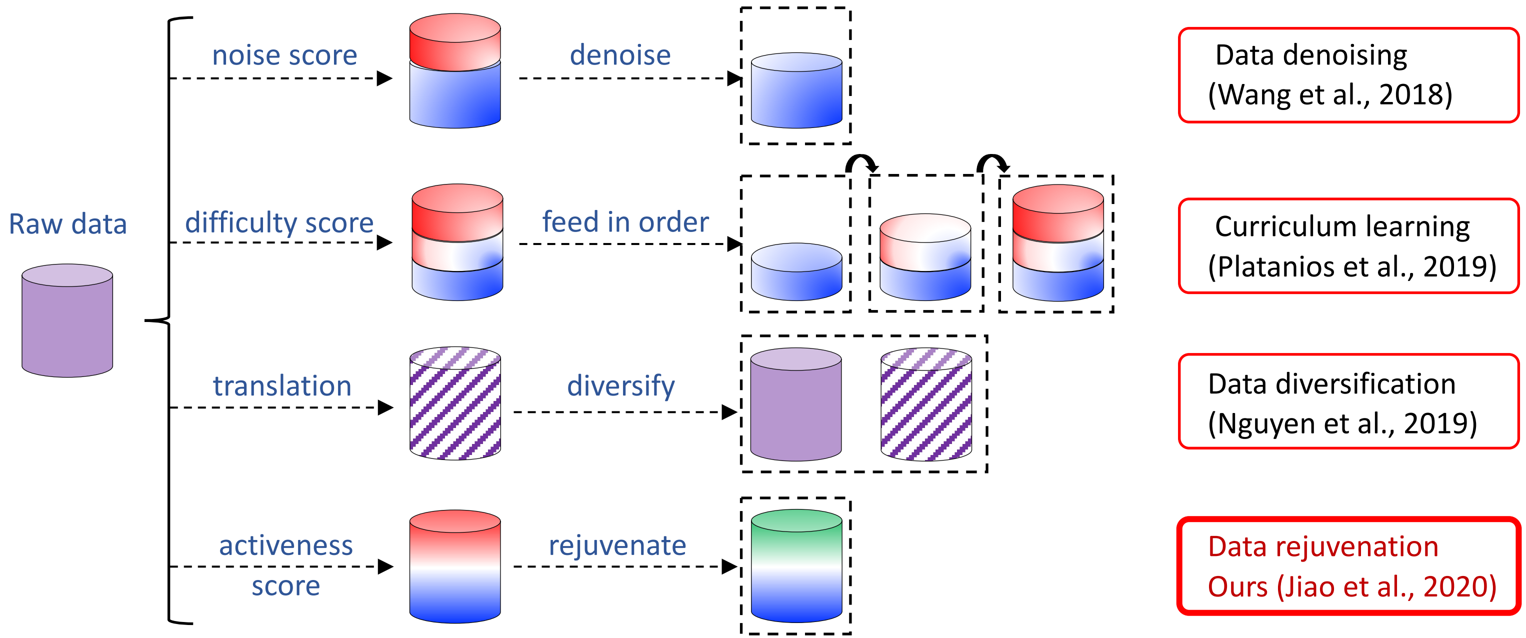
- Challenges on large-scale data
  - Complex patterns
  - Potential noises

  ⇒ Low efficiency

  ⇒ Limited performance

Motivation

- Data manipulation to exploit training data



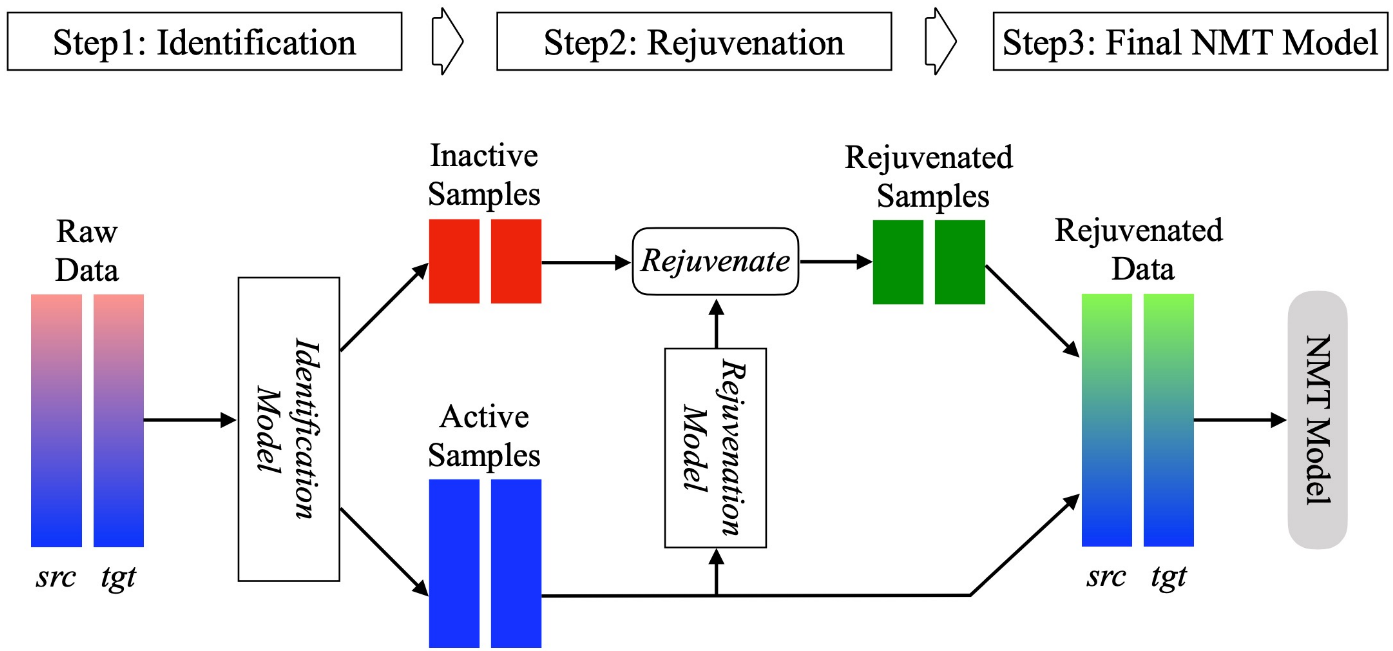
Motivation

- Research problem
  - Inactive samples: training samples that only marginally contribute to or even inversely harm the performance of NMT models

- Main findings
  - Demonstrated the existence of inactive samples in large-scale translation datasets, which mainly depends on the data distribution
  - Proposed a general framework to rejuvenate the inactive samples to improve the training of NMT models
  - Achieved significant improvements over SOTA Transformer and DynamicConv models on WMT14 En-De and En-Fr translation tasks, without model modification

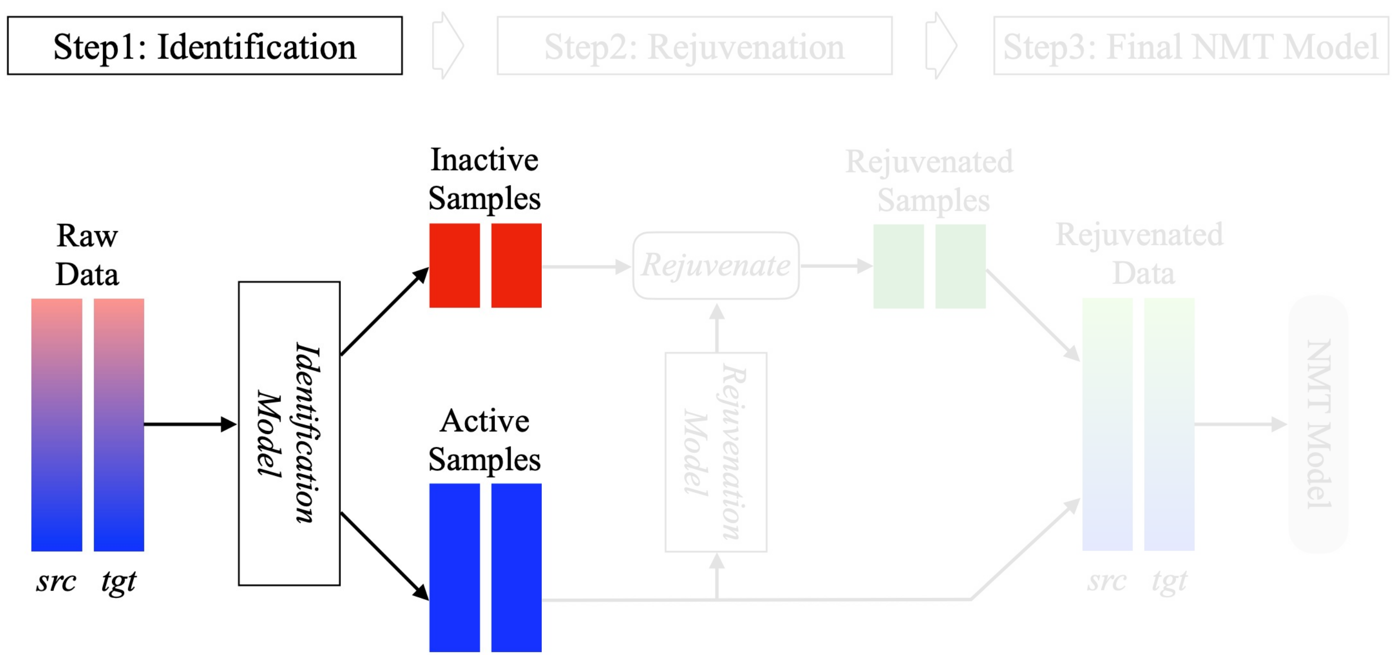
Approach: Data Rejuvenation Framework

- General pipeline
  - Three models: identification model, rejuvenation model, final NMT model



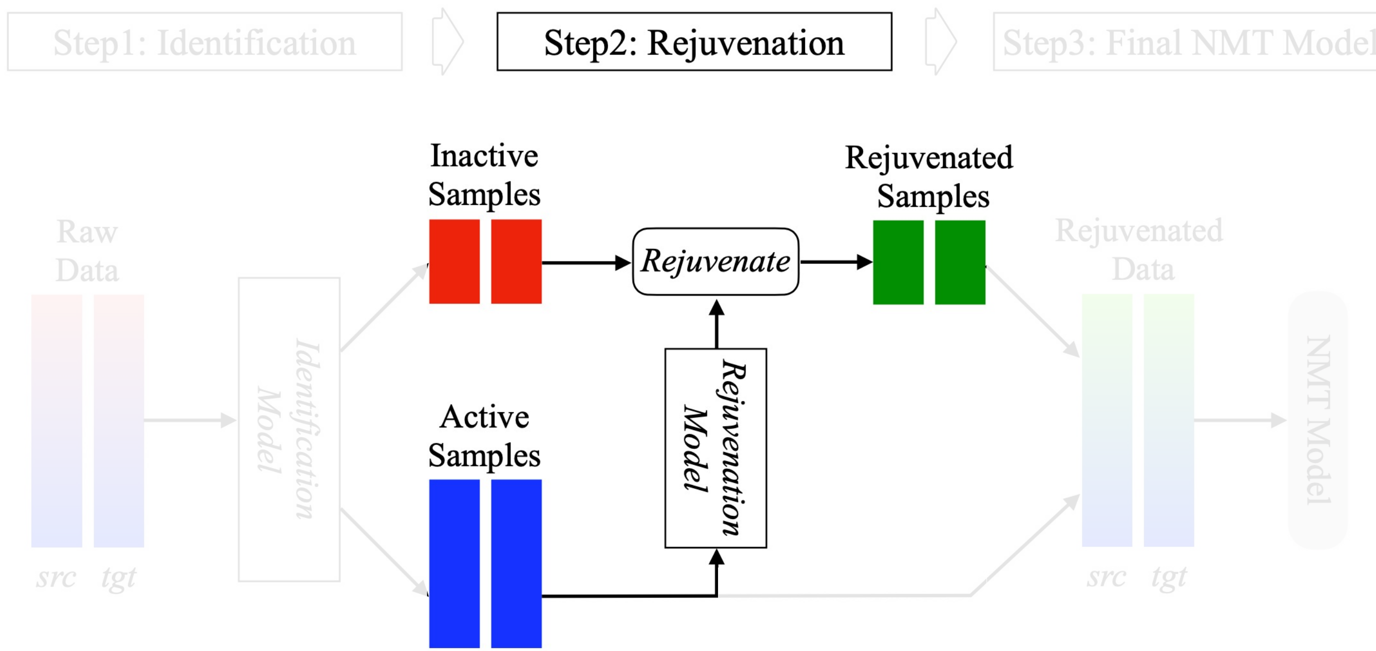
Approach: Data Rejuvenation Framework

- General pipeline
  - Three models: identification model, rejuvenation model, final NMT model



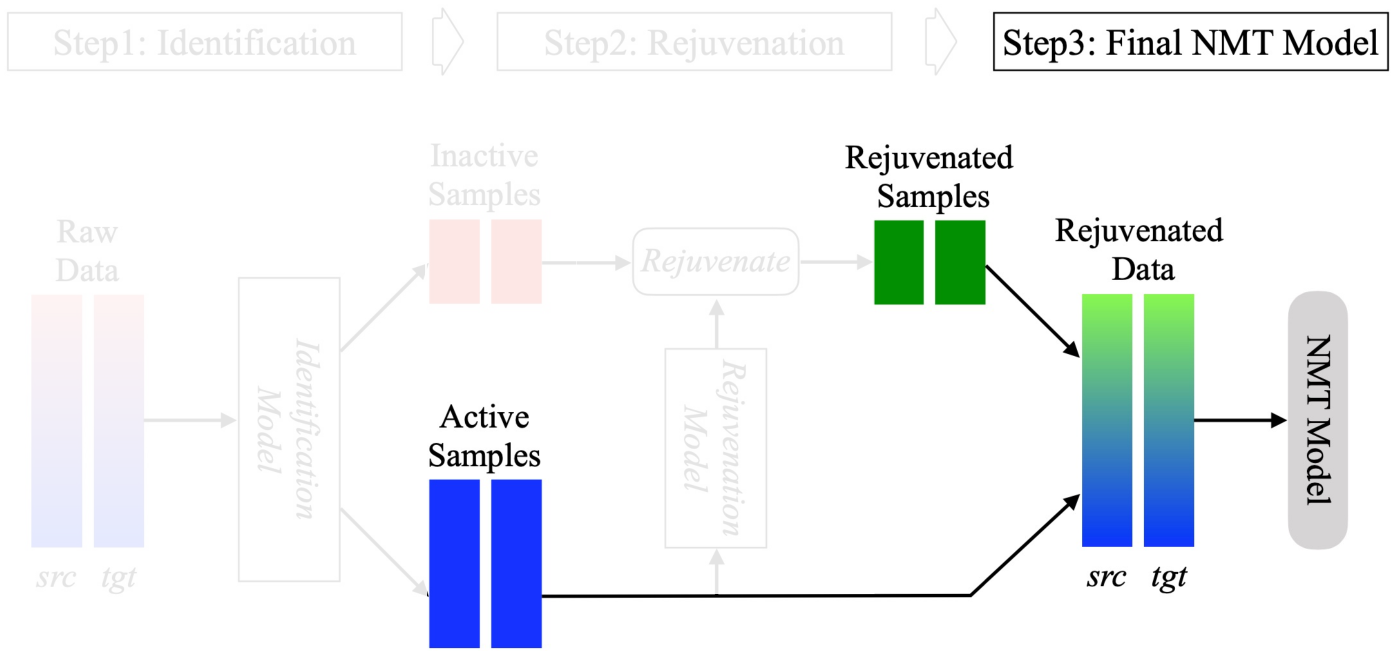
Approach: Data Rejuvenation Framework

- General pipeline
  - Three models: identification model, rejuvenation model, final NMT model



Approach: Data Rejuvenation Framework

- General pipeline
  - Three models: identification model, rejuvenation model, final NMT model

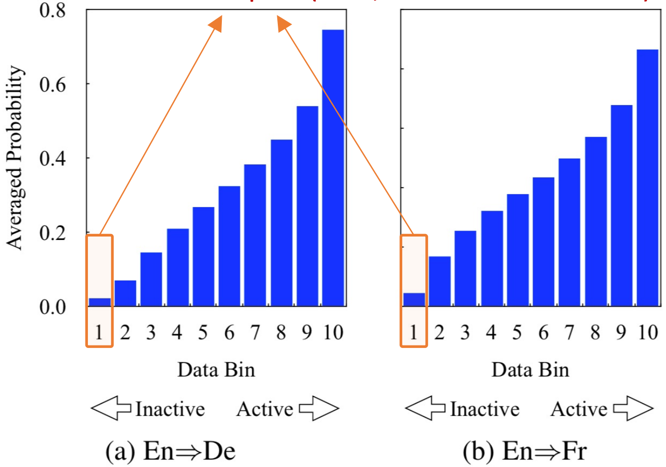


Approach: Identification of Inactive Samples

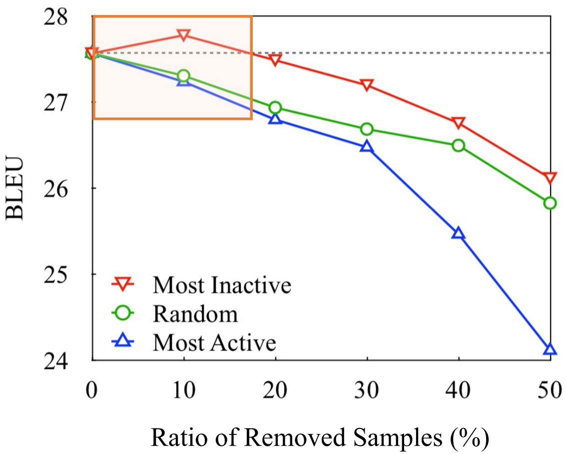
- Identification model: an NMT model trained on raw training data
  - Activeness metric: model output probability

$$I(\mathbf{y}|\mathbf{x}) = \prod_{t=1}^{T} p(y_t|\mathbf{x}, \mathbf{y}_{<t})$$

Inactive samples (10%, Birodkar et al. 2019)



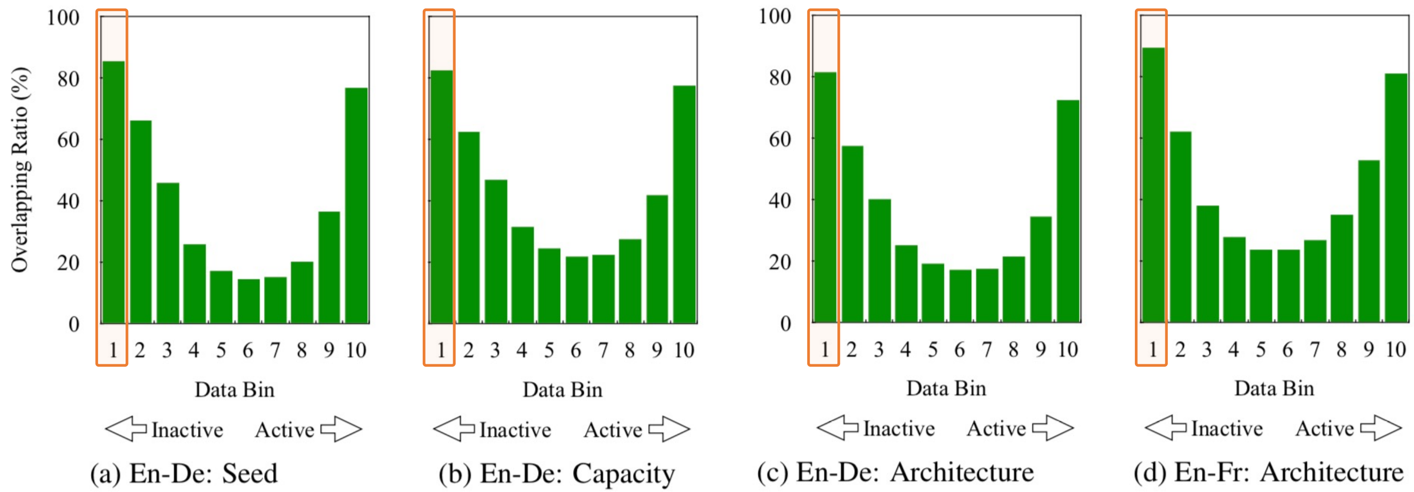
Probability diagram on WMT14 (a) En⇒De and (b) En⇒Fr training data



Translation performance with the most inactive samples removed

Approach: Identification of Inactive Samples

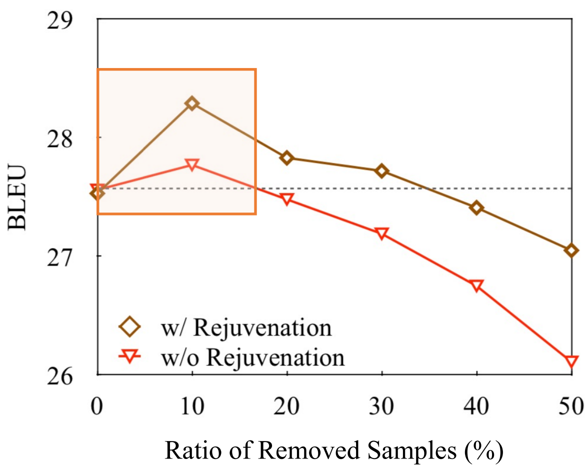
- Consistency: factors that may affect the identification NMT model
  - Random seed: 1, 12, 123, 1234, 12345
  - Model capacity: tiny (256 x 3), base (512 x 6), big (1024 x 6)
  - Architecture: LSTM, Transformer, DynamicConv



Ratio of samples that are shared by different model variants: random seed (a), model capacity (b), model architecture on En⇒De (c) and En⇒Fr (d) datasets

Approach: Rejuvenation of Inactive Samples

- Rejuvenation model: an NMT model trained on active samples
  - Forward translation: simplify the target sentences of inactive samples



Effect of the ratio of samples regarded as inactive samples for rejuvenation

Training Data	BLEU	△
Raw Data	27.5	–
- 10% *Inactive* Samples	27.8	+0.3
+ Rejuvenated Samples	28.3	+0.8
- 10% *Random* Samples	27.4	-0.1
+ Rejuvenated Samples	27.3	-0.2

Comparing data rejuvenation on identified inactive samples and forward translation on randomly selected samples

Experiments: Setup

- Datasets
  - Bitext: WMT14 English=>German (4.5M), English=>French (35.5M)
  - Evaluation: newstest2013 as the valid set, newstest2014 as the test set

- Models
  - LSTM: 32K tokens/batch, 100K steps
  - Transformer-base: 32K tokens/batch, 100K steps; ablation study
  - Transformer-big:
    - Normal: 32K tokens/batch, 300K steps
    - Large-batch: 460K tokens/batch, 30K steps
  - DynamicConv: 32K tokens/batch, 100K steps

- Evaluation metrics
  - BLEU score: n-gram matches between each candidate translation and the reference translations
  - Compare-mt: significance test

Experiments: Main Results

- Comparison with vanilla baseline models

System	Architecture	En⇒De BLEU	En⇒De △	En⇒Fr BLEU	En⇒Fr △
Existing NMT Systems					
Vaswani et al. [14]	Transformer-Base	27.3	–	38.1	–
	Transformer-Big	28.4	–	41.0	–
Ott et al. [15]	Scale Transformer	29.3	–	43.2	–
Wu et al. [91]	DynamicConv	29.7	–	43.2	–
Our NMT Systems					
This work	Lstm	26.5	–	40.6	–
	+ Data Rejuvenation	$27.0^{\uparrow}$	+0.5	$41.1^{\uparrow}$	+0.5
	Transformer-Base	27.5	–	40.2	–
	+ Data Rejuvenation	$28.3^{\Uparrow}$	+0.8	$41.0^{\Uparrow}$	+0.8
	Transformer-Big	28.4	–	42.4	–
	+ Data Rejuvenation	$29.2^{\Uparrow}$	+0.8	$43.0^{\uparrow}$	+0.6
	+ Large Batch	29.6	–	43.5	–
	+ Data Rejuvenation	$30.3^{\Uparrow}$	+0.7	$44.0^{\uparrow}$	+0.5
	DynamicConv	29.7	–	43.3	–
	+ Data Rejuvenation	$30.2^{\uparrow}$	+0.5	$43.9^{\uparrow}$	+0.6

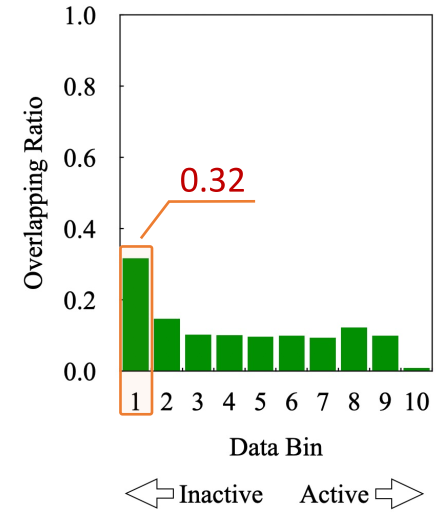
Evaluation of translation performance across model architectures and language pairs. "↑ / ⇑": indicate statistically significant improvement over the corresponding baseline $p < 0.05/0.01$ respectively

Experiments: Main Results

- Comparison with related data manipulation methods

Model	BLEU	△
Transformer-Base	27.5	–
+ *Data Rejuvenation*	28.3	+0.8
+ Data Diversification-BT	26.9	-0.6
+ *Data Rejuvenation*	27.9	+0.4
+ Data Diversification-FT	28.1	+0.6
+ *Data Rejuvenation*	28.5	+1.0
+ Data Denoising	28.1	+0.6
+ *Data Rejuvenation*	28.6	+1.1

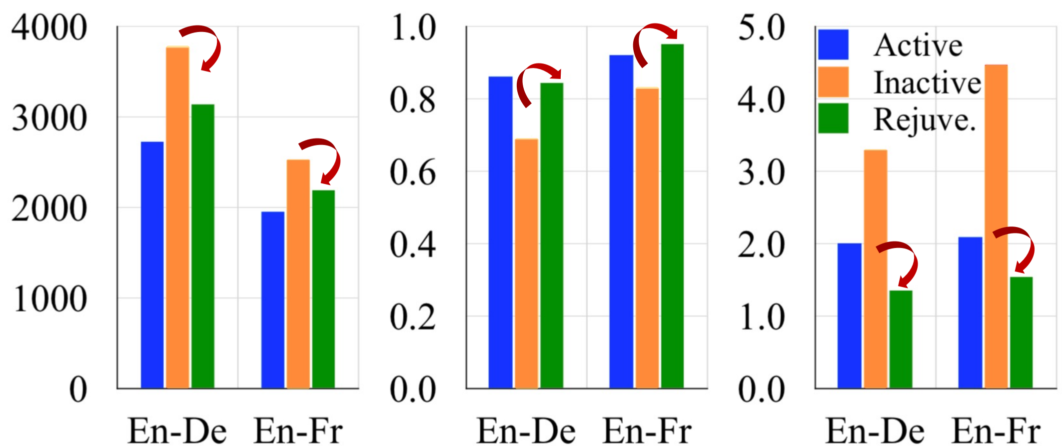
Evaluation of Comparison with related data manipulation approaches. Results are reported on the En⇒De test set.



Overlapping of samples in the order identified by us and that by data denoising

Experiments: Analysis

- Linguistic properties



(a) Frequency Rank (b) Coverage (c) Uncertainty

Linguistic properties of different training samples: frequency rank (↑ more difficult), coverage (↓ more difficult), and uncertainty (↑ more difficult)

Experiments: Analysis

- Inactive sample cases

	Side	Sentence
En⇒De	X	The Second World War finished the destruction of the first .
	Y	Der zweite Weltkrieg tat dann das seine und zerstörte den Rest . =>En: The Second World War then did his and destroyed the rest .
	Y'	Der Zweite Weltkrieg beendete die Zerstörung des ersten . =>En: The Second World War ended the destruction of the first .
En⇒Fr	X	Anything denied by the latter was effectively confirmed as true .
	Y	Tout ce que démentait cette agence se révélait dans la pratique bien réel . =>En: Everything that this agency denied turned out to be very real in practice .
	Y'	Toute chose niée par ce dernier a été effectivement confirmée comme vraie . =>En: Anything denied by the latter has actually been confirmed to be true .

reasoning

passive voice

active voice

Inactive samples from the training sets of En⇒De and En⇒Fr:

- X, Y and Y' represent the source sentence, target sentence, and the rejuvenated target sentence, respectively
- Y and Y' are also translated into English (=>En:) by Google Translate for reference
- For either sample, the underlined phrases correspond to the same content

Summary

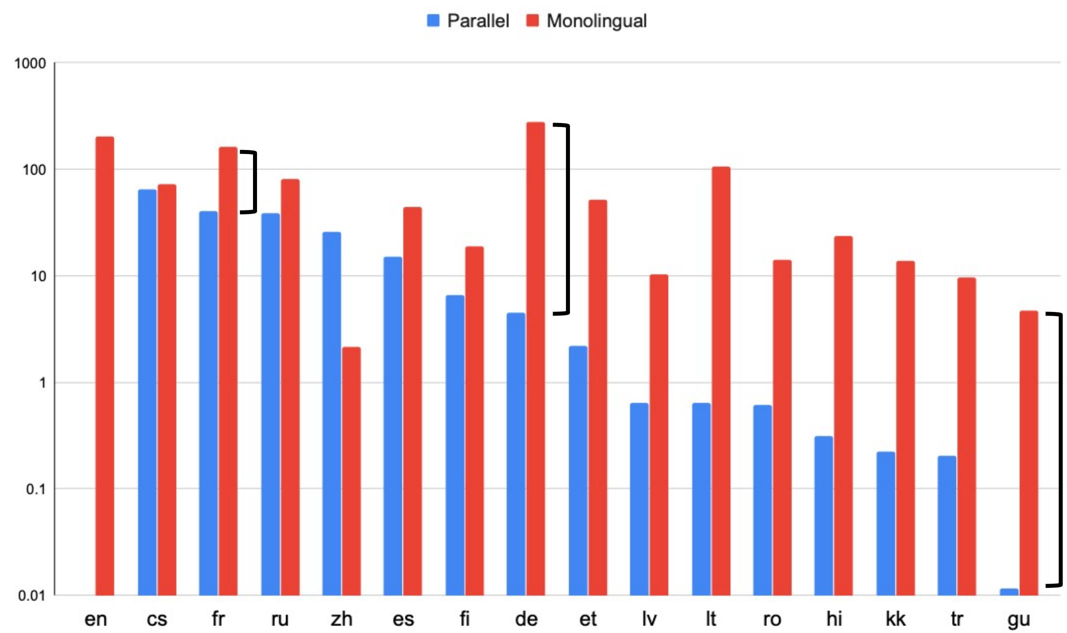
- Demonstrated the existence of inactive samples in large-scale translation datasets, which mainly depends on the data distribution
- Proposed a general framework to rejuvenate the inactive samples to improve the training of NMT models
- Achieved significant improvements over SOTA Transformer and DynamicConv models on WMT14 En-De and En-Fr translation tasks, without model modification

Outline

- ❑ Introduction
- ❑ Context Enhancement with Intra-Sample Structure Mining
- ❑ Self-Supervised Learning with Intra-Sample Structure Mining
- ❑ Data Rejuvenation with Inter-Sample Quality Mining
- ❑ **Self-Training Sampling with Inter-Sample Quality Mining**
- ❑ Conclusion

Motivation

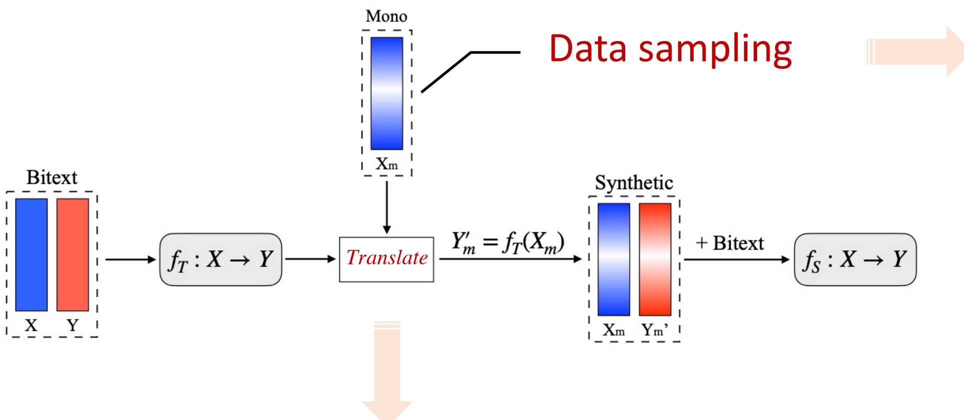
- Unlabeled monolingual data is a huge resource for NMT

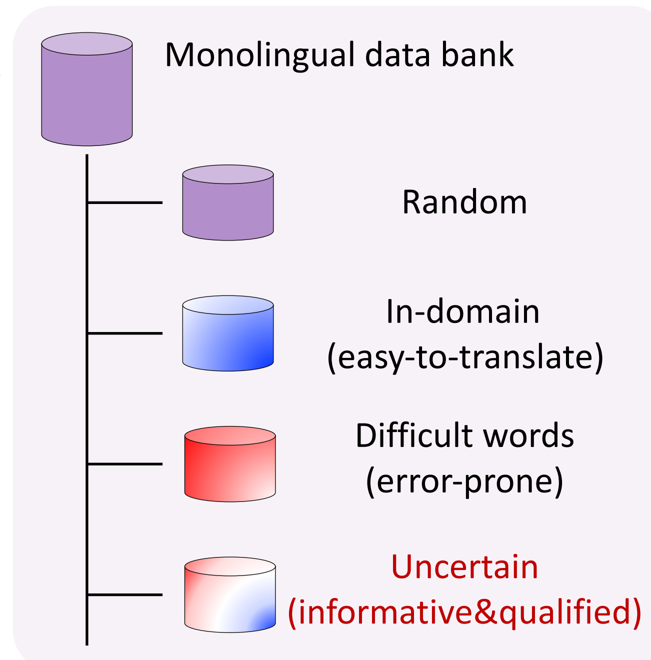


Number of parallel and monolingual training samples in millions for languages in WMT training corpora (Siddhant et al. ACL2020)

Motivation

- Leverage monolingual data by data augmentation
  - Self-training: pair each monolingual sentence at source-side with a synthetic sentence at target-side by translating





  - Inefficient translation:
    - GPU resource
    - Time-consuming generation

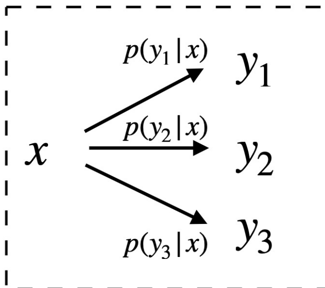
Motivation

- Research Problem
  - Monolingual data sampling: select a subset from the large-scale monolingual data bank for self-training to further improve the performance of NMT models

- Main findings
  - Demonstrated that random sampling is sub-optimal when performing self-training
  - Proposed an uncertainty-based sampling strategy to prefer monolingual sentences with relatively high uncertainty
  - Achieved significant improvements on large-scale translation tasks, WMT English=>German and English=>Chinese, with large-scale monolingual data
  - Demonstrated that the proposed approach particularly improves the translation quality of uncertain sentences, and the prediction accuracy of low-frequency words

Preliminary

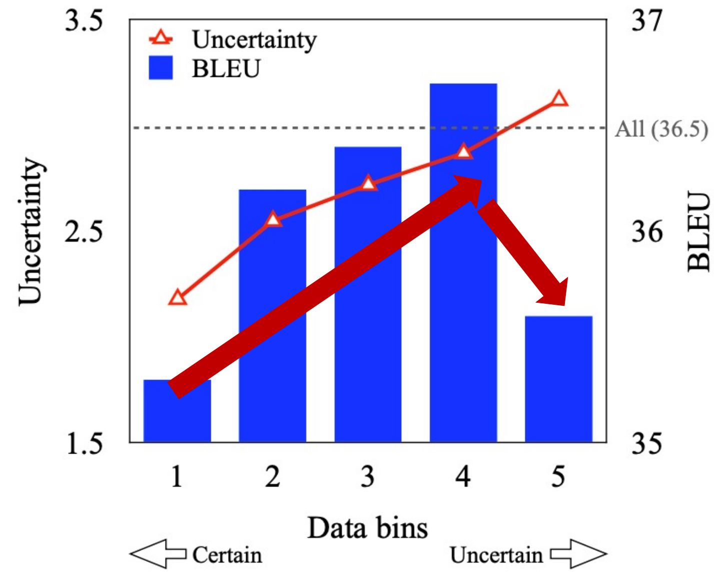
- Monolingual data uncertainty
  - Translation entropy of a source sentence to the target language



Bilingual Dictionary

$$\mathrm{U}(\mathbf{x}^j|\mathcal{A}_b) = \frac{1}{T_x}\sum_{t=1}^{T_x}\mathcal{H}(y|\mathcal{A}_b, x = x_t),$$

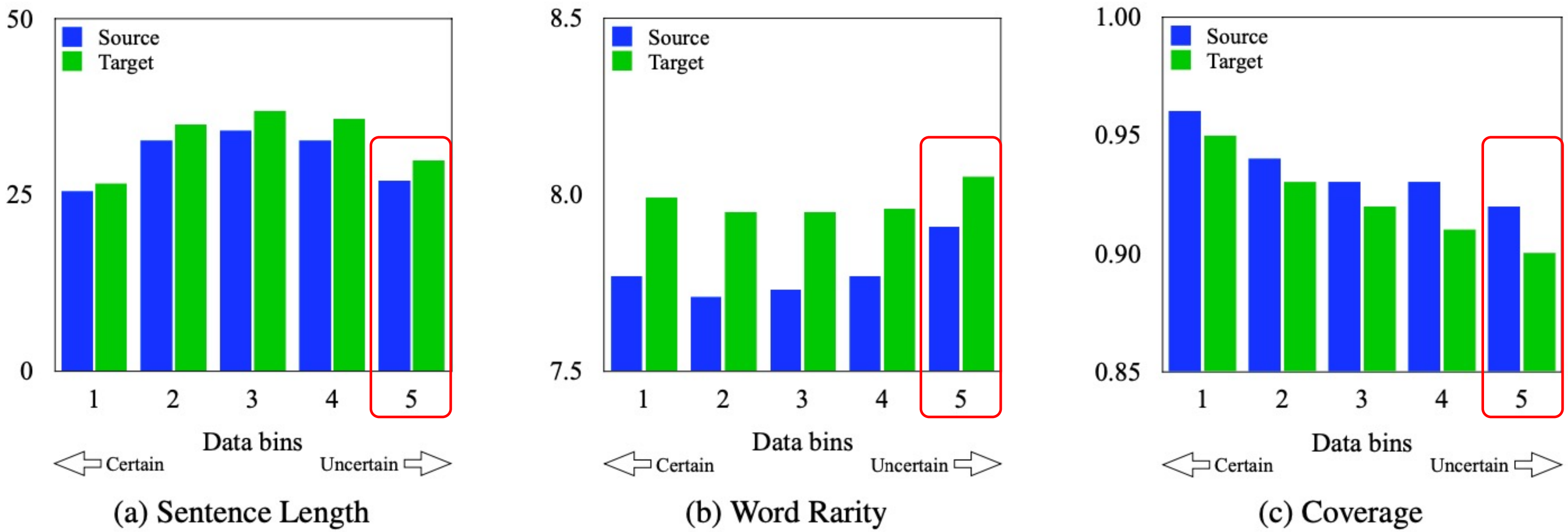
$$\mathcal{H}(y|\mathcal{A}_b, x_i) = -\sum_{y_j \in \mathcal{A}_b(x_i)} p(y_j|x_i)\log p(y_j|x_i).$$



Translation performance vs. the uncertainty of monolingual data, evaluated on WMT19 En-De

Preliminary

- Linguistic properties versus translation uncertainty
  - Monolingual sentences with higher uncertainty are usually longer (except for bin 5)
  - Bin 5 contains noticeably more rare words than the other bins
  - The overall coverage in bin 5 is the lowest



Comparison of monolingual sentences with varied uncertainty in terms of linguistic properties

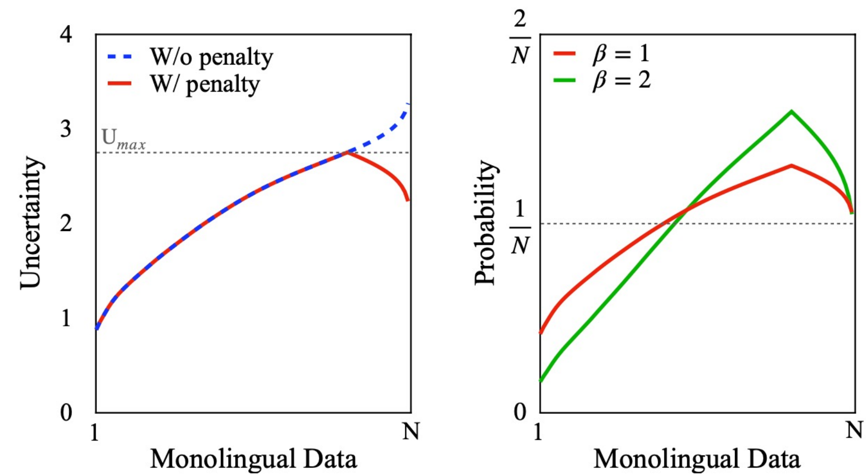
Approach: Uncertainty-Based Sampling

- Sample monolingual sentences according to translation uncertainty
  - Prefer high-uncertainty sentences
  - Penalize sentences with excessively high uncertainty

$$p = \frac{\left[\alpha \cdot \mathrm{U}(\mathbf{x}^j|\mathcal{A}_b)\right]^{\beta}}{\sum_{\mathbf{x}^j \in \mathcal{M}_x}\left[\alpha \cdot \mathrm{U}(\mathbf{x}^j|\mathcal{A}_b)\right]^{\beta}},$$

$$\alpha = \begin{cases} 1, & \mathrm{U}(\mathbf{x}^j|\mathcal{A}_b) \leq \mathrm{U}_{max}, \\ max(\frac{2\boxed{\mathrm{U}_{max}}}{\mathrm{U}(\mathbf{x}^j|\mathcal{A}_b)} - 1, 0), & \text{otherwise.} \end{cases}$$

- U_{max}: threshold for penalizing
  - Uncertainty of the sample at the R% percentile of the parallel data
  - For En=>De, $U_{max}(R = 90) = 2.90$
- β: steepness of distribution

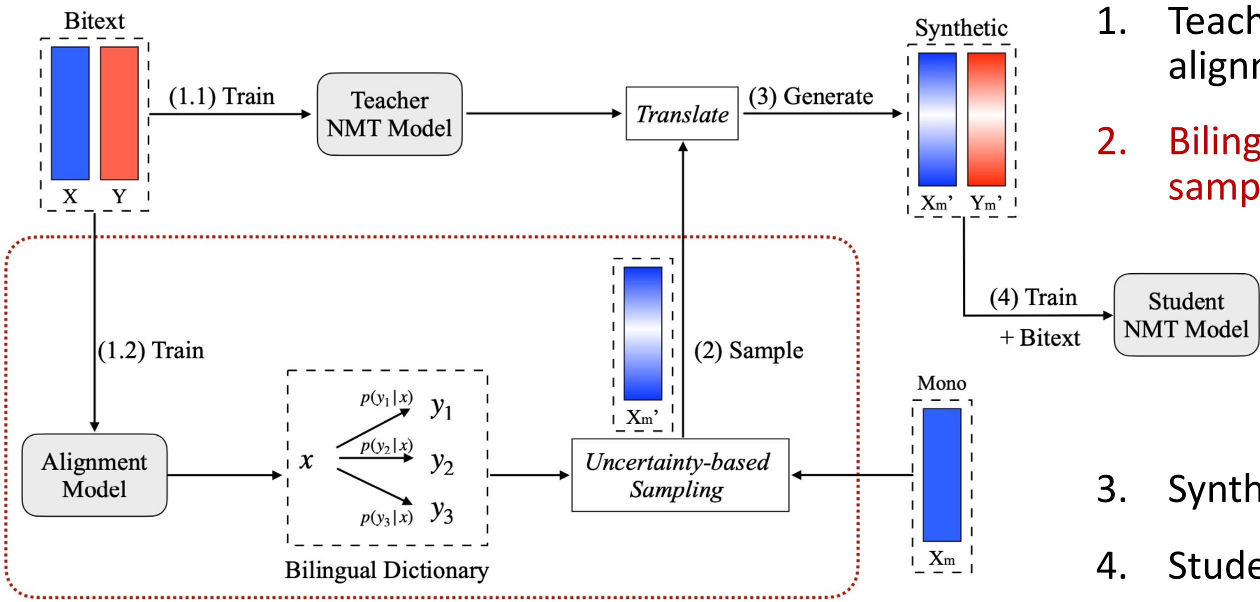


(a) Uncertainty (b) Sampling Probability

Distribution of modified monolingual uncertainty and sampling probability

Approach: Uncertainty-Based Sampling

- Overall framework
  - Add only one step to the standard self-training pipeline, i.e., data sampling



1. Teacher NMT model and alignment model
2. Bilingual dictionary and sampling
3. Synthetic parallel data
4. Student NMT model

Experiments: Setup

- Datasets
  - Bitext: WMT English=>German (36.8M), English=>Chinese (22.1M)
  - Monolingual: newscrawl 2011 – 2019 English (200M)
  - Evaluation: newstest2018 as the valid set, newstest2019/2020 as test sets

- Models
  - Transformer-base: ablation study; 32K tokens/batch, 150K steps
  - Transformer-big: large-scale scenario; 460K tokens/batch, 30K steps

- Evaluation metrics
  - SacreBLEU: detokenized BLEU score
  - Compare-mt: significance test, output analysis

Experiments: Constrained Scenario

- Hyper-parameters
  - Best values: $R = 90,\ \beta = 2$
  - $\mathrm{U}_{max}(R = 90) = 2.90$

BLEU		R		
		100	90	80
β	1	36.6	36.7	36.6
	2	36.7	**36.9**	36.6
	3	36.5	36.5	36.5

Translation performance vs. β and R on En=>De

- Comparison with related methods
  - Random sampling (RandSamp)
  - Difficult word by frequency (DWF)
  - Source language model (SrcLM)
  - Our uncertainty-based sampling (UncSamp)

Data	2019	2020	Avg
RandSamp	40.9	31.6	36.2
DWF	39.6	30.1	34.8
SrcLM	41.1	32.0	36.5
UncSamp	41.6	32.3	36.9
+ Filtering	41.5	32.7	**37.1**

Comparison with related methods on En=>De

Experiments: Unconstrained Scenario

- Sample from the full set of large-scale monolingual data (200M)
  - Transformer-big + large batch training
  - Full bitext training data

System	Data	En⇒De			En⇒Zh		
		2019	**2020**	**Avg**	**2019**	**2020**	**Avg**
Wu et al. (2019b)	BITEXT	37.3	–	–	–	–	–
	+RANDSAMP	39.8	–	–	–	–	–
Shi et al. (2020)	BITEXT	–	–	–	–	38.6	–
	+RANDSAMP	–	–	–	–	41.9	–
This Work	BITEXT	39.6	31.0	35.3	37.1	42.5	39.8
	+RANDSAMP	41.6	33.1	37.3	37.6	43.8	40.7
	+SRCLM	41.7	33.1	37.4	37.3	44.0	40.7
	+UNCSAMP	42.5$^{\Uparrow}$	34.4$^{\Uparrow}$	**38.4**	38.2$^{\Uparrow}$	44.3$^{\uparrow}$	**41.3**

Translation performance on WMT En⇒De and WMT En⇒Zh test sets. Our UncSamp approach achieves further improvements over the RandSamp method.

Experiments: Analysis

- Understand which aspects of translation outputs are improved
  - Uncertain sentences
  - Low-frequency words

Unc	BITEXT	RANDSAMP	UNCSAMP BLEU	UNCSAMP △(%)
Low	38.1	39.7	41.5	8.9
Med	34.2	36.7	37.4	9.3
High	31.0	33.4	34.4	**10.9**

Translation performance vs. sentence uncertainty

Freq	BITEXT	RANDSAMP	UNCSAMP Fmeas	UNCSAMP △(%)
Low	52.3	53.8	54.7	**4.5**
Med	65.2	66.5	66.9	2.6
High	70.3	71.6	72.0	2.4

Prediction accuracy vs. word frequency

Uncertain monolingual sentences contain more medium- to low-frequency words at the target side

Summary

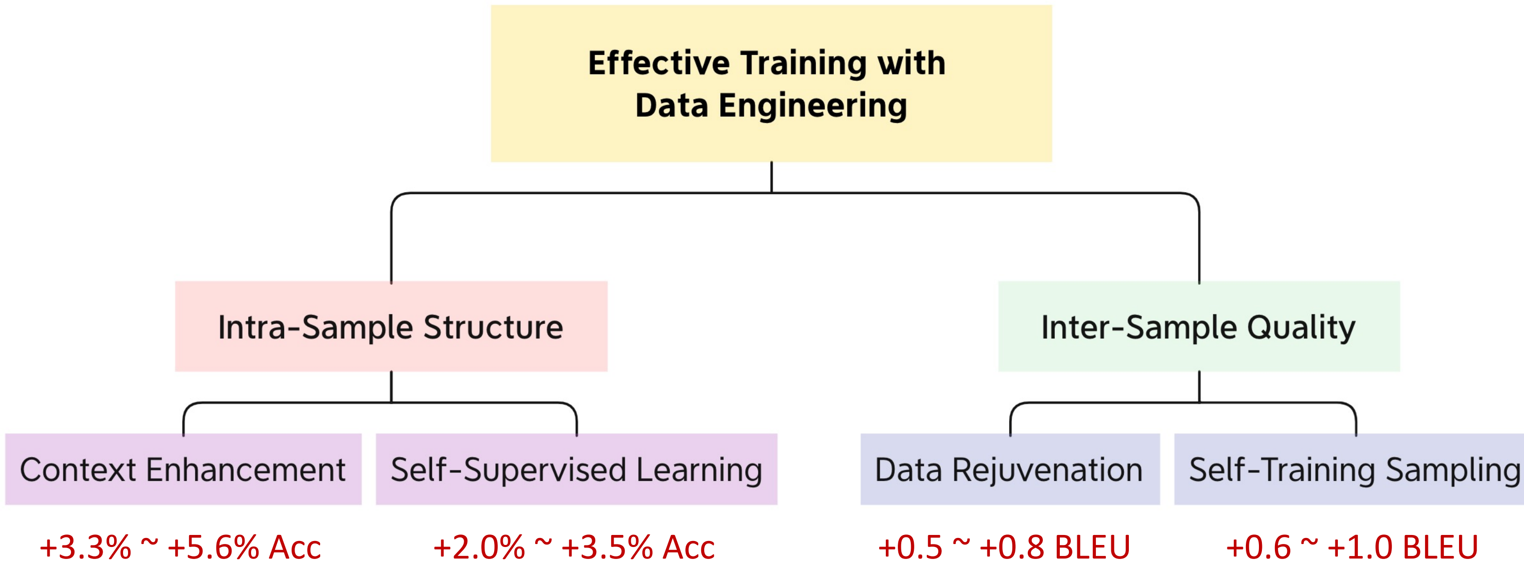
- Demonstrated that random sampling is sub-optimal when performing self-training
- Proposed an uncertainty-based sampling strategy to prefer monolingual sentences with relatively high uncertainty
- Achieved significant improvements on large-scale translation tasks, WMT English=>German and English=>Chinese, with large-scale monolingual data
- Demonstrated that the proposed approach particularly improves the translation quality of uncertain sentences, and the prediction accuracy of low-frequency words

Outline

- ❑ Introduction
- ❑ Context Enhancement with Intra-Sample Structure Mining
- ❑ Self-Supervised Learning with Intra-Sample Structure Mining
- ❑ Data Rejuvenation with Inter-Sample Quality Mining
- ❑ Self-Training Sampling with Inter-Sample Quality Mining
- ❑ **Conclusion**

Conclusion

- We exploit the training data with intra-sample structure and inter-sample quality information for effective training of NLP models



All the four studies consistently improve the effectiveness of model training!

Acknowledgement

- **Supervisors:**

  Prof. Irwin King and Prof. Michael R. Lyu

- **Committees:**

  Prof. Kevin Yip, Prof. Xunying Liu, and Prof. Shou-De Lin

- **Mentors** during internship in Tencent AI Lab:

  Dr. Xing Wang and Dr. Zhaopeng Tu

- **Group fellows**
- **My girlfriend:**

  Miss Yuye Wang

- **My family**

Publications During Ph.D. Study

1. **Wenxiang Jiao**, Haiqin Yang, Irwin King, Michael R. Lyu. "HiGRU: Hierarchical Gated Recurrent Units for Utterance- Level Emotion Recognition". In Proceedings of the 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), pp. 397-406, Minneapolis, USA, June 2 - June 7, 2019.
2. **Wenxiang Jiao**, Michael R. Lyu, Irwin King. "Real-Time Emotion Recognition via Attention Gated Hierarchical Memory Network". In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020), pp. 8002-8009, New York, USA, February 7 - February 12, 2020.
3. **Wenxiang Jiao**, Michael R. Lyu, Irwin King. "Exploiting Unsupervised Data for Emotion Recognition in Conversations". In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Findings of EMNLP (EMNLP- Findings 2020), pp. 4839-4846, Online, USA, November 16 - November 20, 2020.
4. **Wenxiang Jiao**, Xing Wang, Shilin He, Irwin King, Michael R. Lyu, Zhaopeng Tu. "Data Rejuvenation: Exploiting Inactive Training Examples for Neural Machine Translation". In Proceed- ings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), pp. 2255-2266, Online, USA, November 16 - November 20, 2020.
5. **Wenxiang Jiao**, Xing Wang, Zhaopeng Tu, Shuming Shi, Michael R. Lyu, Irwin King. "Self-training Sampling with Monolingual Data Uncertainty for Neural Machine Translation". In Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), To appear, Online, Thailand, August 1 - August 6, 2021.

Thank You!

Q&A: General

- Question: The first two studies are more related to model design, right?
- Answer:
  1. Indeed, the first two studies involve some design of the models. However, as mentioned, the data format plays an important role in determining the model architecture. We focus on the intra-sample structure information, which provides signals for developing more advanced models and learning more accurate representations. Therefore, it inevitably includes some improvement of previous models. The first study on HiGRU can be considered as the preparation for the second study on Pre-CODE.
  2. Nevertheless, there are still efforts required for data engineering in Pre-CODE, including the cleaning and filtering of data, and the construction of the questions and answers based on the unlabeled conversation data.

Q&A: General

- Question: Did you try to investigate inter-sample quality on ERC or intra-sample structure on MT?
- Answer:
  1. We tried to study inter-sample quality on ERC, e.g., applying self-training on ERC, but did not attain any improvement. Because the dataset is too small, and the teacher model is not strong enough to produce high-quality synthetic data.
  2. We did not investigate intra-sample structure on MT because the sentence structure does not contain information as rich as conversations and is also already well modeled by current self-attention networks.

Q&A: General

- Question: Why didn't you choose generation tasks that involve conversations for inter-sample quality? May be better to include intra-sample structure studies?
- Answer:
  1. The main reason is that tasks like conversation generation are not defined well. There is no consistent framework for training, or standard criteria for evaluation. Therefore, it will be hard to assess the advantages of our studies convincingly.
  2. In contrast, MT is a classic generation tasks, with standard frameworks (e.g., Transformer), datasets (e.g., WMT shared tasks), and evaluation metrics (e.g., BLEU score). Besides, MT involves multiple languages, which help us to gain more understanding on the interaction between languages.

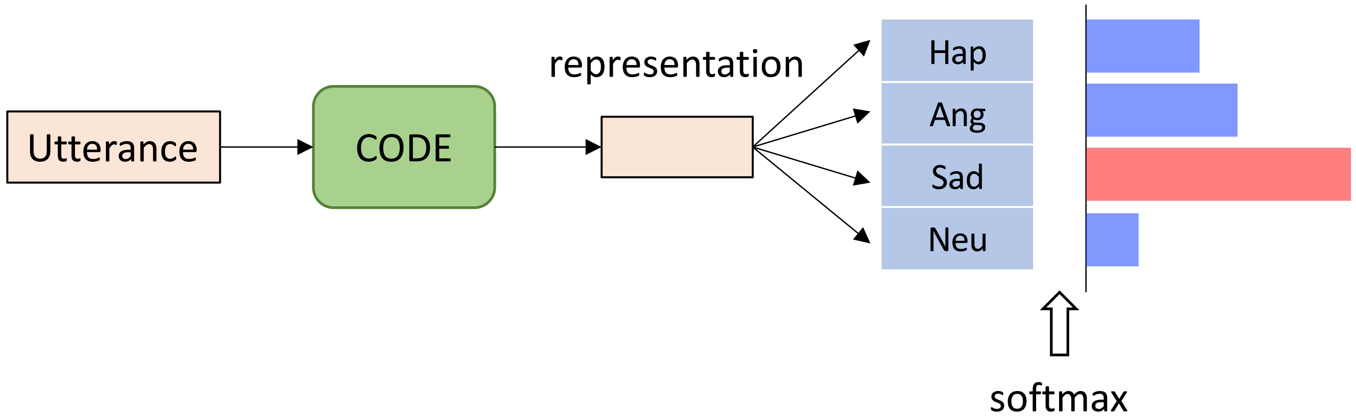
Q&A: General

- Question: Computation cost?
- Answer: It is mainly related to the model size and data size.
    1. The models for HiGRU and Pre-CODE are shallow networks with RNNs, while that for DataReju and UncSamp are Transformer-big models.
    2. Training sets for the first two contain less than 40K sentences while contain more than 10M sentence pairs for the latter two.

Methods	Params	Num of Sent	GPU x Num	Prep Time/h	Train Time/h	Total/h
HiGRU	<10M	< 11K	1080Ti x 1	N/A	0.5	0.5
Pre-CODE	<10M	< 40K + < 2M	1080Ti x 1	4.0	8.0 + 0.5	12.5
DataReju	213M	< 35M + < 3.5M	V100 x 8	1.5	16.0 + 16.0	34.0
UncSamp	213M	< 36.8M + < 40M	V100 x 8	12.0	16.0 + 16.0	44.0

Q&A: ERC

- Question: How to make the prediction of emotions?
- Answer:
  1. Maximum probability prediction. For a conversation, we feed it into the model and obtain the representation of each utterance. Calculate the relevance with all emotion embeddings, use softmax to normalize, and select the best.



Q&A: ERC

- Question: How to calculate accuracy, and F1-score?
- Answer: Take a binary classification as an example.
  1. Precision: The ratio of correctly predicted positive observations to the total predicted positive observations.
  2. Recall: The ratio of correctly predicted positive observations to the total observations in actual Class=Yes, i.e., accuracy.
  3. F1-score: The harmonic average of Precision and Recall.

		Predicted Class	
		Class=Yes	Class=No
Actual Class	Class=Yes	True Positive	False Negative
	Class=No	False Positive	True Negative

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F1} - \text{score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

Q&A: ERC

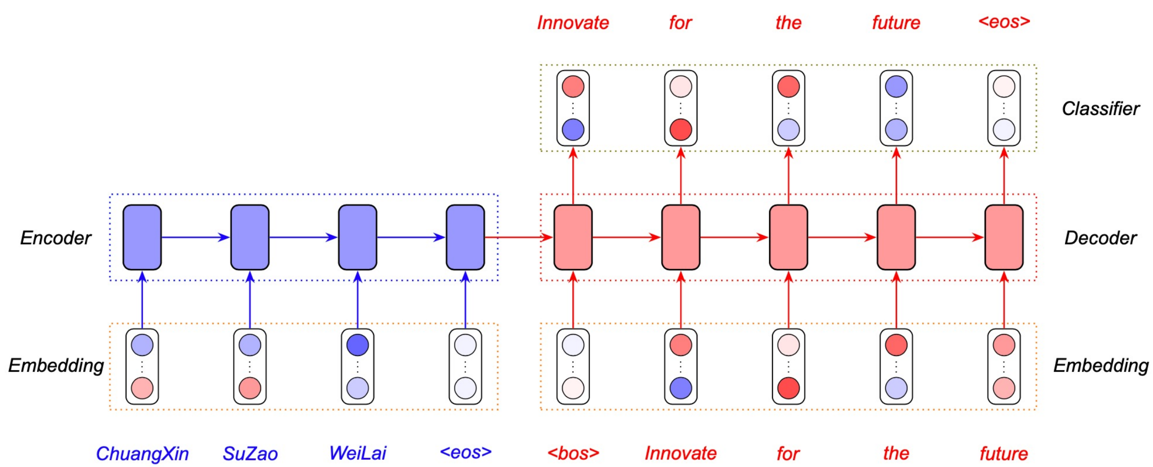
- Question: How to choose and use the noise utterances in Pre-CODE?
- Answer:
  1. For each conversation, we randomly sample 10 noise utterances from the other conversations.
  2. These 10 noise utterances are shared by all the masked ground-truth utterances in the conversations.
  3. We dynamically sample the 10 noise utterances in each epoch.

Q&A: ERC

- Question: How about using documents for pre-training?
- Answer:
  1. It is doable to pre-train on normal documents rather than conversations, since both text formats contain the hierarchical context. The learned representations could be a useful initialization for downstream ERC tasks.
  2. However, there might be domain or style mis-matching issues, as documents are usually more formal and less emotional. The domain gap may make this kind of pre-training less optimal.

Q&A: NMT

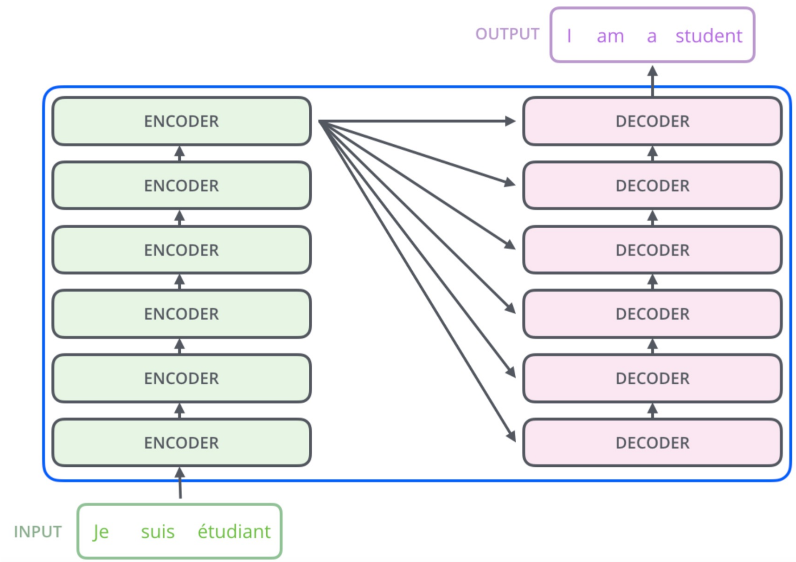
- Question: What is an encoder-decoder framework?
- Answer:
  1. The encoder-decoder framework represents a type of neural networks for learning the pattern of paired sentences. It contains an encoder and a decoder. The encoder encodes the source sentence, and the decoder predicts the output to match the target sentence. Conventionally, the encoder and decoder are RNNs. But now, it could be CNNs or self-attention networks.

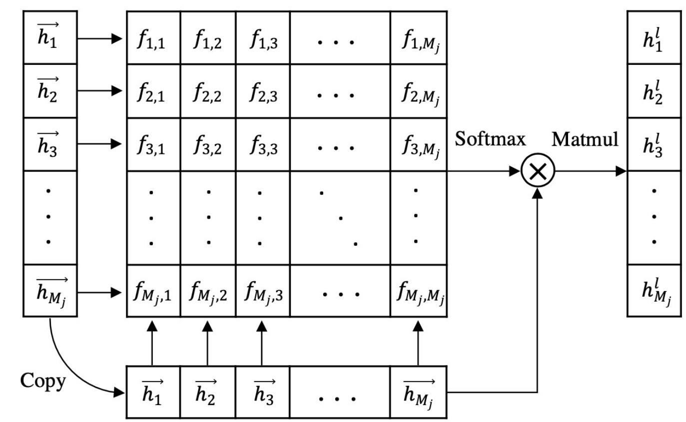


Tan et al. (AI Open 2020)

Q&A: NMT

- Question: What is the Transformer model?
- Answer:
  1. Transformer is a kind of encoder-decoder model that adopts the self-attention mechanism for modeling the text sequence.



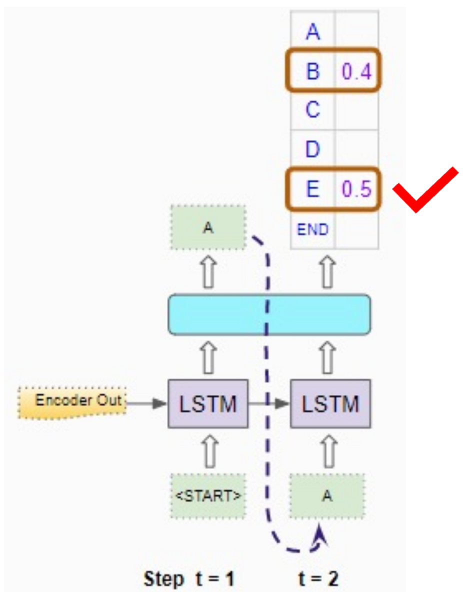


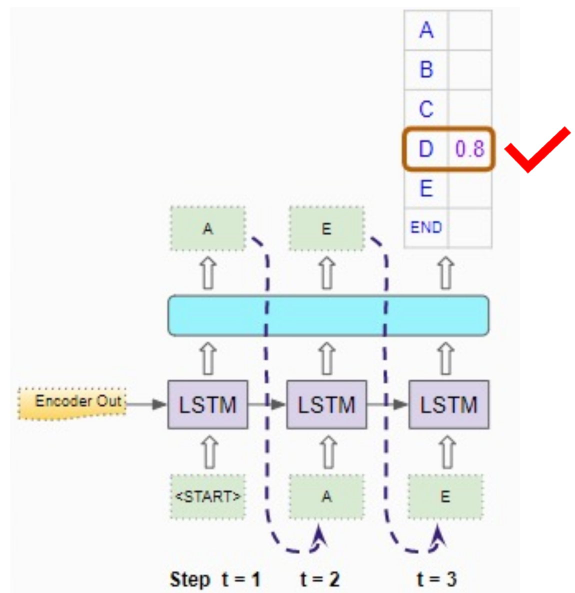
$$h_k^l = \sum_{p=1}^{M_j} a_{kp} \overrightarrow{h_p} \qquad a_{kp} = \frac{\exp(f(\overrightarrow{h_k}, \overrightarrow{h_p}))}{\sum_{p'=1}^{M_j} \exp\left(f(\overrightarrow{h_k}, \overrightarrow{h_{p'}})\right)}$$

[Jay Alammar's Blog]

Q&A: NMT

- Question: How to generate the translation during inference?
- Answer:
  1. Auto-regressive decoding, i.e., to predict the word at step t based on previously predicted words. Greedy search, i.e., to choose the word with the highest probability.

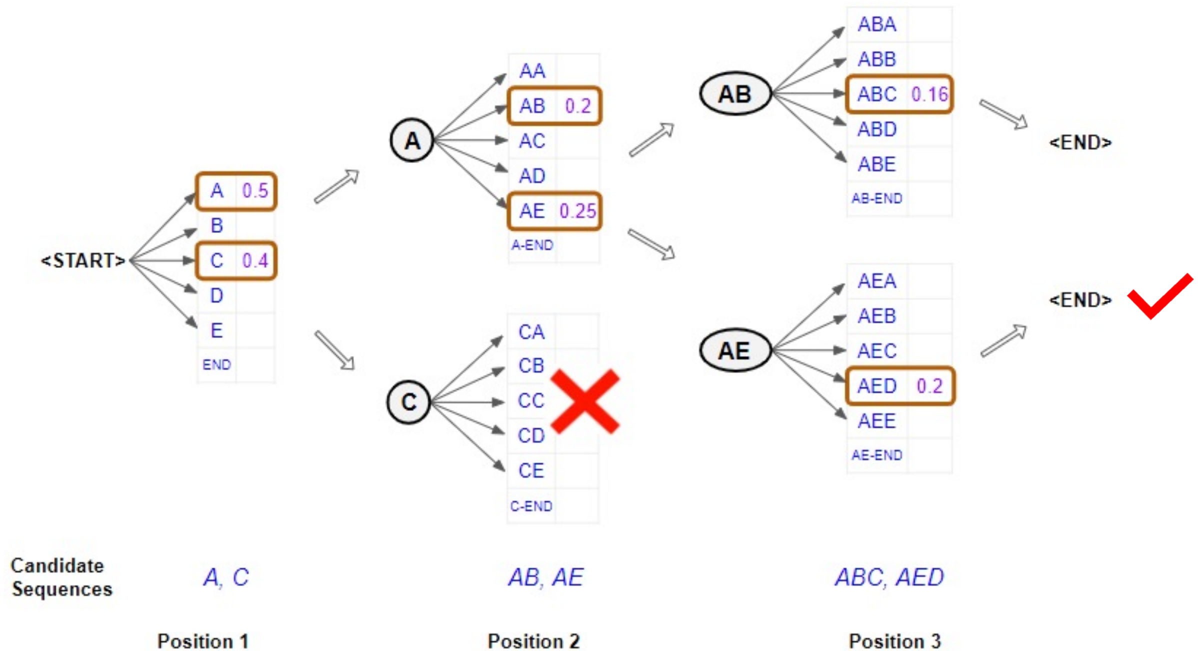




Ketan Doshi's Blog

Q&A: NMT

- Question: How to generate the translation sentence?
- Answer:
  1. Beam search algorithm, i.e., to maintain k partial translations with the highest joint probabilities at each step, where k is the beam width.



[Ketan Doshi's Blog]

Q&A: NMT

- Question: How to calculate the BLEU score?
- Answer:
  1. BLEU denotes "Bilingual Evaluation Understudy", which calculates the ratio of matched n-grams between a candidate translation and a (or multiple) reference translation. Usually, we use 4-gram BLEU, i.e., N = 4.

Source: 今天天气不错
Candidate: It is a nice day today
Reference: Today is a nice day

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$\mathrm{BP} = \begin{cases} 1, & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right), & \text{if } c \le r \end{cases}$$

Candidate: It is a nice day today

Reference: Today is a nice day

1-gram: p1 = 5 / 6

Candidate: It is a nice day today

Reference: Today is a nice day

3-gram: p3 = 2 / 4

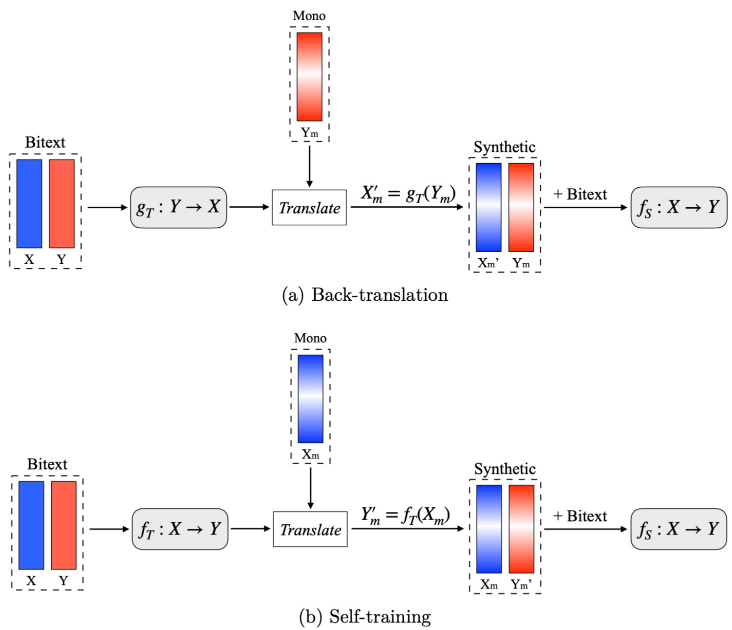
Q&A: NMT

- Question: What is the difference between the multi-BLEU and sacreBLEU?
- Answer:
  1. Multi-BLEU is the tokenized BLEU, which is usually calculated with self-developed tokenizers. For the same candidate translation, the different tokenizers will result in different BLEU scores, making it unfair to compare the results of different institutions.
  2. SacreBLEU is the detokenized BLEU, which aims to eliminate such an inconsistency of tokenizers. It collects all the publicly available test sets. Users can upload their detokenized candidate translations to the API, and SacreBLEU will use a unified tokenizer and report the BLEU score.

Detokenized: It’s a nice day today!
Tokenized-1: It ’ s a nice day today !
Tokenized-2: It ’s a nice day today !

Q&A: NMT

- Question: How to perform forward-translation (i.e., self-training) or back-translation?
- Answer:
  1. Both pair each monolingual sentence with a synthetic sentence by translating through a Teacher NMT model.

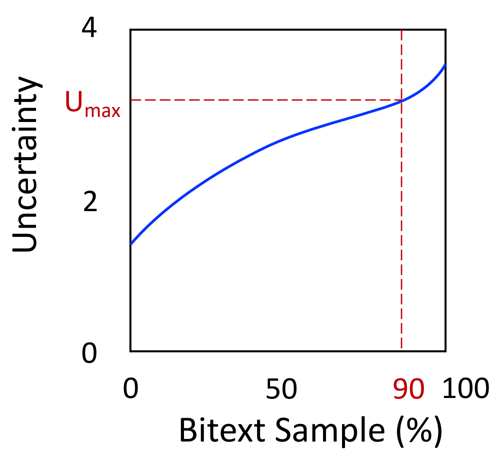


Q&A: NMT

- Question: We see that filtering is applied upon the proposed UncSamp. What and why?
- Answer:
  1. After our UncSamp is applied and we obtain the synthetic parallel data, we train a language model on the target sentences of the original parallel data. This target language model is used to remove synthetic parallel data that is distant to the domain of the original parallel data.
  2. It serves as an additional tool to reduce the effect of unqualified synthetic parallel data.

Q&A: NMT

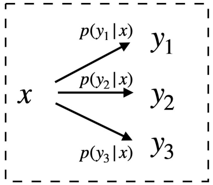
- Question: It is still unclear how to decide the Umax.
- Answer:
  1. Calculate the uncertainty of source sentences in the original parallel data (bitext) and sort them.
  2. Find the sentence ranked at the 90% position of all the sentences, and choose its uncertainty as Umax.
  3. We assume the NMT model cannot learn the sentences with uncertainty over Umax, therefore cannot produce high-quality synthetic data.



Q&A: NMT

- Question: How to calculate the linguistic properties?
- Answer:
  1. Frequency rank: The frequency rank of a word is its position in the dictionary where words are sorted in the descending order of their frequencies.
  2. Word rarity: Word rarity also measures the frequency of words in a sentence with a higher value indicating a more rare sentence.
  3. Coverage: Firstly, we train an alignment model on the training data by *fast-align*, and force-align the source and target sentences of each subset. Then we calculate the ratio of source words being aligned by any target words.
  4. Uncertainty: Translation entropy of a source sentence to the target language.

$$\mathrm{WR}(\mathbf{x}) = -\frac{1}{T_x}\sum_{t=1}^{T_x}\log p(x_t)$$



Bilingual Dictionary

$$\mathrm{U}(\mathbf{x}^j|\mathcal{A}_b) = \frac{1}{T_x}\sum_{t=1}^{T_x}\mathcal{H}(y|\mathcal{A}_b, x = x_t),$$

$$\mathcal{H}(y|\mathcal{A}_b, x_i) = -\sum_{y_j\in\mathcal{A}_b(x_i)} p(y_j|x_i)\log p(y_j|x_i).$$

Q&A: General

- Question: Future directions?
- Answer:
  1. Low-frequency words in self-training.
     - Low-frequency/high-uncertainty sentences cannot be well translated by the teacher model.
  2. Low-frequency issues in multilingual NMT models.
     - Low-frequency issues have been barely discussed in the multilingual settings.
  3. Self-supervised multilingual pre-training.
     - Multilingual pre-training on large-scale monolingual data is a promising direction, which will bring benefits to low-resource translation tasks and low-frequency words translation.