Experiences & Feelings for EMNLP2021

Mieradilijiang Maimaiti
(Hangzhou, 2011.11.24)

Outline

- Conference
- Key points
- Main Conference
- Poster
- Tutorials
- Workshop

It should be ……





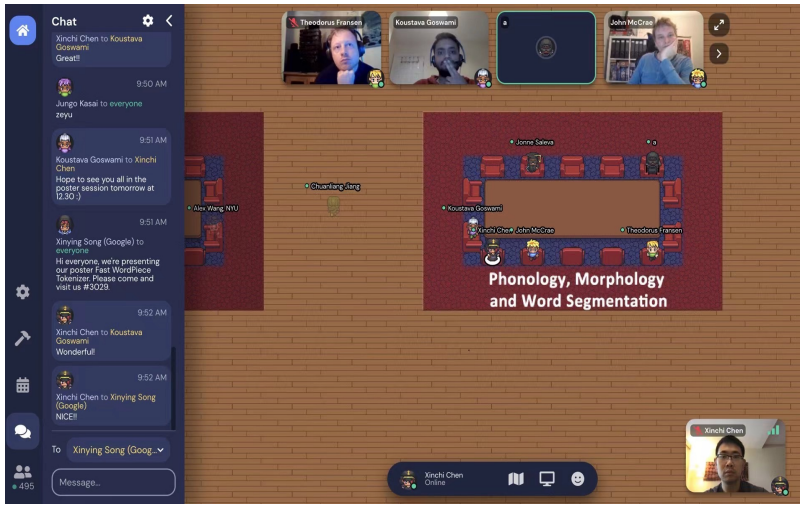

Photos resource: official websites and friends

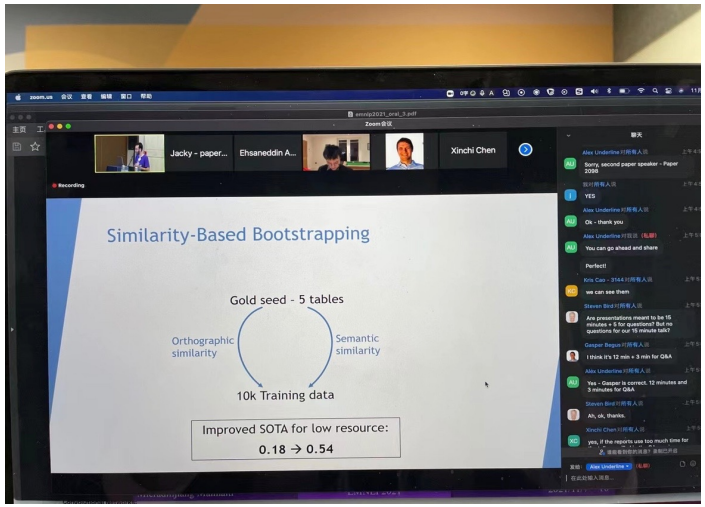
The reality is ……



Xuanjing Huang Lucia Specia Scott Wen-tau Yih

Opening Remarks from PC Chairs





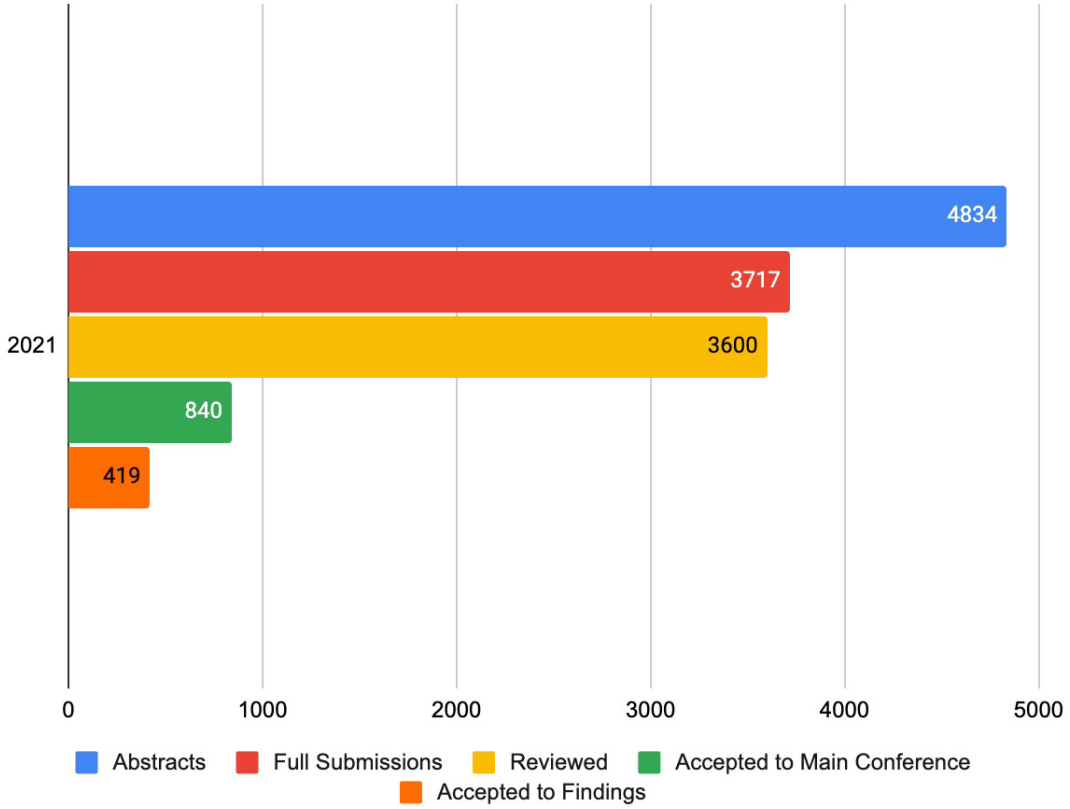
Photos resource: Internet, friends and me

Outline

- Conference
- Key points
- Main Conference
- Poster
- Tutorials
- Workshop

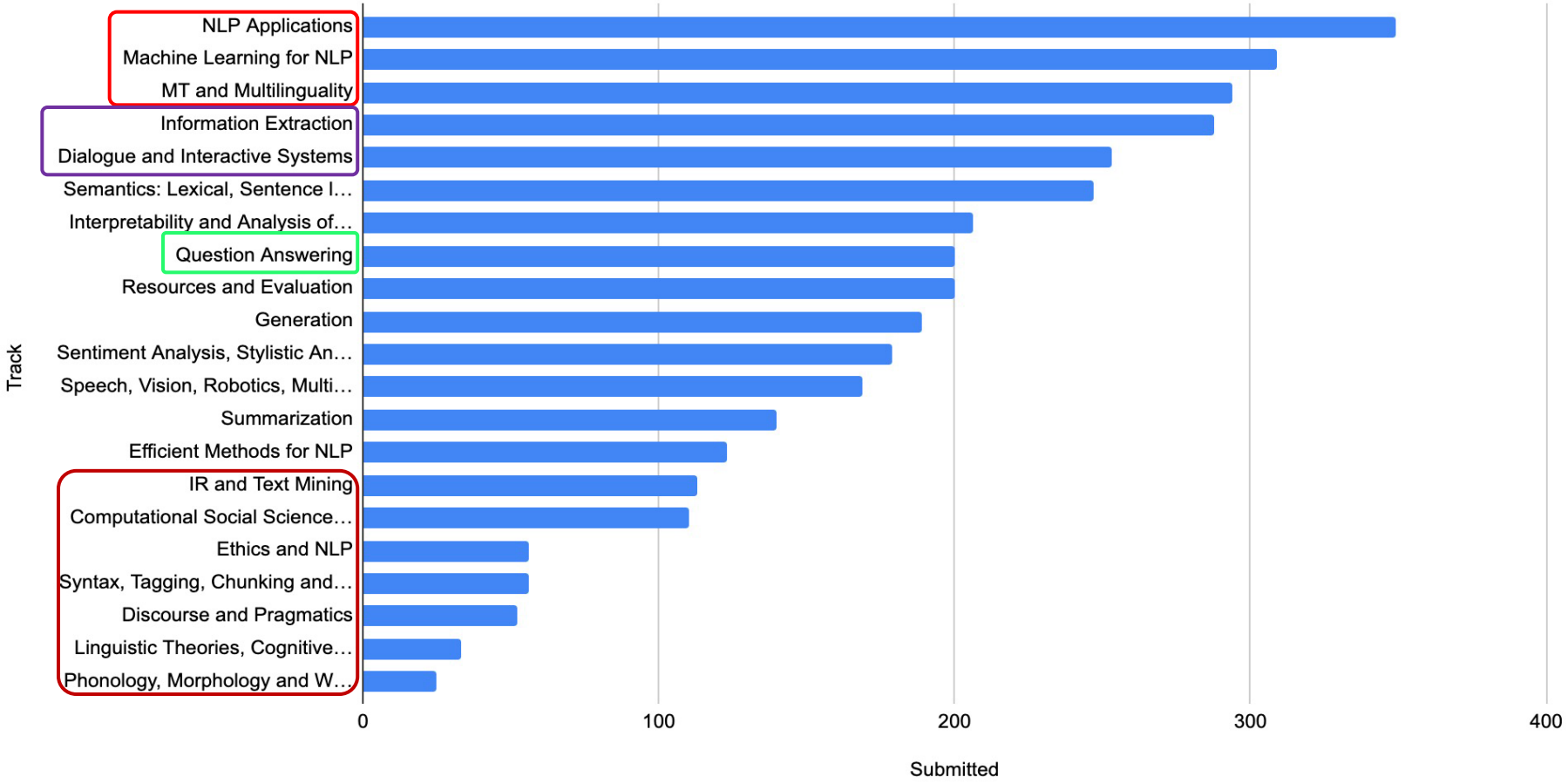
Key points

- 4834 abstracts
- 3717 full submissions
- 3600 papers reviewed
  - 2540 long
  - 1060 short
- 840 Accepted (23.3%)
  - 650 long
  - 190 short
- Plus Findings (11.6%)
  - 300 long
  - 119 short



Prof. Huang EMNLP2021 opening remark

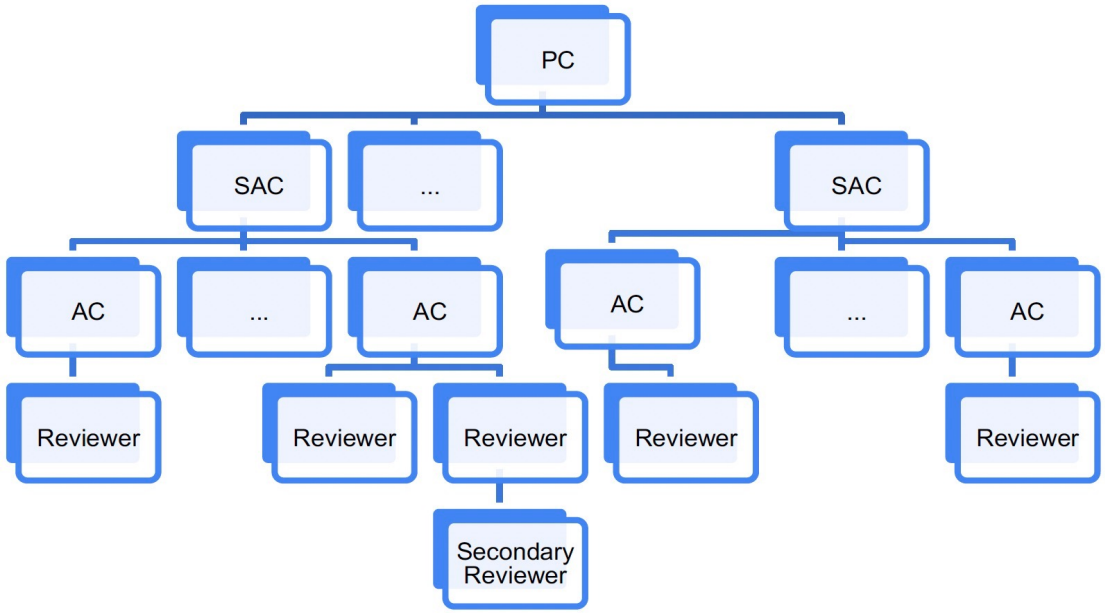
Submission by Track



Prof. Huang EMNLP2021 opening remark

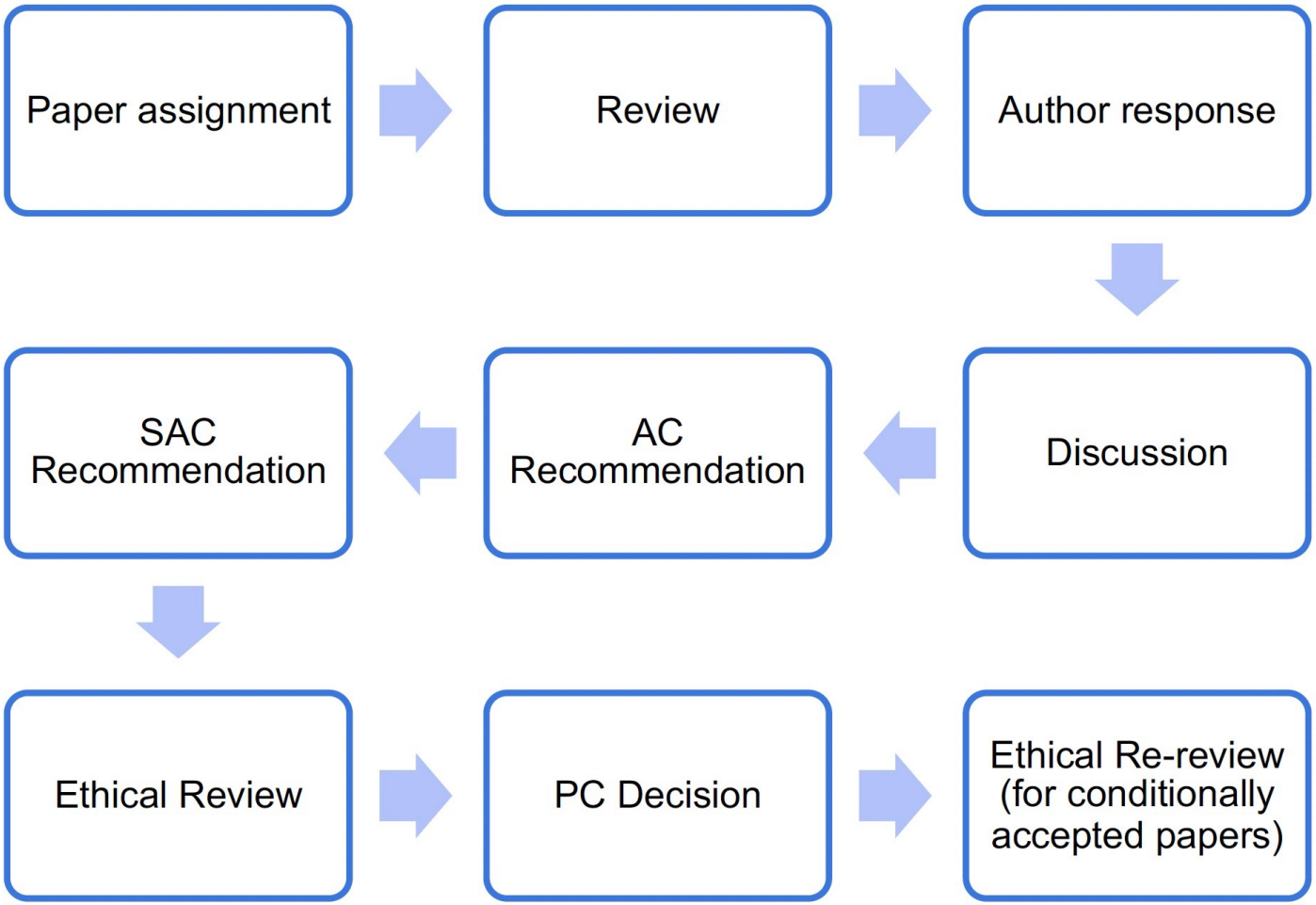
Program Committee Structure

- PC Co-Chairs: 3
- Number of SACs: 46
- Number of ACs: 236
- Number of (primary) reviewers: 3,112
- Number of Secondary reviewers: 370



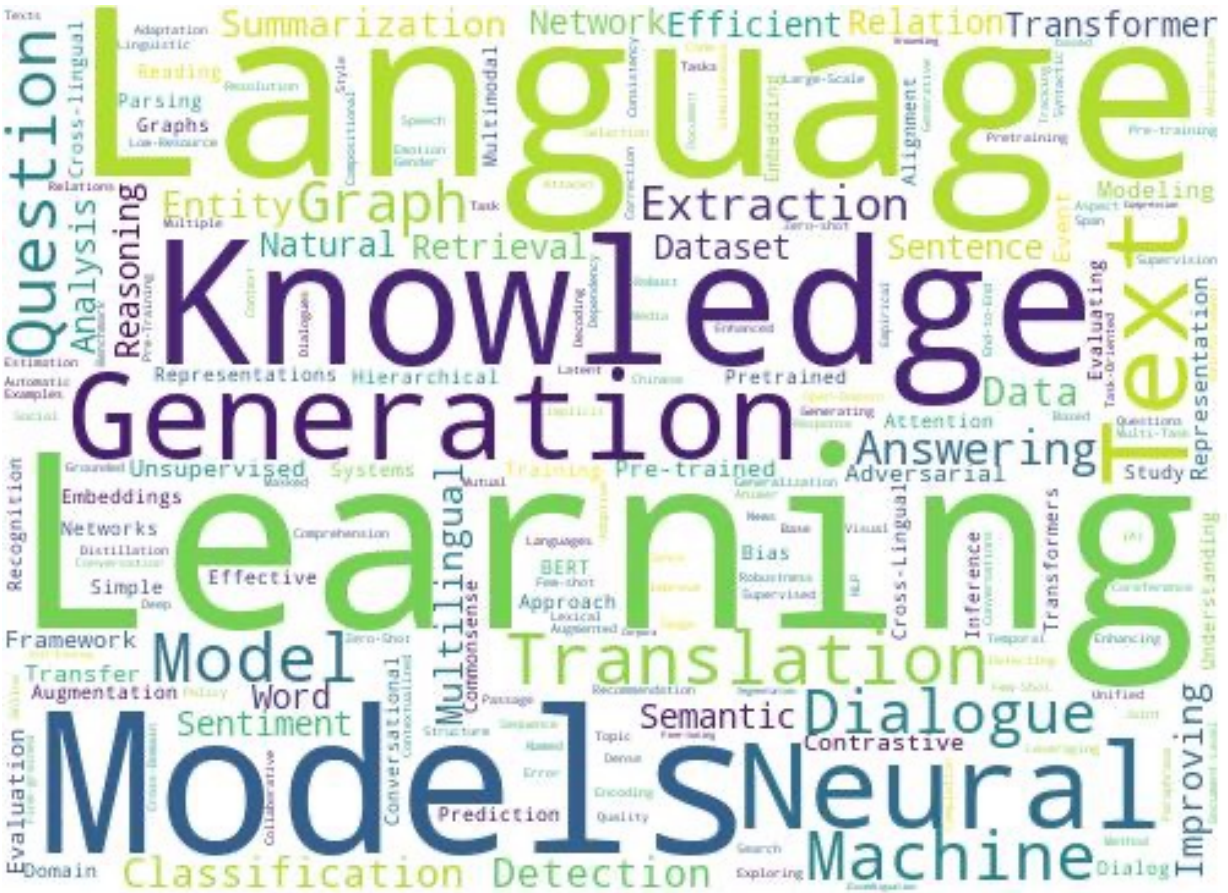
Prof. Huang EMNLP2021 opening remark

Technical Review Process



Prof. Huang EMNLP2021 opening remark

Hot Topics



Prof. Huang EMNLP2021 opening remark

Best Paper Awards

- Selection process for best long / short papers
  - SACs and ACs nominates papers
  - PC identified 21 candidates for the best papers award
  - Best Paper Award Committee assessed the candidates
- Best papers
  - Best Long Paper -- Visually Grounded Reasoning across Languages and Cultures
  - Best Short Paper -- CHoRaL: Collecting Humor Reaction Labels from Millions of Social Media
- What is the meaning of "Best"?
  - Impactful / Insightful / Field-changing/ Very solid

Prof. Huang EMNLP2021 opening remark

Outline

- Conference
- Key points
- Main Conference
- Poster
- Tutorials
- Workshop

Main Conference

Slides/figures are borrowed from the authors of respective presentation talk.

Boosting Cross-Lingual Transfer via Self-Learning with Uncertainty Estimation

Liyan Xu[†] **Xuchao Zhang**[‡] **Xujiang Zhao**[§]
Haifeng Chen[‡] **Feng Chen**[§] **Jinho D. Choi**[†]
[†]Emory University, {liyan.xu,jinho.choi}@emory.edu
[‡]NEC Laboratories America, {xuczhang,haifeng}@nec-labs.com
[§]University of Texas at Dallas, {xujiang.zhao,feng.chen}@utdallas.edu

Background

- **Cross-lingual transfer** (**CLT**): model for one language ⇒ model for other language(s)
- Zero-shot: training on **source** language + inference on **target** languages
- Major direction: embedding alignment

Background

- Embedding alignment:
  - Explicit word-embedding alignment: translation matrix
    - Supervised (Mikolov et al., 2013, etc.)
    - Unsupervised (Conneau et al., 2018, etc.)
  - Shared/joint embedding space: **multilingual pre-trained language models**
    - mBERT (Devlin et al., 2019)
    - XLM-R (Conneau et al., 2020)
    - mT5 (Xue et al., 2021)

Motivation

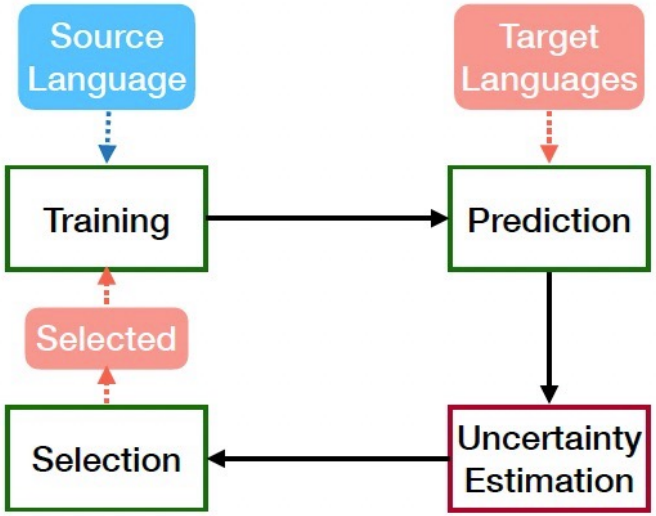
- Practical scenarios:
  - zero-shot?
  - Annotation for target languages?
  - Middle ground: **unlabeled data of target languages**
- Previous work: self-learning for multilingual document classification (Dong and Melo, 2019)
  - Predictions on unlabeled data of target languages
  - a.k.a "**pseudo labels**"

Approach

- **Self-learning** framework for cross-lingual transfer
  - w/ multilingual pre-trained LMs
  - Making use of zero-shot capability
- Explicit **uncertainty estimation**
  - uncertainty estimation ⇒ pseudo label quality ⇒ CLT performance

Approach

- Iterative training and prediction:
  - 1st iteration:
    - Train on gold labels of source language
  - 1+ iteration:
    - Select top-k confident predictions of target languages into training set
    - Need accurate uncertainty estimation
    - New training set: more data for task-specific learning and joint embedding alignment
    - Termination: no more unlabeled data or early stop on dev set



Uncertainties

- Deep learning models are notorious for over-confident predictions
  - High-dimensional space ⇒ sparse data points ⇒ imperfect decision boundary
- Two main types of uncertainties (Kendall and Gal, 2017; Depeweg et al., 2018)
  - ***Aleatoric*** uncertainty: intrinsic **data uncertainty** regardless of models
  - ***Epistemic*** uncertainty: **model uncertainty** that can be explained away with more data
- This work: focus on *aleatoric* uncertainty

Uncertainty Estimation

- Adapt three uncertainty estimation techniques:
  - Language Heteroscedastic Uncertainty (**LEU**)
  - Language Homoscedastic Uncertainty (**LOU**)
  - Evidential Uncertainty (**EVI**)

Results

NER

- Unlabeled data helps even without uncertainty estimation (BL-Single).
- Joint training on all target languages helps low-resource languages (BL-Joint).
- Uncertainty estimation outperforms (best results by **LEU**).

	en	af	ar	bg	bn	de	el	es	et	eu	fa	fi	fr	he	hi	hu	id	it	ja	jv	
BL-Direct	84.0	79.3	45.5	81.4	77.4	78.8	78.9	71.4	79.0	61.0	52.0	78.7	79.3	54.6	70.8	79.4	52.9	81.0	25.0	62.6	
BL-Single	84.0	78.9	56.9	84.5	79.3	80.9	81.6	72.9	80.7	63.2	54.8	80.5	81.9	63.0	73.9	81.7	54.3	82.1	36.5	60.9	
BL-Joint	84.7	79.5	56.7	84.9	80.5	80.5	81.5	73.3	81.2	64.0	55.1	81.2	82.1	62.6	76.6	81.6	54.5	83.0	37.2	63.5	
SL-EVI	**85.2**	83.7	**75.1**	85.8	82.0	83.6	84.4	**86.5**	84.6	72.1	72.9	84.7	**84.1**	61.4	**80.2**	85.7	54.8	83.9	41.3	69.2	
SL-LOU	84.4	**85.3**	61.1	87.1	81.9	83.4	85.4	75.6	85.5	74.6	74.9	84.4	83.3	**68.5**	78.6	84.5	**55.5**	**85.1**	46.2	**70.0**	
SL-LEU	84.7	81.5	70.0	**87.6**	**83.6**	**84.6**	**85.5**	85.0	**85.6**	**77.8**	**81.0**	**86.2**	83.1	62.0	79.5	**87.0**	53.4	84.8	**49.5**	65.3	
	ka	kk	ko	ml	mr	ms	my	nl	pt	ru	sw	ta	te	th	tl	tr	ur	vi	yo	zh	avg
BL-Direct	69.3	51.9	57.9	63.6	62.4	69.6	60.1	83.7	80.9	70.2	69.2	58.2	51.3	1.8	71.0	76.7	55.8	76.2	41.4	33.0	64.4
BL-Single	73.6	52.5	63.6	66.0	66.8	62.6	54.3	84.8	82.6	72.9	67.7	63.2	57.2	3.1	74.7	81.8	69.9	80.9	46.2	43.6	67.5
BL-Joint	73.6	53.4	63.6	67.5	67.9	64.3	53.0	84.8	83.2	73.5	69.7	63.1	57.4	3.6	76.1	81.8	71.5	81.4	**54.8**	43.7	68.3
SL-EVI	81.0	56.4	69.4	**76.3**	77.9	72.5	71.7	**87.1**	85.5	**80.6**	71.2	69.4	61.5	**6.7**	80.7	85.3	79.8	**86.2**	42.7	48.9	73.3
SL-LOU	78.8	58.7	70.2	75.4	**79.4**	73.8	71.2	86.4	86.2	79.2	**73.3**	**69.5**	68.8	4.7	**83.4**	88.4	**85.9**	85.8	49.1	50.5	73.8
SL-LEU	**81.1**	**63.7**	**71.8**	76.0	76.2	**75.9**	**71.5**	**87.1**	**87.6**	79.9	70.4	64.0	**69.9**	2.2	81.3	**89.1**	**85.9**	85.9	43.5	**54.8**	**74.4**

Results

XNLI

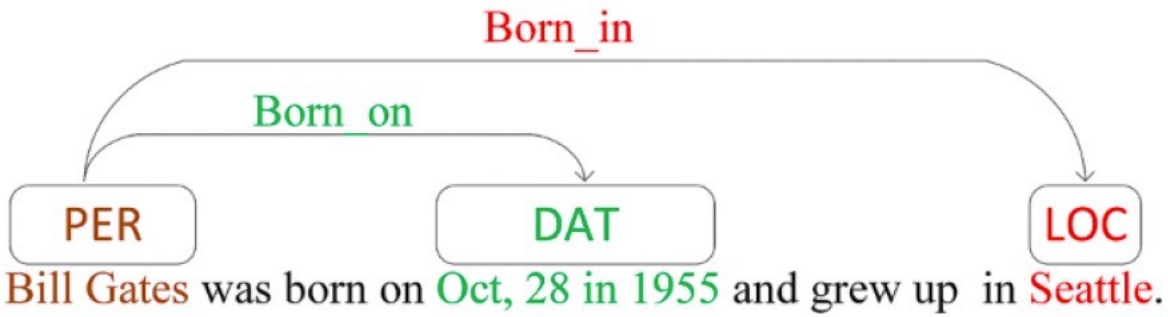
- Unlabeled data does not help without uncertainty estimation (BL-Single).
- Uncertainty estimation outperforms (best results by **LEU/LOU**).

	en	ar	bg	de	el	es	fr	hi	ru	sw	th	tr	ur	vi	zh	avg
BL-Direct	88.5	78.0	82.5	81.8	80.5	83.8	82.9	74.8	78.7	67.5	76.7	78.1	71.5	79.4	78.2	78.9
BL-Single	88.5	77.6	82.4	82.0	79.6	82.5	82.1	76.1	79.1	69.1	76.6	77.9	71.5	77.9	78.2	78.7
BL-Joint	88.2	78.8	82.0	82.2	80.4	83.1	82.2	76.1	79.6	68.8	76.2	78.0	71.4	79.1	78.5	79.0
SL-EVI	88.1	79.5	84.4	83.4	82.4	**84.8**	83.7	78.0	81.6	71.1	78.2	79.2	74.4	80.8	80.4	80.7
SL-LOU	88.2	**81.0**	84.4	**83.5**	82.3	**84.8**	**83.9**	78.9	**81.8**	**73.9**	79.3	80.1	**75.7**	81.6	**81.4**	**81.4**
SL-LEU	88.1	80.7	**84.9**	83.4	**82.8**	84.5	83.8	**79.2**	**81.8**	73.0	**79.7**	**80.5**	**75.7**	**81.9**	81.3	**81.4**

A Partition Filter Network for Joint Entity and Relation Extraction

Zhiheng Yan, Chong Zhang, Jinlan Fu, Qi Zhang and Zhongyu Wei

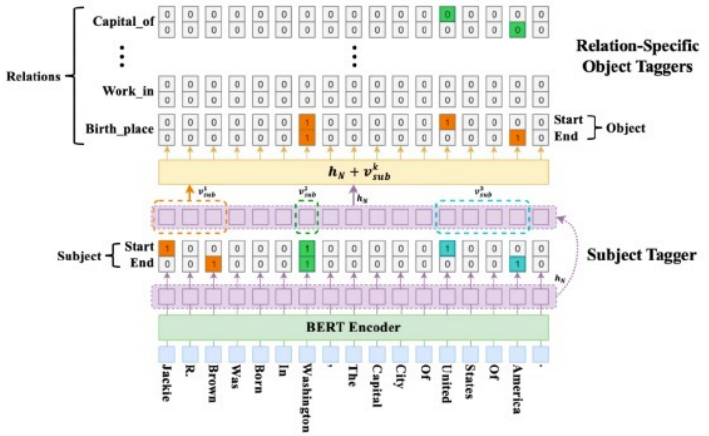
Task Definition



NER Identify boundary and type of entity
RE Identify relation between entity pair
Joint Extract Relational triples
(Bill Gates (PER), Born_in, Seattle(LOC))

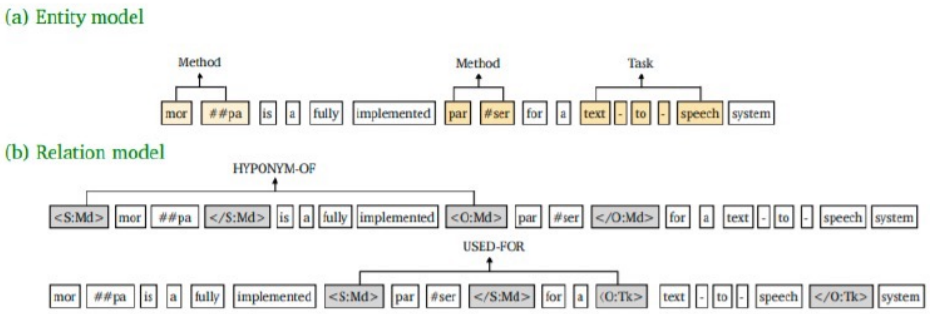
Limitations of Current Encoding Schemes

Sequential Encoding: Encoding NER and RE with a pre-defined order -> imbalanced interaction



A Novel Cascade Binary Tagging Framework for Relational Triple Extraction ACL 2020

Parallel Encoding: Encoding NER and RE with independent encoders -> insufficient interaction



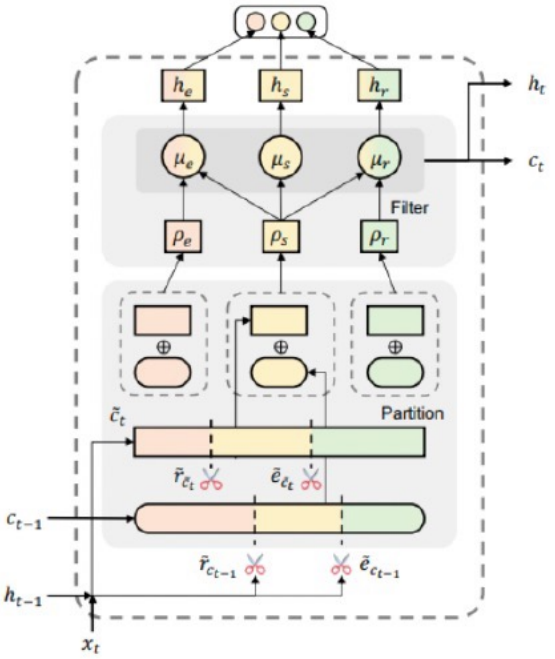
A Frustratingly Easy Approach for Entity and Relation Extraction NAACL 2021

Our Solution – Joint Encoding

Joint Encoding: Encoding task-specific feature with a joint encoder – Partition Filter Encoder

Partition: Segment a set of cell neurons into two intra-task partition and one inter-task partition with two scissor gates

Filter: Generate task-specific feature with shared partition and corresponding task partition



(b) Inner Mechanism of Partition Filter

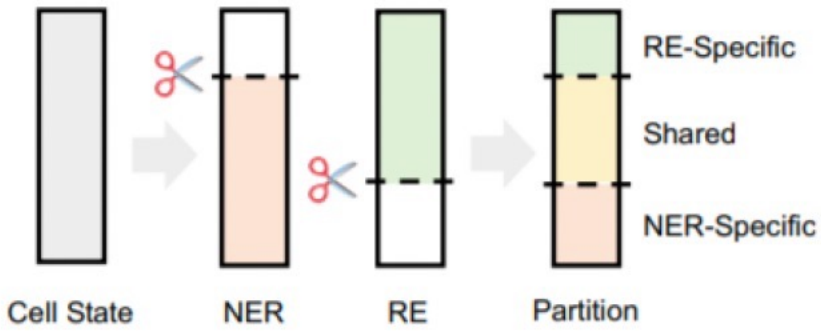
Enforcing bilateral interaction between NER and RE:

- **Partitions are generated with the joint effort of entity and relation gates**
- **Information in shared partition in evenly accessible to both tasks**

Encoder - Partition

1. Identify cut-off points (scissor gates)

2. Generate three partitions with the gates



Encoder - Filter

Partition Result
Entity Partition - Information reserved for NER only
Relation Partition – Information reserved for RE only
Shared Partition – Information valuable to both tasks

Feature Extraction
Entity Feature = Entity Partition + Shared Partition
Relation Feature = Relation Partition + Shared Partition

Decoder – Table Filling

NER Table
Word Pair prediction : Entity Span
Table Type : Entity Type

RE Table
Word Pair prediction : Subj-Obj
Triple (head token)
Table Type : Relation Type

Triple Extraction: Subj-Obj
Triple (head + tail token)

Method	NER	RE
NYT △		
CopyRE (Zeng et al., 2018)	86.2	58.7
GraphRel (Fu et al., 2019)	89.2	61.9
CopyRL (Zeng et al., 2019)	-	72.1
Casrel (Wei et al., 2020) †	(93.5)	89.6
TpLinker (Wang et al., 2020b) †	-	91.9
PFN†	**95.8**	**92.4**
WebNLG △		
CopyRE (Zeng et al., 2018)	82.1	37.1
GraphRel (Fu et al., 2019)	91.9	42.9
CopyRL (Zeng et al., 2019)	-	61.6
Casrel (Wei et al., 2020) †	(95.5)	91.8
TpLinker (Wang et al., 2020b) †	-	91.9
PFN†	**98.0**	**93.6**
ADE ▲		
Multi-head (Bekoulis et al., 2018b)	86.4	74.6
Multi-head + AT (Bekoulis et al., 2018a)	86.7	75.5
Rel-Metric (Tran and Kavuluru, 2019)	87.1	77.3
SpERT (Eberts and Ulges, 2019) †	89.3	79.2
Table-Sequence (Wang and Lu, 2020) ‡	89.7	80.1
PFN†	89.6	80.0
PFN‡	**91.3**	**83.2**
ACE05 △		
Structured Perceptron (Li and Ji, 2014)	80.8	49.5
SPTree (Miwa and Bansal, 2016)	83.4	55.6
Multi-turn QA (Li et al., 2019) †	84.8	60.2
Table-Sequence (Wang and Lu, 2020) ‡	89.5	64.3
PURE (Zhong and Chen, 2021) ‡	**89.7**	65.6
PFN‡	89.0	**66.8**
ACE04 △		
Structured Perceptron (Li and Ji, 2014)	79.7	45.3
SPTree (Miwa and Bansal, 2016)	81.8	48.4
Multi-turn QA (Li et al., 2019) †	83.6	49.4
Table-Sequence (Wang and Lu, 2020) ‡	88.6	59.6
PURE (Zhong and Chen, 2021) ‡	88.8	60.2
PFN‡	**89.3**	**62.5**
SciERC △		
SPE (Wang et al., 2020a) §	**68.0**	34.6
PURE (Zhong and Chen, 2021) §	66.6	35.6
PFN§	66.8	**38.4**

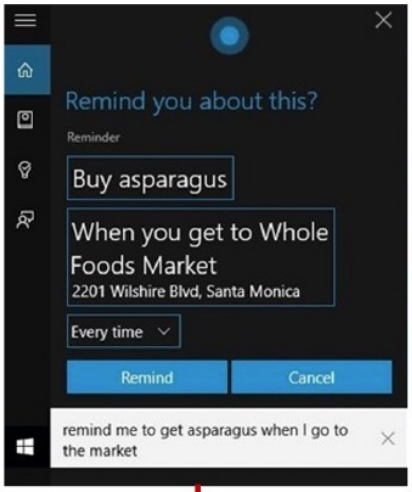
Table 1: Experiment results on six datasets. †, ‡ and § denotes the use of BERT, ALBERT and SCIBERT(Devlin et al., 2019; Lan et al., 2020; Beltagy et al., 2019) pre-trained embedding. △ and ▲ denotes the use of micro-F1 and macro-F1 score. NER results of Casrel are its reported average score of head and tail entity. Results of PURE are reported in single-sentence setting for fair comparison.

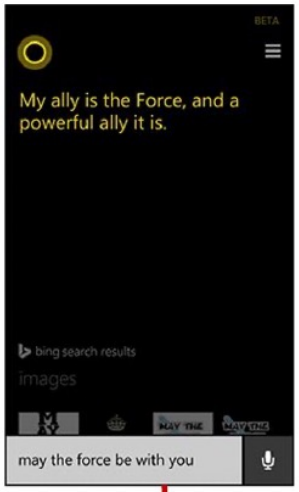
Few-Shot Intent Detection via Contrastive Pre-Training and Fine-Tuning

Jian-Guo Zhang[1]*, Trung Bui[2], Seunghyun Yoon[2], Xiang Chen[2], Zhiwei Liu[1]
Congying Xia[1], Quan Hung Tran[2], Walter Chang[2], Philip Yu[1]
[1] University of Illinois at Chicago, Chicago, USA
[2]Adobe Research, San Jose, USA

Intelligent Assistants

Google Now Siri Cortana echo





Task-Oriented

Chit-Chat (social bots)

Image Credit: Jason Wu

Intent Detection

Is a key component in task-oriented dialog systems

- Given a collection of utterances with labels, train a model to estimate labels for new utterances.

Find_movie
Buy_tickets
Find_restaurant
Book_table
Find_lyrics
...

Challenges for Intent Detection

- **Data Scarcity:** It is expensive to have many training examples

- **Balanced K-shot learning:** Assume that we have K examples for each of the N classes in our training data

Challenges for Intent Detection

- It could be more challenging when there exist many fine-grained intents

How can I go about ordering a virtual card?	getting_virtual_card
How do I get a disposable virtual card as well?	get_disposable_virtual_card
Where is this card accepted?	card_acceptance
My new card hasn't come in.	card_arrival

Challenges for Intent Detection

- It could be more challenging when there exist many fine-grained intents
- User utterances may be semantically similar across different intents

Will my card be here soon?	card_delivery_estimate
When will I get my card shipped to me?	card_delivery_estimate
Can I track when my card will be delivered?	card_arrival
My new card hasn't come in.	card_arrival

Few-shot Intent Detection

- Data scarcity: There are only a few examples for each intent class
- Fine-grained scenario: There are semantically similar intents
- How to correctly identify (fine-grained) user intents with limited training examples?

A few-shot Intent Detection Schema

- Stage-1: Self-supervised contrastive pre-training on collected intent datasets, which implicitly learns to discriminative semantically similar utterances without using any labels.

- Stage-2: Few-shot supervised contrastive learning to help the model explicitly learn to pull utterances from the same intent close and push utterances across different intents apart.

Stage-1: Self-Supervised Pre-Training

- Train a model on collected intent datasets without using any labels
- **Self-supervised contrastive learning:**
  - **Positive pair:** Sentence_A, sentence_A with a few randomly and dynamically masked tokens.
  - **Negative pair:** All other pairs that are non-diagonal in the batch.
- **Mask language modeling** loss to enhance the token-level utterance understanding:

$$\mathcal{L}_{\text{mlm}} = -\frac{1}{M}\sum_{m=1}^{M} \log P(x_m),$$

Stage-2: Supervised Fine-Tuning

Given very limited training examples for each intent

- **Supervised contrastive learning:**
  - **Positive pair:** Two utterances from the same intent class; the utterance and itself.
  - **Negative pair:** Two utterances from different classes, i.e., all other pairs in the batch.
- **Intent classification** loss learns to classify utterances on a few training examples:

$$\mathcal{L}_{\text{intent}} = -\frac{1}{N}\sum_{j=1}^{C}\sum_{i=1}^{N}\log P(C_j \mid u_i),$$

Overall Performance

- **CPFT vs. CONVBERT+MLM:** +3.21% (BANKING77, 10-shot), +2.61% (HWU64, 10-shot)
  - It is much better than task-adaptive learning.

Model	CLINC150		BANKING77		HWU64	
	5-shot	10-shot	5-shot	10-shot	5-shot	10-shot
RoBERTa+Classifier (Zhang et al., 2020a)	87.99	91.55	74.04	84.27	75.56	82.90
USE (Casanueva et al., 2020)	87.82	90.85	76.29	84.23	77.79	83.75
CONVERT (Casanueva et al., 2020)	89.22	92.62	75.32	83.32	76.95	82.65
USE+CONVERT (Casanueva et al., 2020)	90.49	93.26	77.75	85.19	80.01	85.83
CONVBERT (Mehri et al., 2020a)	-	92.10	-	83.63	-	83.77
CONVBERT + MLM (Mehri et al., 2020a)	-	92.75	-	83.99	-	84.52
CONVBERT + Combined (Mehri et al., 2020b)	-	93.97	-	85.95	-	86.28
DNNC (Zhang et al., 2020a)	91.02	93.76	80.40	86.71	80.46	84.72
CPFT	**92.34**	**94.18**	**80.86**	**87.20**	**82.03**	**87.13**

Table 2: Testing accuracy (×100%) on three datasets under 5-shot and 10-shot settings.

Is contrastive pre-training beneficial to the target intent dataset?

- Stage-1: Contrastive pre-training on the datasets except for HWU64
- Stage-2: few-shot learning on HWU64
- Stage-1 + Stage-2 vs. Stage-2 only: +1.98% (5-shot), +1.21 (10-shot)
  - Answer: Yes
- The performance drops when compared to Stage-1 with contrastive pre-training on all datasets including HWU64

Outline

- Conference
- Key points
- Main Conference
- Poster
- Tutorials
- Workshop

Poster

Slides/figures are borrowed from the authors of respective presentation talk.

達摩院 ALIBABA DAMO ACADEMY

Alibaba Group 阿里巴巴集团

Improving Question Answering Model Robustness with Synthetic Adversarial Data Generation

†UCL ‡Facebook AI Research *USC


Max Bartolo †


Tristan Thrush ‡


Robin Jia ‡*


Sebastian Riedel †‡


Pontus Stenetorp †


Douwe Kiela ‡

Max Bartolo

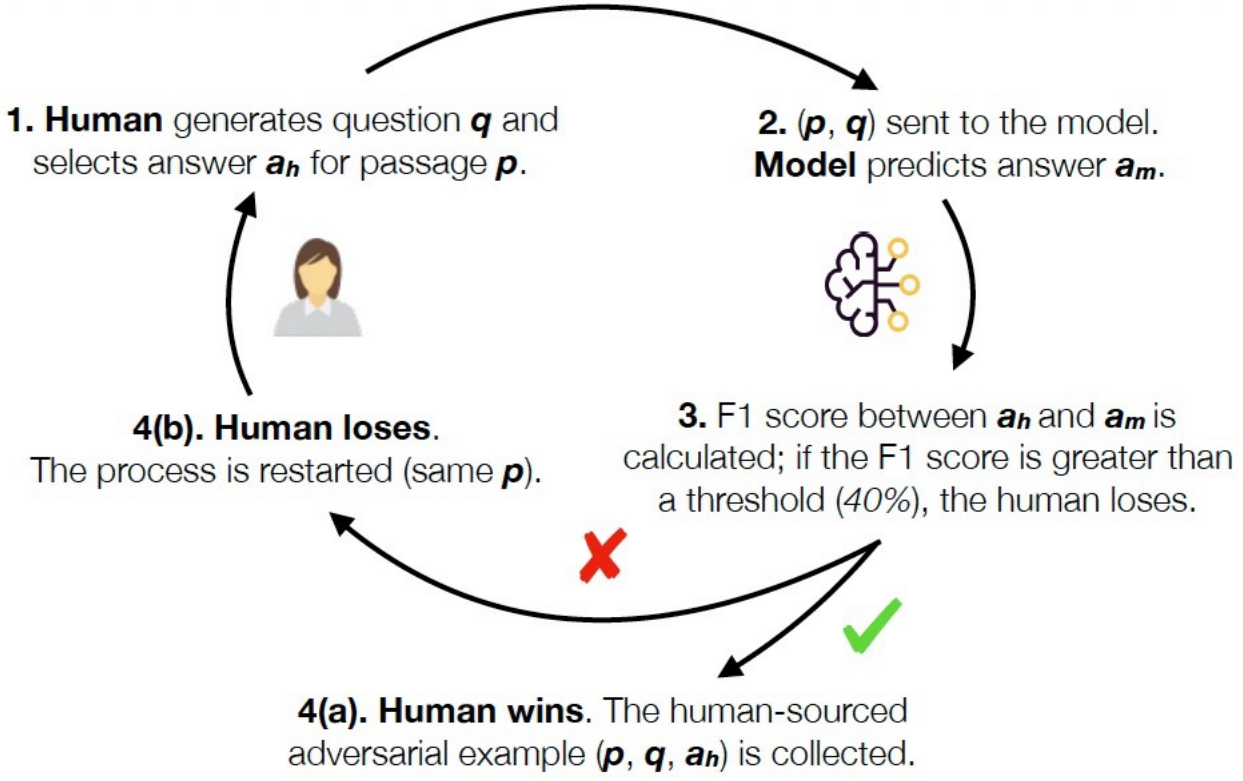
maxbartolo.com

max_nlp

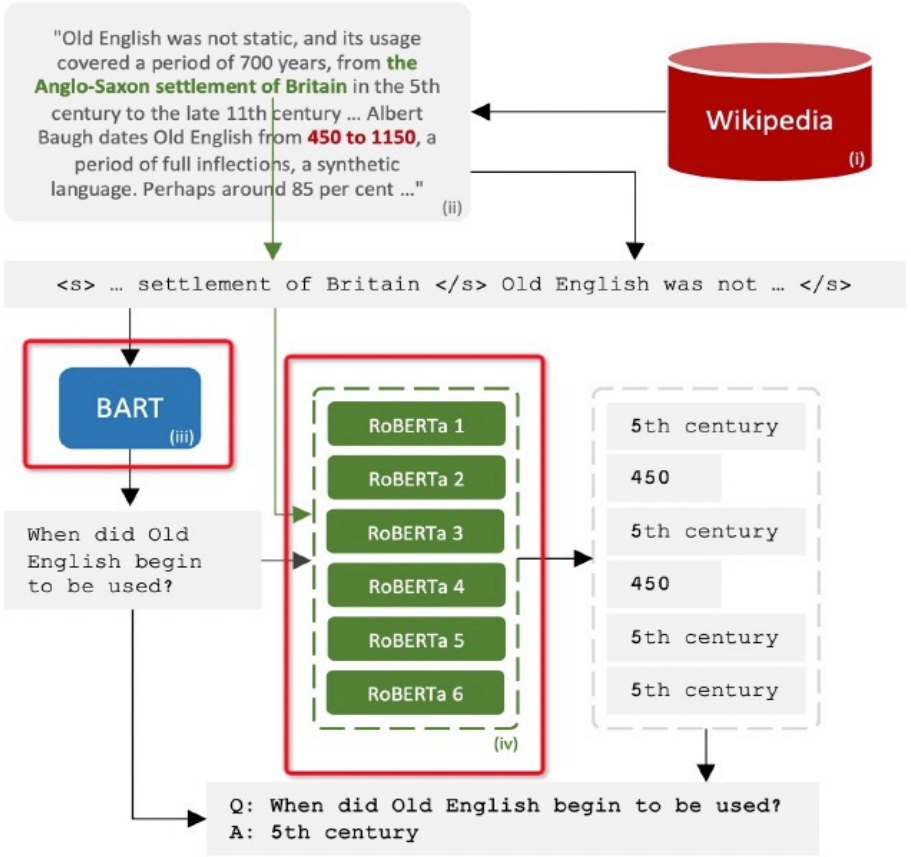
FACEBOOK AI

NLP UCL

Beat the AI (Bartolo et al., 2020)



Synthetic Adversarial Data Pipeline



Better Domain Generalisation

MRQA in-domain

Model	SQuAD		NewsQA		TriviaQA		SearchQA		HotpotQA		NQ		Avg	
	EM	F_1	*EM*	F_1	*EM*	F_1	*EM*	F_1	*EM*	F_1	*EM*	F_1	*EM*	F_1
R_{SQuAD}	$84.1_{1.3}$	$90.4_{1.3}$	$41.0_{1.2}$	$57.5_{1.6}$	$60.2_{0.7}$	$69.0_{0.8}$	$16.0_{1.8}$	$20.8_{2.7}$	$53.6_{0.8}$	$68.9_{0.8}$	$40.5_{2.7}$	$58.5_{2.0}$	49.2	60.9
$R_{SQuAD+AQA}$	$84.4_{1.0}$	$90.2_{1.1}$	$41.7_{1.6}$	$58.0_{1.7}$	$\mathbf{62.7}_{0.4}$	$\mathbf{70.8}_{0.3}$	$20.6_{2.9}$	$25.5_{3.6}$	$56.3_{1.1}$	$72.0_{1.0}$	$54.4_{0.5}$	$68.7_{0.4}$	53.3	64.2
SynQA	$88.8_{0.3}$	$\mathbf{94.3}_{0.2}$	$42.9_{1.6}$	$60.0_{1.4}$	$62.3_{1.1}$	$70.2_{1.1}$	$23.7_{3.7}$	$29.5_{4.4}$	$\mathbf{59.8}_{1.1}$	$75.3_{1.0}$	$55.1_{1.0}$	$68.7_{0.8}$	55.4	66.3
$SynQA_{Ext}$	$\mathbf{89.0}_{0.3}$	$\mathbf{94.3}_{0.2}$	$\mathbf{46.2}_{0.9}$	$\mathbf{63.1}_{0.8}$	$58.1_{1.8}$	$65.5_{1.9}$	$\mathbf{28.7}_{3.2}$	$\mathbf{34.3}_{4.1}$	$59.6_{0.6}$	$\mathbf{75.5}_{0.4}$	$\mathbf{55.3}_{1.1}$	$\mathbf{68.8}_{0.9}$	**56.2**	**66.9**

MRQA out-of-domain

Model	BioASQ		DROP		DuoRC		RACE		RelationExt.		TextbookQA		Avg	
	EM	F_1	*EM*	F_1	*EM*	F_1	*EM*	F_1	*EM*	F_1	*EM*	F_1	*EM*	F_1
R_{SQuAD}	$53.2_{1.1}$	$68.6_{1.4}$	$39.8_{2.6}$	$52.7_{2.2}$	$49.3_{0.7}$	$60.3_{0.8}$	$35.1_{1.0}$	$47.8_{1.2}$	$74.1_{3.0}$	$84.4_{2.9}$	$35.0_{3.8}$	$44.2_{3.7}$	47.7	59.7
$R_{SQuAD+AQA}$	$54.6_{1.2}$	$\mathbf{69.4}_{0.8}$	$59.8_{1.3}$	$68.4_{1.5}$	$\mathbf{51.8}_{1.1}$	$\mathbf{62.2}_{1.0}$	$38.4_{0.9}$	$51.6_{0.9}$	$75.4_{2.3}$	$85.8_{2.4}$	$40.1_{3.1}$	$48.2_{3.6}$	53.3	64.3
SynQA	$\mathbf{55.1}_{1.5}$	$68.7_{1.2}$	$64.3_{1.5}$	$72.5_{1.7}$	$51.7_{1.3}$	$62.1_{0.9}$	$\mathbf{40.2}_{1.2}$	$\mathbf{54.2}_{1.3}$	$78.1_{0.2}$	$87.8_{0.2}$	$40.2_{1.3}$	$49.2_{1.5}$	**54.9**	**65.8**
$SynQA_{Ext}$	$54.9_{1.3}$	$68.5_{0.9}$	$\mathbf{64.9}_{1.1}$	$\mathbf{73.0}_{0.9}$	$48.8_{1.2}$	$58.0_{1.2}$	$38.6_{0.4}$	$52.2_{0.6}$	$\mathbf{78.9}_{0.4}$	$\mathbf{88.6}_{0.2}$	$\mathbf{41.4}_{1.1}$	$\mathbf{50.2}_{1.0}$	54.6	65.1

Table 8: Domain generalisation results on the in-domain (top) and out-of-domain (bottom) subsets of MRQA.

MetaTS: Meta Teacher-Student Network for Multilingual Sequence Labeling with Minimal Supervision

Zheng Li[1], Danqing Zhang [1], Tianyu Cao[1], Ying Wei[2], Yiwei Song[1], Bing Ying[1]

[1]Amazon.com Inc

[2]City University of Hong Kong

Multilingual Sequence Labeling

Sequence labeling:
- token-level classification (NER, ABSA, SRL,…)
- expensive and time-consuming human labeling

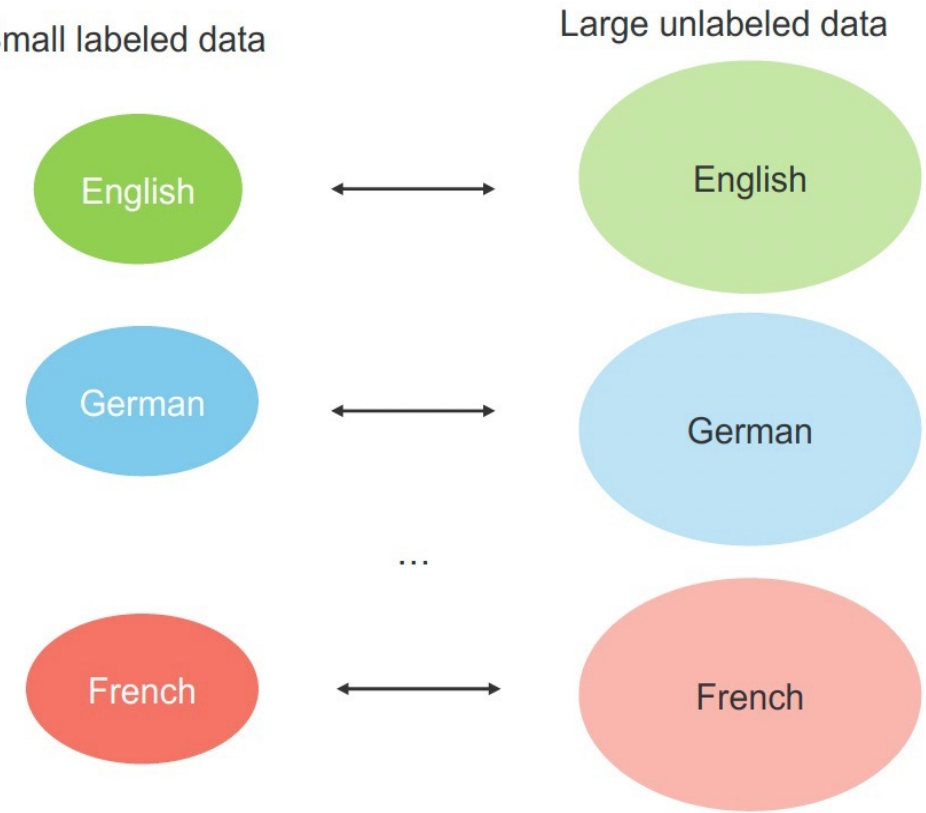
NER

Label: Brand ProductLine Size ProductType

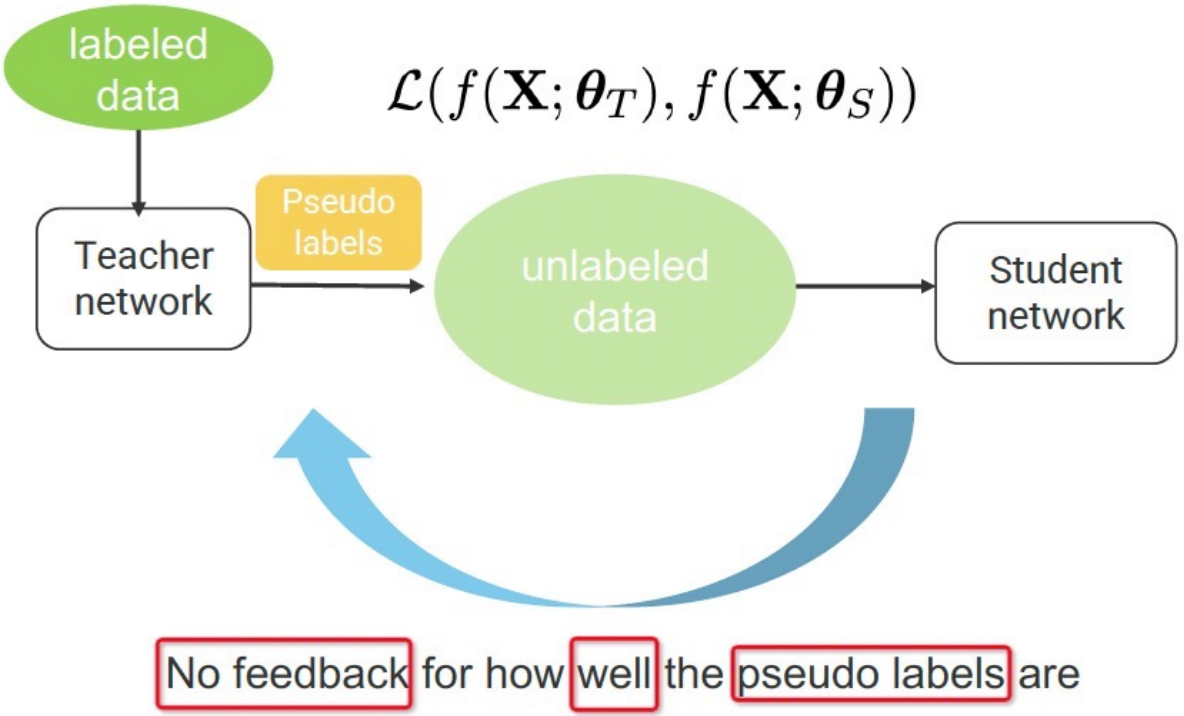
Query: mackie profx6v3 6-channel mixer

Multilingual
- Exacerbated data scarcity when we meet multiple languages

Semi-supervised Learning (SSL)



Self-training via Teacher-Student Network



Student Network

- Hard pseudo labels from the teacher

$$\widetilde{y}_{m,n}^{l_i,(t)} = \arg\max_c f_{n,c}(\widetilde{\mathbf{X}}_m^{l_i}; \boldsymbol{\theta}_T^{(t)})$$

- Student training on pseudo-labeled data

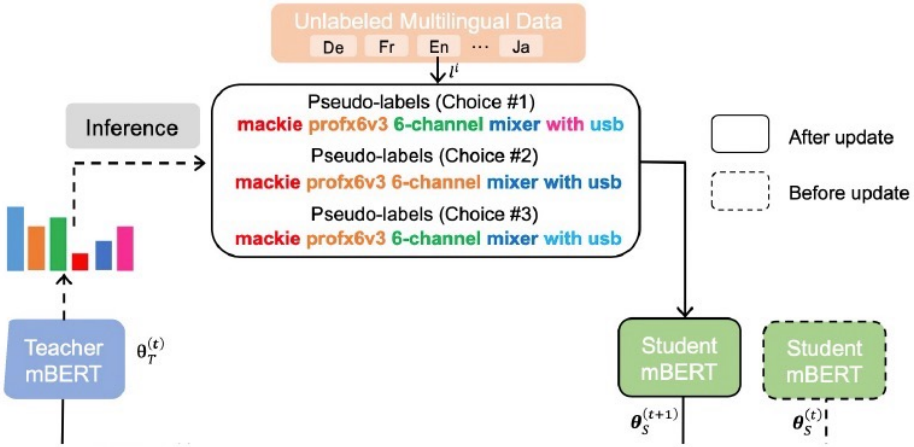
$$\boldsymbol{\theta}_S^{(t+1)} = \arg\min_{\boldsymbol{\theta}} \frac{1}{\widetilde{M}^{l_i}} \sum_{m=1}^{\widetilde{M}^{l_i}} \ell(\widetilde{\mathbf{Y}}_m^{l_i,(t)}, f(\widetilde{\mathbf{X}}_m^{l_i}; \boldsymbol{\theta}_S^{(t)}))$$

labeled data for language l_i $\{(\mathbf{X}_m^{l_i}, \mathbf{Y}_m^{l_i})\}_{m=1}^{M^{l_i}}$

unlabeled data for language l_i $\{\widetilde{\mathbf{X}}_m^{l_i}\}_{m=1}^{\widetilde{M}^{l_i}}$

teacher $\boldsymbol{\theta}_T$

student $\boldsymbol{\theta}_S$



Teacher Network

- Teacher loss

$$\mathcal{L}_{\mathrm{T}} = \mathcal{L}_{\mathrm{sup}} + \mathcal{L}_{\mathrm{reg}} + \mathcal{L}_{\mathrm{meta}}$$

- Supervised loss on labeled data

$$\mathcal{L}_{\mathrm{sup}} = \frac{1}{M^{l_i}} \sum_{m=1}^{M^{l_i}} \ell(\mathbf{Y}_m^{l_i}, f(\mathbf{X}_m^{l_i}; \boldsymbol{\theta}_T^{(t)}))$$

- Regularization loss on unlabeled data

$$\mathcal{L}_{\mathrm{reg}} = -\frac{1}{M^{l_i} N} \sum_{m,n=1}^{M^{l_i},N} \mathbb{I}(z_{m,n}^{\max}) \frac{\mathbf{z}_{m,n}}{\tau} \log \mathbf{z}_{m,n}^{G}$$

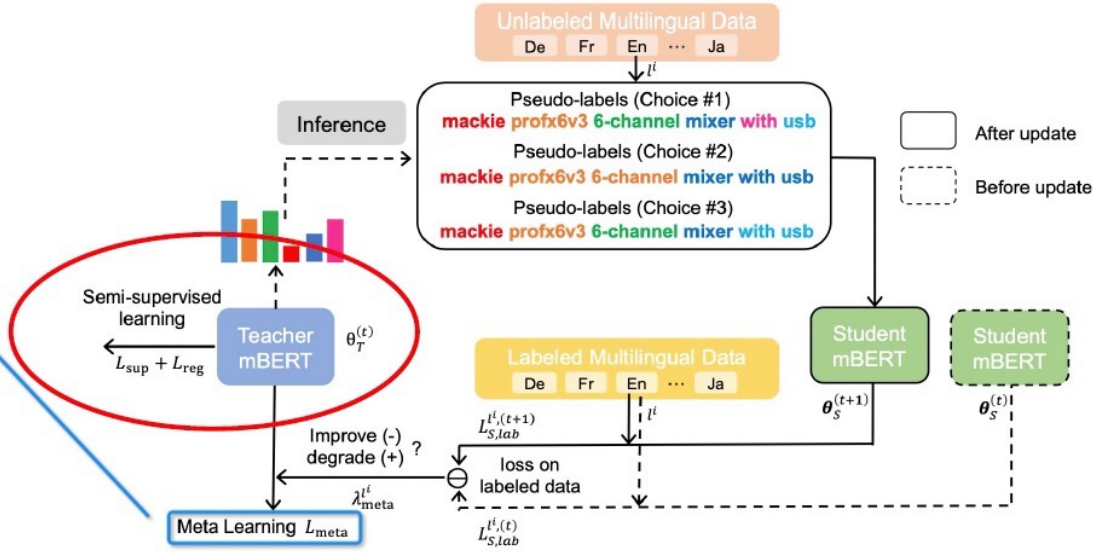


Figure 1: The framework of the Meta Teacher-Student Network (MetaTS).

enforce consistent prediction on unlabeled samples and augmented version (Gaussian noises)

Main Results – Public Datasets

Method (*Span F1*)	En	Es	De	Nl	Avg	Δ
Fully-supervised Baselines (1% labeled data)						
mBERT (Single)	83.03	75.62	67.31	73.67	74.91	(+5.91)
mBERT (Multi)	82.54	79.90	73.32	79.78	78.88	(+1.94)
Semi-supervised Baselines (1% labeled data)						
MT (KL)	83.52	77.99	73.40	80.71	78.91	(+1.91)
MT (MSE)	84.25	79.45	73.95	79.98	79.46	(+1.36)
VAT	83.70	78.27	73.02	81.00	79.00	(+1.82)
NoisyStudent	82.54	79.21	71.08	78.38	77.80	(+3.02)
BOND (hard)	82.75	78.31	75.74	80.29	79.27	(+1.55)
BOND (soft)	85.26	78.39	75.21	78.40	79.32	(+1.50)
BOND (soft-high)	84.62	79.87	72.68	80.31	79.37	(+1.45)
MetaTS (Ours)	**85.67**†	**80.05**	**76.23**†	**81.31**†	**80.82**†	-
Upper Bound (100% labeled data)						
mBERT (Full)	90.34	85.99	81.66	89.43	86.85	-

Table 4: The results (%) on multilingual open-domain NER. Δ refers to the improvements. † means the statistically significant improvement over the best baseline with paired sample t-test $p < 0.01$.

Method (*Span F1*)	En	Fr	Es	Tr	Nl	Ru	Avg	Δ
Fully-supervised Baselines (10% labeled data)								
mBERT (Single)	49.39	40.89	52.38	27.75	38.06	44.12	42.10	(+10.50)
mBERT (Multi)	55.85	47.61	58.37	29.24	46.51	46.15	47.29	(+5.31)
Semi-supervised Baselines (10% labeled data)								
MT (KL)	56.64	47.06	60.76	28.38	46.80	49.56	48.20	(+4.40)
MT (MSE)	54.56	48.53	60.88	30.65	47.26	50.28	48.69	(+3.91)
VAT	54.12	46.03	58.84	33.99	46.47	50.35	48.30	(+4.30)
NoisyStudent	55.90	47.13	56.89	34.92	47.53	49.11	48.58	(+4.02)
BOND (hard)	57.36	48.84	59.71	36.62	46.98	48.56	49.68	(+2.92)
BOND (soft)	56.34	50.40	61.95	33.78	**50.62**	48.14	50.21	(+2.39)
BOND (soft-high)	56.70	49.74	61.08	35.62	47.48	51.42	50.34	(+2.26)
MetaTS (Ours)	**59.45**†	**54.29**†	**62.90**†	**37.15**†	50.27	**51.51**	**52.60**†	-
Upper Bound (100% labeled data)								
mBERT (Full)	61.54	57.76	65.80	43.11	58.19	56.44	57.14	-

Table 5: The results (%) on multilingual E2E-ABSA.

Outline

- Conference
- Key points
- Main Conference
- Poster
- Tutorials
- Workshop

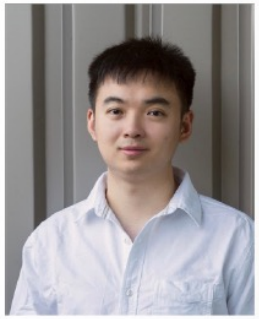
Tutorials

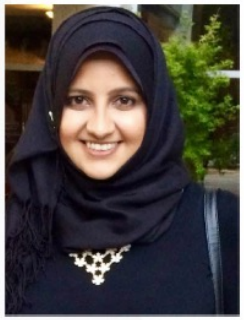
Slides/figures are borrowed from the speakers of respective presentation talk.

EMNLP 2021 Tutorial

Knowledge-Enriched Natural Language Generation





Wenhao Yu[1], Meng Jiang[1], Zhiting Hu[2], Qingyun Wang[3], Heng Ji[3,4], Nazneen Rajani[5]

1 University of Notre Dame 2 University of California San Diego
3 University of Illinois at Urbana-Champaign 4 Amazon Scholar 5 Salesforce Research

Text generation is needed everywhere in our life!

Image Captioning:
Image-to-Text Generation

85	Final	95

ROCKETS | LAKERS | STATS | NEWS

Player	Min	Reb	Ast	Pts
6 LeBron James · F	35	7	8	15
9 Kent Bazemore · F	27	2	1	9
3 Anthony Davis · C	33	13	2	16
0 Russell Westbrook · G	35	8	9	20
20 Avery Bradley · G	30	3	0	2
7 Carmelo Anthony	25	3	0	23
15 Austin Reaves	20	2	0	2
11 Malik Monk	19	2	0	0
10 DeAndre Jordan	17	3	2	8

HOU 1-5 **85** Final **95** LAL 4-3

LOS ANGELES -- — Carmelo Anthony scored 23 points in a reserve role and Russell Westbrook added 20 points in the Los Angeles Lakers' 95-85 victory over the Houston Rockets on Sunday night.

Anthony Davis had 16 points and 13 rebounds, and LeBron James scored 15 points in the Lakers' fourth win in five games after opening the season with two losses. Westbrook added eight rebounds and nine assists.

Eric Gordon scored 17 points and Christian Wood had 16 for the Rockets, who lost their fourth straight at the start of a five-game road trip.

Sports News Generation:
Table-to-Text Generation

Text-to-text generation is needed everywhere!

- $P(Y|X) = P(y_1, \cdots, y_m | x_1, \cdots, x_n) = \prod_{t=1}^{m} p(y_t | X, y_1, \cdots, y_{t-1})$, when Y is text and X is text.

Dialog systems

What should I wear outside today? X

The temperature will be 85 at noon. You can wear short sleeve shorts. Y

Summarization

X: *"... Her husband was one of 17 people killed in January's terror attacks in Paris. ... Valerie Braham said to the assembled crowd. ..."*

Y: "… Philippe Braham was killed in the January's terror attacks. …"

Machine Translation

X: *"... (Two years after, Maria was born ...) … Her brother gave her a hug. ..."*

Y: "… 她的哥哥给了她一个大大的拥抱. …"

Paraphrasing

X: *"... The game ended with the umpire making a bad call, and if the call had gone the other way, the Blue Whales might have actually won the game. It wasn't a victory, but I say the Blue Whales look like they have a shot at the championship, especially if they continue to improve."*

Y: *"... The match ended with the referee calling wrongly, and if the call went the other way, the Blue Whales could already win the match. It was not a victory, but I say that the Blue Whales seem to have a chance in the championship, especially if they keep improving."*

Knowledge is needed everywhere in text generation!

- Mistakes may occur due to lack of knowledge.

What should I wear outside today?

The temperature will be ~~65~~ 85 at noon. You can wear short sleeve shorts.

Dialog systems

"... Her husband was one of 17 people killed in January's terror attacks in Paris. ... Valerie Braham said to the assembled crowd. ..."

→ "... ~~Valerie Braham~~ Philippe Braham was killed in the January's terror attacks. ..."

Summarization

"... (Two years after, Maria was born ...) ... Her brother gave her a hug. ..."

↓

"... 她的~~弟弟~~哥哥给了她一个大大的拥抱. ..."

Machine Translation

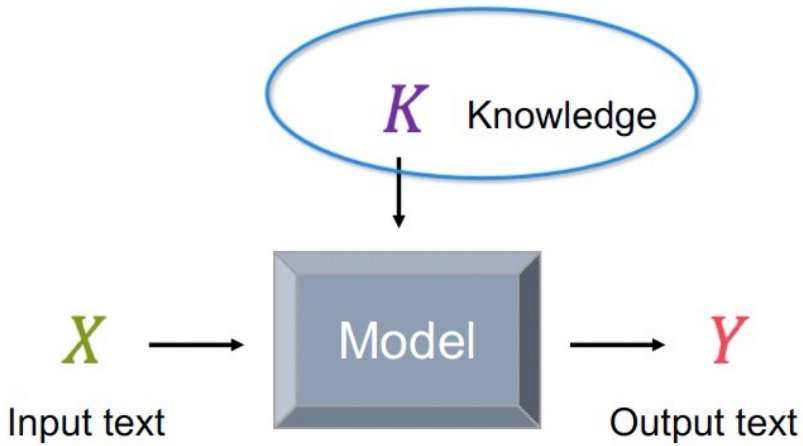
"... The game ended with the umpire making a bad call, and if the call had gone the other way, the Blue Whales might have actually won the game. It wasn't a victory, but I say the Blue Whales look like they have a shot at the championship, especially if they continue to improve."

→ *"... The match ended with the referee calling wrongly, and if the call went the other way, the Blue Whales could ~~never~~ already win the match. It was not a victory, but I say that the Blue Whales seem to have a chance in the championship, especially if they keep improving."*

Paraphrasing

6

Knowledge-Enriched Natural Language Generation (KE-NLG)



Key questions: (Outline of this tutorial)

- Where does the Knowledge come from? In other words, what are the Knowledge sources?
- What are the data/representations of Knowledge?
- What are the methods for integrating Knowledge into NLG models? How to utilize different types of Knowledge?
- What are the concrete NLG models/techniques that have been enhanced by Knowledge? What are their advantages and disadvantages?
- Where can we find benchmark datasets, code library, and hands-on tutorial for doing research on this topic?
- What are the remaining challenges and future directions?

7

達摩院 ALIBABA DAMO ACADEMY

Alibaba Group 阿里巴巴集团

Google DeepMind IBM Research

Multi-Domain Multilingual Question Answering

Sebastian Ruder and Avirup Sil
Google, IBM Research AI

Twitter : @seb_ruder, @aviaviavi__

Date of the tutorial: November 11, 2021

Who are we?

Sebastian

- Research scientist at DeepMind
- Author of multilingual benchmarks (XQuAD, XTREME) and multilingual models (RemBERT)

Previously at: DeepMind

Now at: Google

Avi

- Principal Research Scientist & Manager
- Area Chair for QA at NAACL, ACL etc.
- Several QA publications incl. the IBM GAAMA system
  - Top ranked on Natural Questions, TyDI, XOR TyDI

IBM Research

2

Disclaimer: This tutorial is our **own opinion & findings**

- Not Google's, DeepMind's or IBM's
- Please look into the licensing documentation of the individual models + datasets
- We're not promoting the use of any particular model and/or datasets
- Slides / figures are borrowed from the authors of respective papers (either via personal communication or from the internet)
- This tutorial is by no means exhaustive: we've tried our best to include relevant materials
  - there might still be some publications which we may have missed.

Links

- Check out the latest information regarding the tutorial on the GitHub page:
  https://github.com/sebastianruder/emnlp2021-multiqa-tutorial

- You can find the slides at:
  https://tinyurl.com/multi-qa-tutorial

- Tweet : #multiqa-tutorial

Outline

- Conference
- Key points
- Main Conference
- Poster
- Tutorials
- Workshop

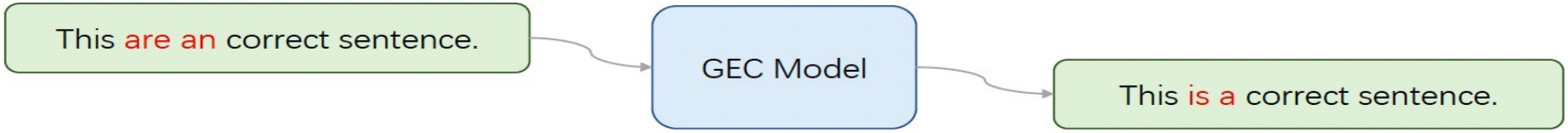
Workshop

Slides/figures are borrowed from the authors of respective presentation talk.

Data Augmentation of Incorporating Real Error Patterns and Linguistic Knowledge for Grammatical Error Correction

Xia Li and **Junyi He**
School of Information Science and Technology
Guangdong University of Foreign Studies, Guangzhou, China
xiali@gdufs.edu.cn, zeonngai_ho@163.com

Example:



2. Motivation

(2) Limitations of Current GEC DA Methods

Many GEC studies augment training data by introducing random errors (random deletion, random replacement, etc.) into texts.

It will generate unreal and low-quality synthetic errors which could be further propagated into the GEC model and harm its performance.

Example:

original sentence: **This is a correct sentence.**

augmented sentence: **point This what a correct.**

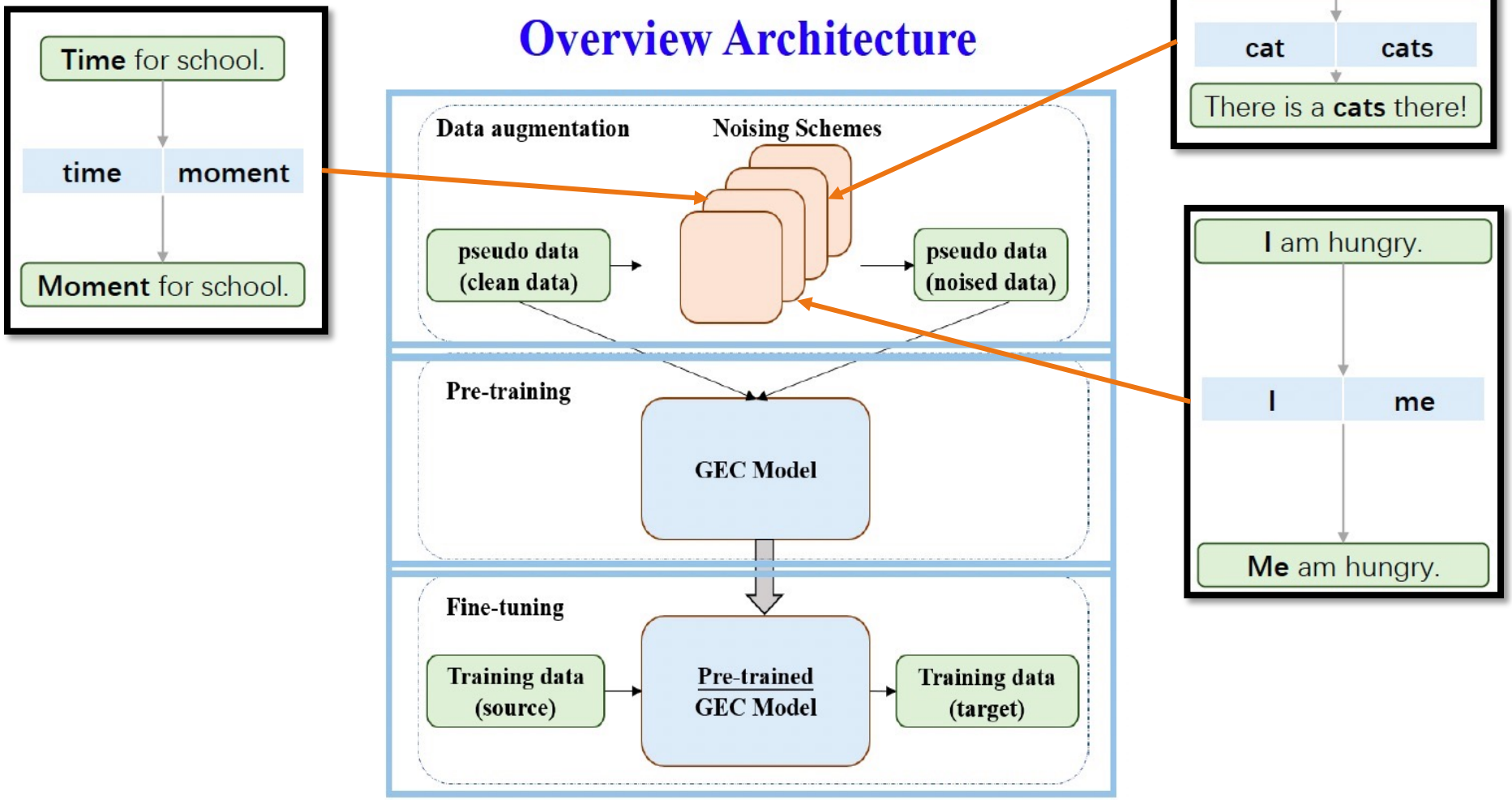
3. Our Method

We propose a novel data augmentation method for the GEC task containing **four noising schemes**, which generates:

- more **real** & **high-quality** synthetic data.
- more **representative** & **diverse** synthetic data.

Noising Scheme	Characteristics
Real error pattern based noising scheme	• Uses **real error patterns** manually annotated by human experts • Introduces **real** & **high-quality** grammatical errors
Synonym noising scheme	• Incorporates **linguistic knowledge** • Introduces **representative** & **diverse** grammatical errors
Inflection noising scheme	
Functional word noising scheme	

3. Our Method


Time for school.
time
moment
Moment for school.
Overview Architecture
Data augmentation
Noising Schemes
pseudo data (clean data)
pseudo data (noised data)
Pre-training
GEC Model
Fine-tuning
Training data (source)
Pre-trained GEC Model
Training data (target)
There is a cat there!
cat
cats
There is a cats there!
I am hungry.
I
me
Me am hungry.

Dataset	#Pairs	Split
FCE-train	32,073	Train
NUCLE	57,119	Train
Lang-8	1,097,274	Train
CoNLL-2013	1,381	Valid
CoNLL-2014	1,312	Test
OBC (partial)	4.5M	Pretrain

Table 4: Summary of each dataset. #Pairs indicates the number of sentence pairs in each dataset.

Noising Scheme	Precision	Recall	$F_{0.5}$
All	**28.9**	**7.1**	**18.0**
- **Real Pattern**	24.1	6.6	15.7
- **Synonym**	28.1	6.9	17.4
- **Inflection**	25.0	6.6	16.1
- **Functional Word**	26.6	6.8	16.8

Table 8: Contribution of each noising scheme. The first row (**All**) is the result of using all noising schemes. And the remaining are the results of removing one of the noising schemes each.

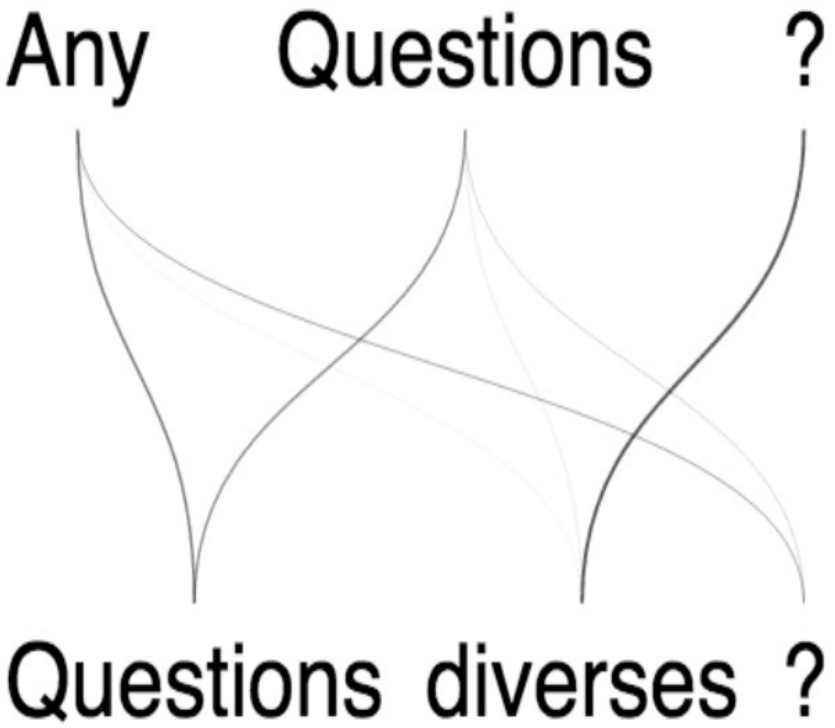
Model	#Pseudo Data	P	R	$F_{0.5}$
Vanilla Transformer-copy	0M (w/o pre-training)	65.8	29.3	52.7
Pre-training Decoder (Zhao et al., 2019)	30M	68.0	35.0	57.2
Denoising Auto-encoder (Zhao et al., 2019)	30M	69.0	37.0	58.8
Error type w/ Data selection (Takahashi et al., 2020)	10M	**69.1**	34.5	57.6
Real Error Patterns & Linguistic Knowledge (Ours)	**5M**	69.0	**38.3**	**59.4**
Lichtarge et al., 2019 (Lichtarge et al., 2019)	170M	65.5	37.1	56.8
Kiyono et al., 2019 (Kiyono et al., 2019)	70M	67.9	**44.1**	**61.3**

Table 6: Experimental results of the GEC models. The first section lists the result of the vanilla Transformer-copy model without pretraining. The second section lists results of our proposed method and several baselines for comparison, all of which are based on the Transformer-copy architecture. "Ours" denotes our method. We also collect results of some relevant works in the third section. #Pseudo Data indicates the amount of pretraining data used by each model. **Bold** indicates the highest score in each column.

Conclusion

- Self-supervised Learning
  - MT, QA, NER
  - + PTM, + Contrastive learning
- Contrastive Learning
  - Data augmentation + Curriculum learning
- Prompting/prompted
- Curriculum learning + sequence generation
- Jointly training
  - Joint encoding
  - Joint Pre-training

Thank You!



This inspiration comes from Dzmitry Bahdanau @ ICLR2014