Advances and Challenges in Unsupervised Neural Machine Translation

Rui Wang and Hai Zhao

Department of Computer Science and Engineering

Shanghai Jiao Tong University

https://wangruinlp.github.io/unmt

Menu

- About Us
- Towards Unsupervised Neural Machine Translation (UNMT)
  - Background of Machine Translation (MT)
  - Supervision in MT
  - Unsupervised MT
- Advances in UNMT
  - Pre-trained (Cross-lingual) Language Model
  - Multilingual UNMT
- Challenges in UNMT
  - Reproductive Baselines
  - UNMT & Supervised NMT
  - Distance Language Pairs

Menu

- About Us
- Towards Unsupervised Neural Machine Translation (UNMT)
  - Background of Machine Translation (MT)
  - Supervision in MT
  - Unsupervised MT
- Advances in UNMT
  - Pre-trained (Cross-lingual) Language Model
  - Multilingual UNMT
- Challenges in UNMT
  - Reproductive Baselines
  - UNMT & Supervised NMT
  - Distance Language Pairs

Rui Wang

- Associate Professor, Shanghai Jiao Tong University, Shanghai, China
- Research Interest:
  - Machine Translation
  - Multilingual NLP
- Homepage: https://wangruinlp.github.io/

Hai Zhao

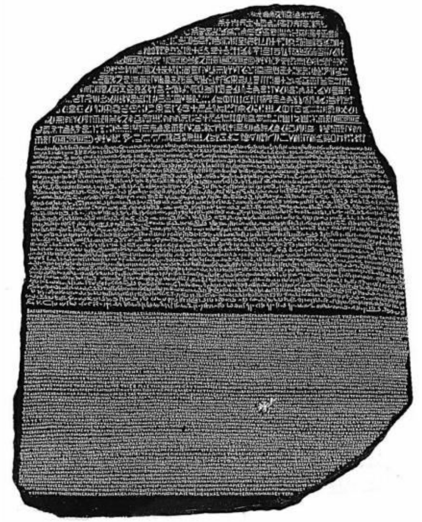
- Professor, Shanghai Jiao Tong University, Shanghai, China
- Research Interest:
  - Natural Language Processing
  - Machine Learning
  - Data Mining
  - Bioinformatics and Artificial Intelligence
- Homepage: http://bcmi.sjtu.edu.cn/~zhaohai/

Menu

- About Us
- Towards Unsupervised Neural Machine Translation (UNMT)
  - Background of Machine Translation (MT)
  - Supervision in MT
  - Unsupervised MT
- Advances in UNMT
  - Pre-trained (Cross-lingual) Language Model
  - Multilingual UNMT
- Challenges in UNMT
  - Reproductive Baselines
  - UNMT & Supervised NMT
  - Distance Language Pairs

MT: History

- Human Translation
  - 3rd~1st BC: Bible Translation in West
  - 1st AD: Buddhism Translation in China



Ancient Egyptian (hieroglyphic)

Ancient Egyptian (Demotic)

Ancient Greek

Rosetta Stone (196 BC)

- Machine Translation:
  - Starting from 1949, treat the source language as an *encrypted* target language.
  - 1970s- Rule based MT.
  - 1980s- Example based MT.
  - 1990s- Statistical MT.
  - 2010s- Neural MT.

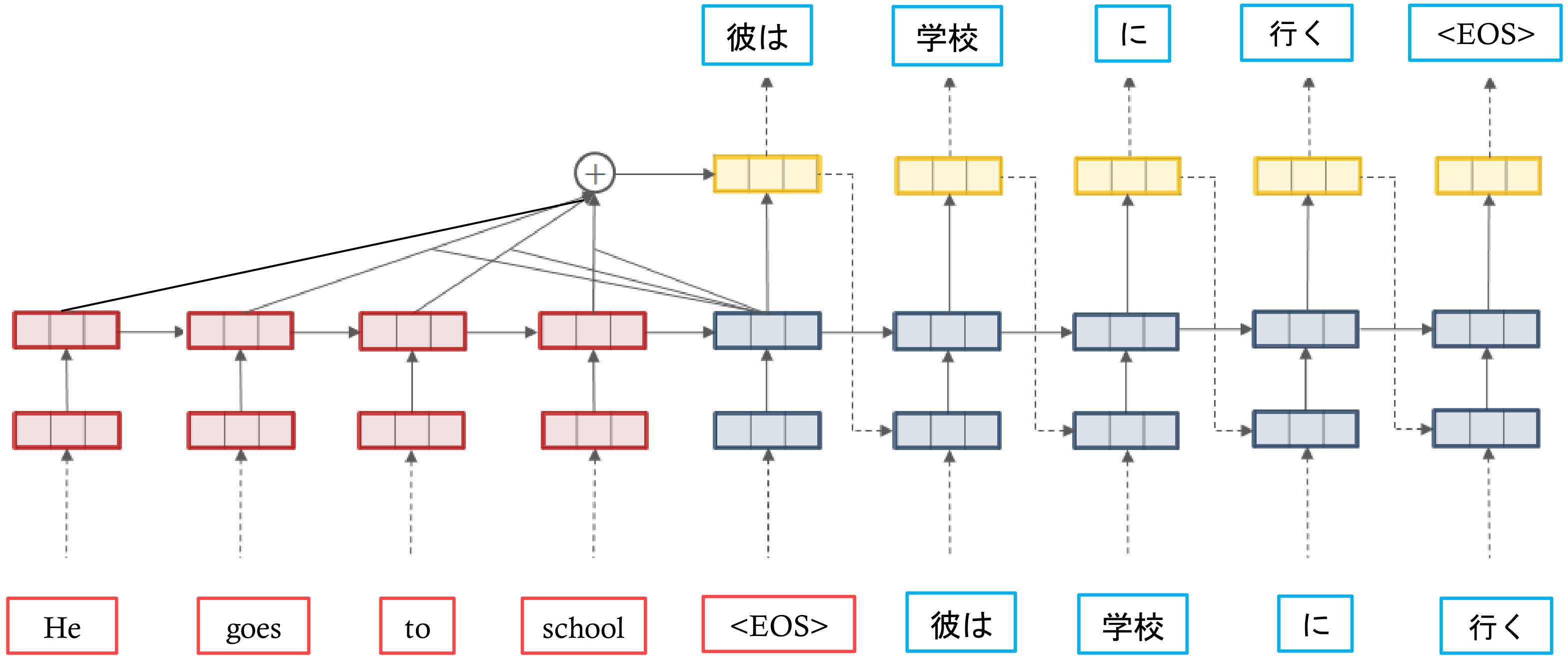
MT: from ML aspect

- MT is a typical text generation task.
  - $\boldsymbol{x}$: source sentence; $\boldsymbol{y}$: target sentence.
  - maximum likelihood estimation (MLE):

$$\mathcal{L}_{\mathbf{MLE}}(\theta) = -\log p_\theta(\boldsymbol{y}|\boldsymbol{x}) = -\sum_{i=1}^{l} \log p_\theta(y_i|\boldsymbol{x}, \boldsymbol{y}_{<i})$$

- MT has a standard evaluation metric:
  - n-gram: contiguous sequence of n words.

$$BLEU = \frac{\sum ngram_{correct}}{\sum ngram_{in\_reference}}$$



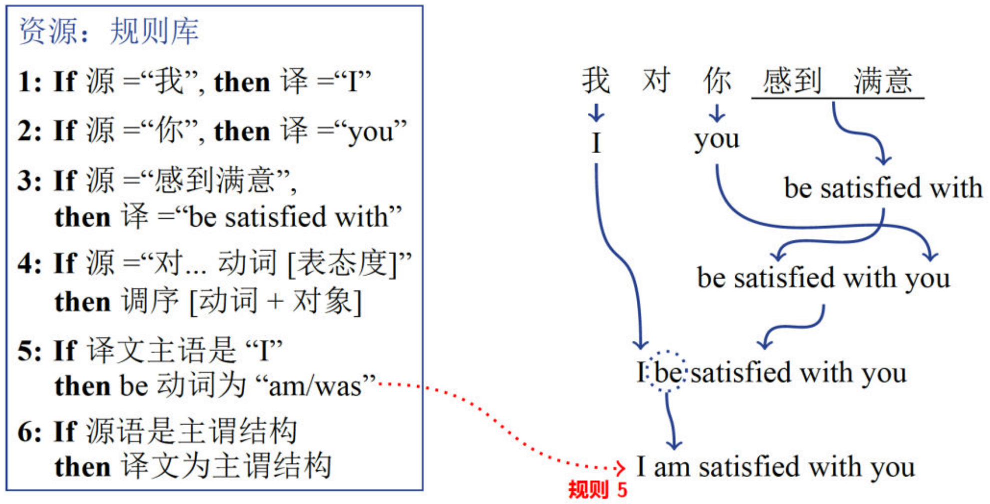
Menu

- About Us
- Towards Unsupervised Neural Machine Translation (UNMT)
  - Background of Machine Translation (MT)
  - Supervision in MT
  - Unsupervised MT
- Advances in UNMT
  - Pre-trained (Cross-lingual) Language Model
  - Multilingual UNMT
- Challenges in UNMT
  - Reproductive Baselines
  - UNMT & Supervised NMT
  - Distance Language Pairs

Supervision in MT

- Rule-based MT:
  - Annotated linguistic rules

Linguistic Rules



- Example-based MT:
  - Translation examples

Example Database

Lexicon

Matching and Replacement

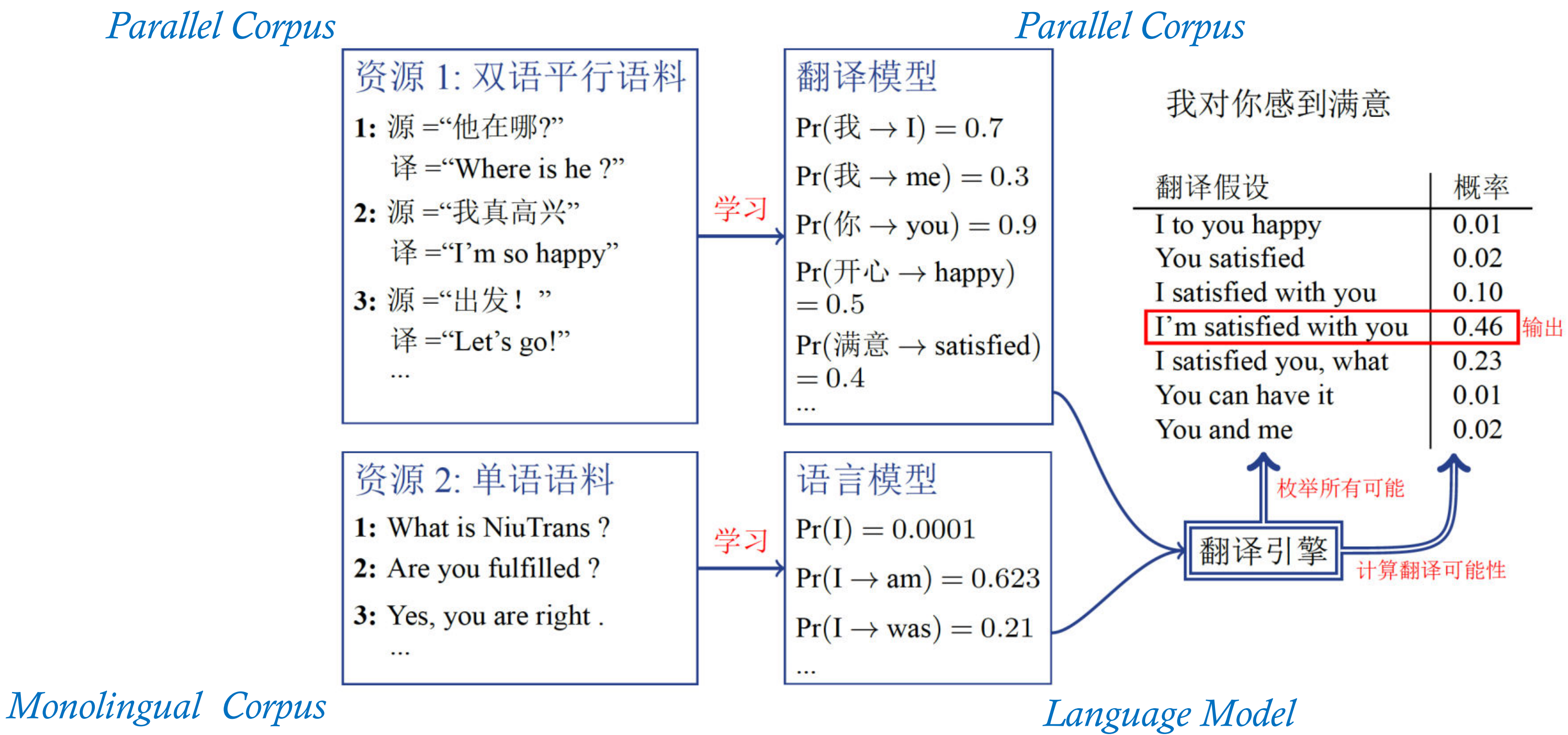
源 : source

译 : target

[Examples from Xiao and Zhu, SMT-Book]

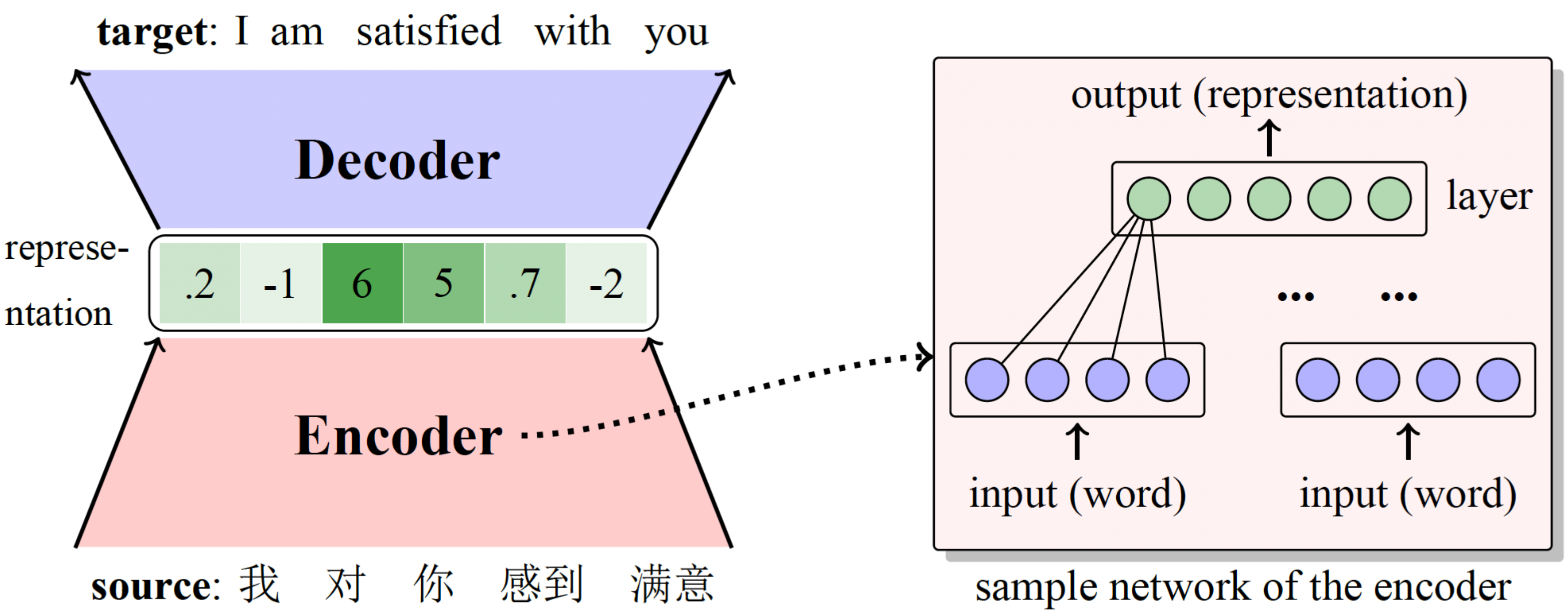
Supervision in MT

- Statistical Machine Translation (SMT)
  - Parallel corpus: sentence-level alignment.
  - Monolingual corpus: *n*-grams probability.
  - To learn the translation rules statistically.



Supervision in MT

- Neural Machine Translation (NMT):
  - Parallel corpus as sequence-to-sequence input.
  - Rules are not necessary any more.



What Is Supervision in MT

- Supervision in linguistic:
  - Shared words or subwords: *restaurant* in French and English. 一般 in Chinese and Japanese
  - The same or similar syntactic structure
  - The same or similar pronunciation
  - …

- **Supervision in machine learning: parallel input {X, Y} or monolingual input {X} and {Y}**
  - Bilingual lexicon
  - Phrase table
  - Parallel sentences
  - Comparable corpus/document
  - …

Does Supervised Always Necessary?

- My understanding
  - Supervision in linguistic is always necessary.
  - Supervision in machine learning is not always necessary.
- Definition of unsupervised MT in machine learning
  - No parallel training corpus is given.
  - Dev corpus is only used to select model.
- We will discuss this topic in the section “Challenges in UNMT”

Menu

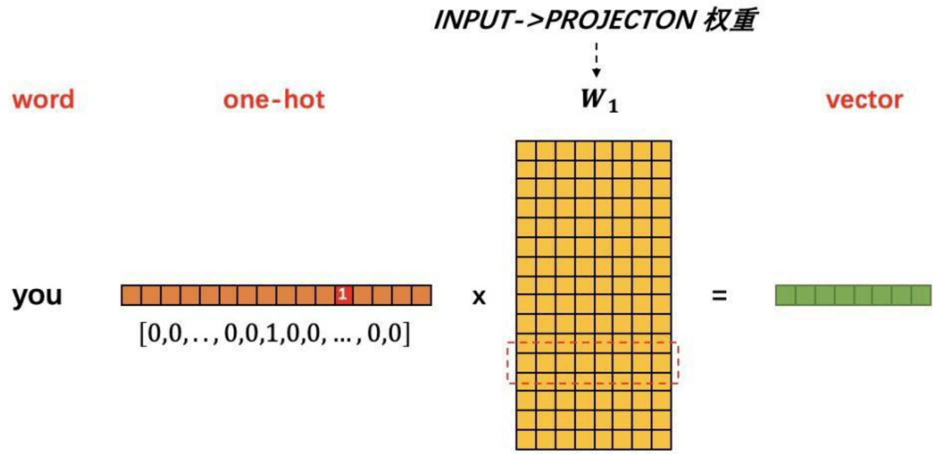
- About Us
- Towards Unsupervised Neural Machine Translation (UNMT)
  - Background of Machine Translation (MT)
  - Supervision in MT
  - Unsupervised MT
- Advances in UNMT
  - Pre-trained (Cross-lingual) Language Model
  - Multilingual UNMT
- Challenges in UNMT
  - Reproductive Baselines
  - UNMT & Supervised NMT
  - Distance Language Pairs

Monolingual Word Embedding

- As the development of neural network technology in NLP, words can be represented in continuous space.
- However, too sparse…

$$
\begin{aligned}
\mathrm{I} \Leftrightarrow V_{\mathrm{I}} &= [1,0,0,0,0,0,0,\ldots,0] \\
\text{you} \Leftrightarrow V_{\text{you}} &= [0,1,0,0,0,0,0,\ldots,0] \\
\text{is} \Leftrightarrow V_{\text{is}} &= [0,0,1,0,0,0,0,\ldots,0] \\
\text{are} \Leftrightarrow V_{\text{are}} &= [0,0,0,1,0,0,0,\ldots,0] \\
\text{very} \Leftrightarrow V_{\text{very}} &= [0,0,0,0,1,0,0,\ldots,0] \\
\text{wise} \Leftrightarrow V_{\text{wise}} &= [0,0,0,0,0,1,0,\ldots,0] \\
\text{smart} \Leftrightarrow V_{\text{smart}} &= [0,0,0,0,0,0,1,\ldots,0]
\end{aligned}
$$

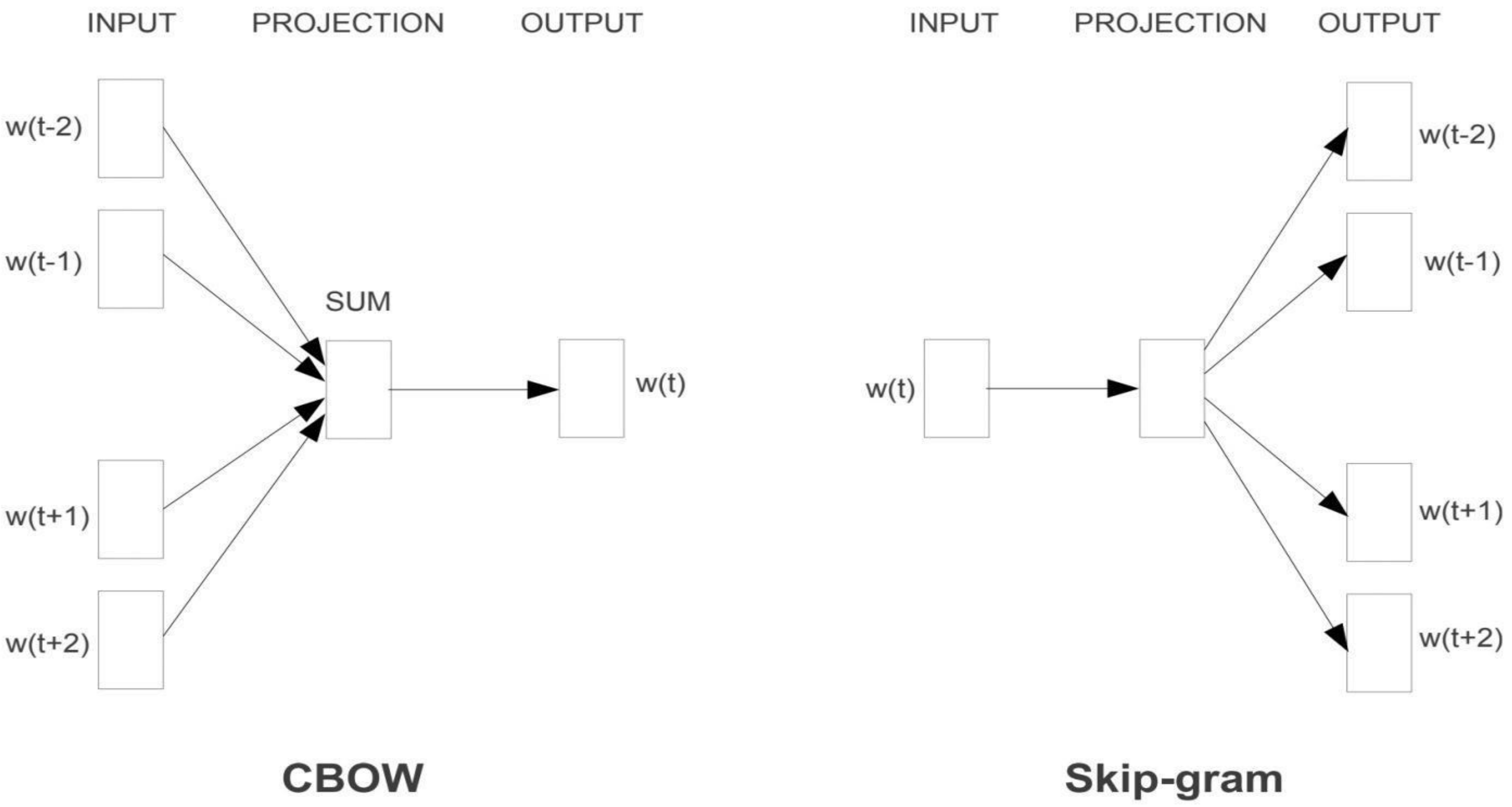
One-hot Representation



Projection

Monolingual Word Embedding

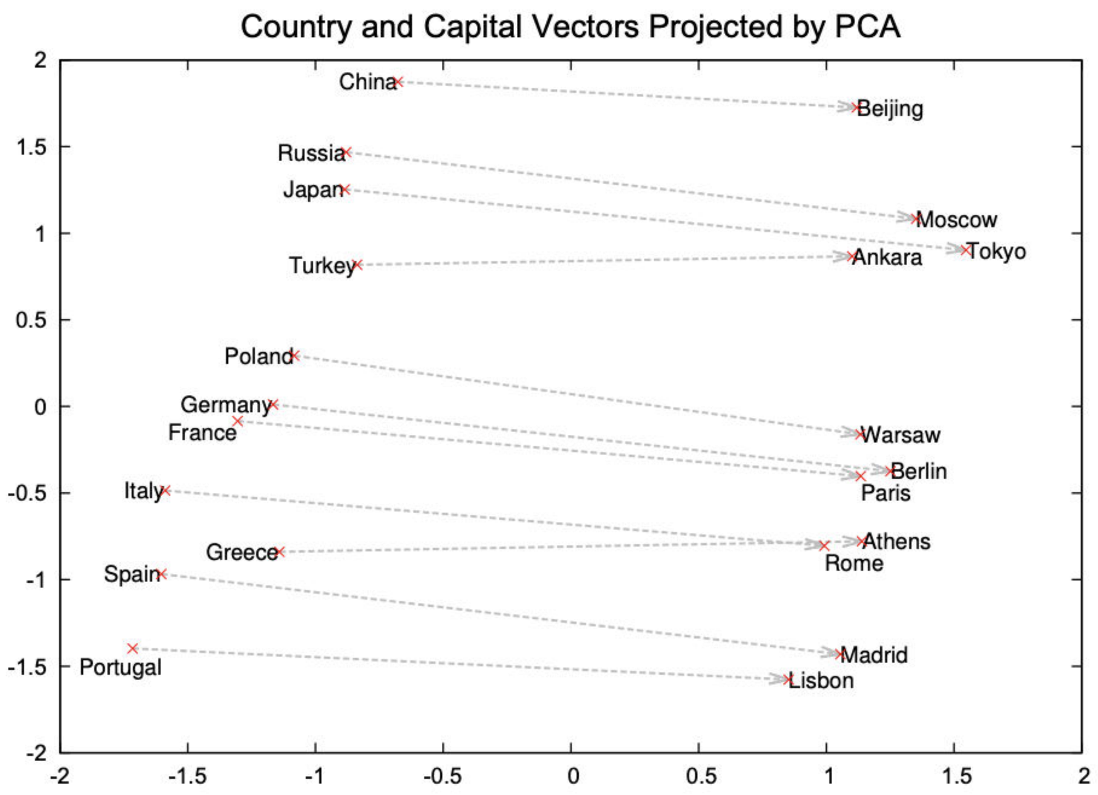
- Word2Vec



[Mikolov et al., NeurIPS-2013]

Monolingual Word Embedding

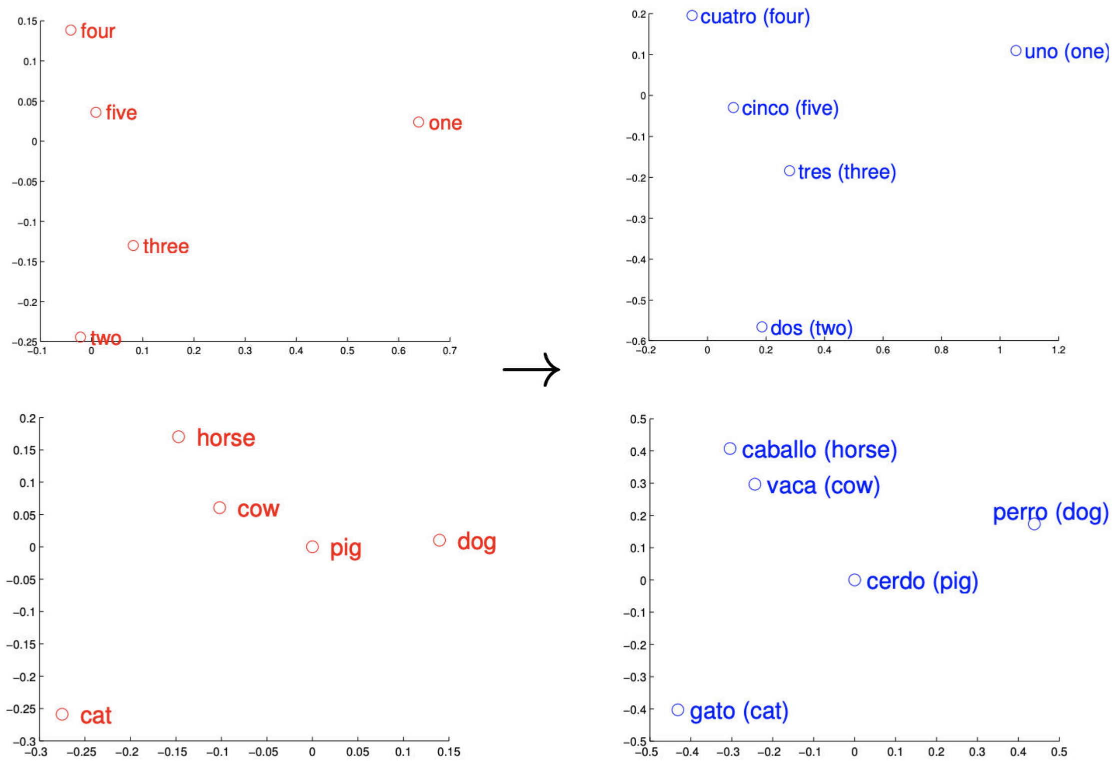
- Then, there is some interesting findings.



[Mikolov et al., NeurIPS-2013]

Bilingual Word Embedding (BWE)

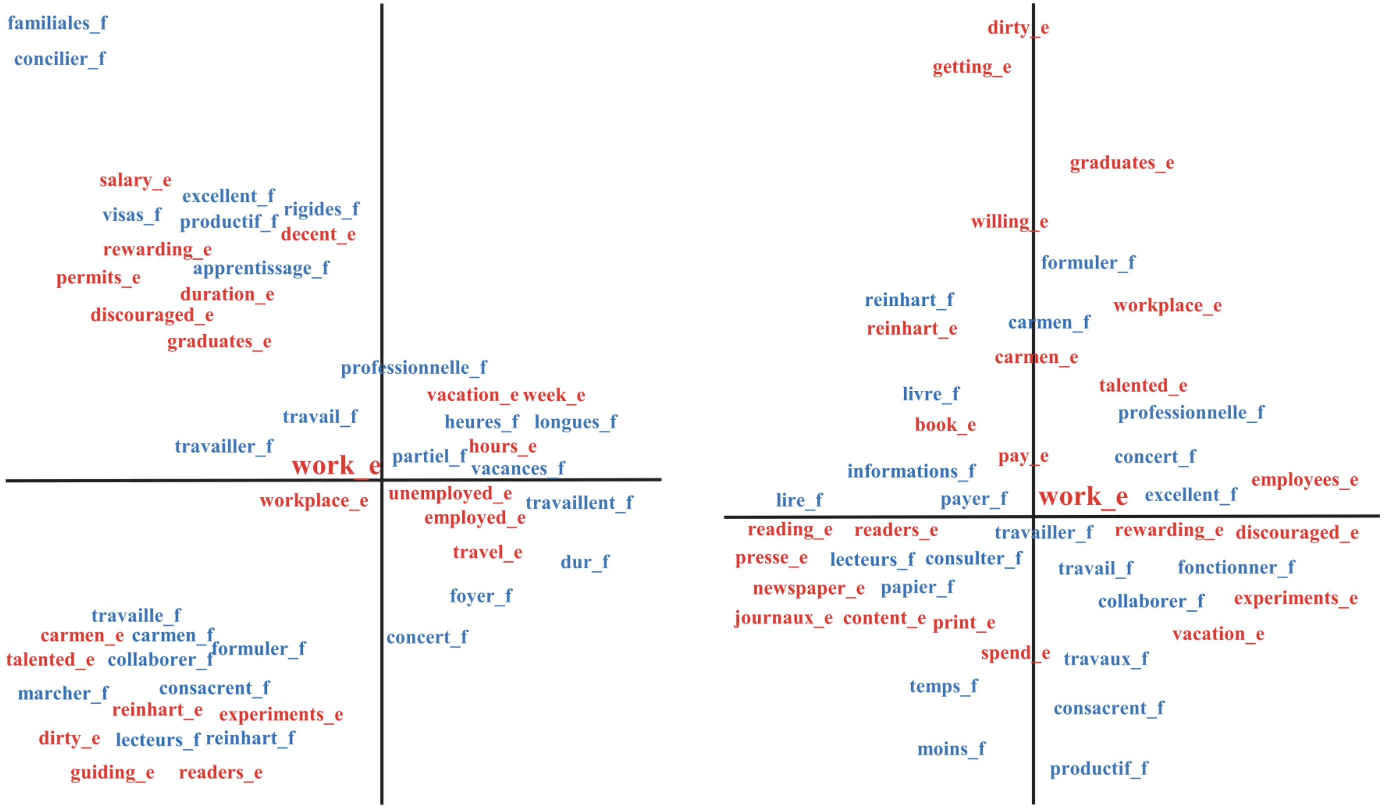
- To project one language space onto anther, researchers have to learn a translation map (matrix).
- The most typical supervision is an annotated lexicon (i.e., 5000 words).



[Mikolov et al., ArXiv-2013]

Bilingual Word Embedding (BWE)

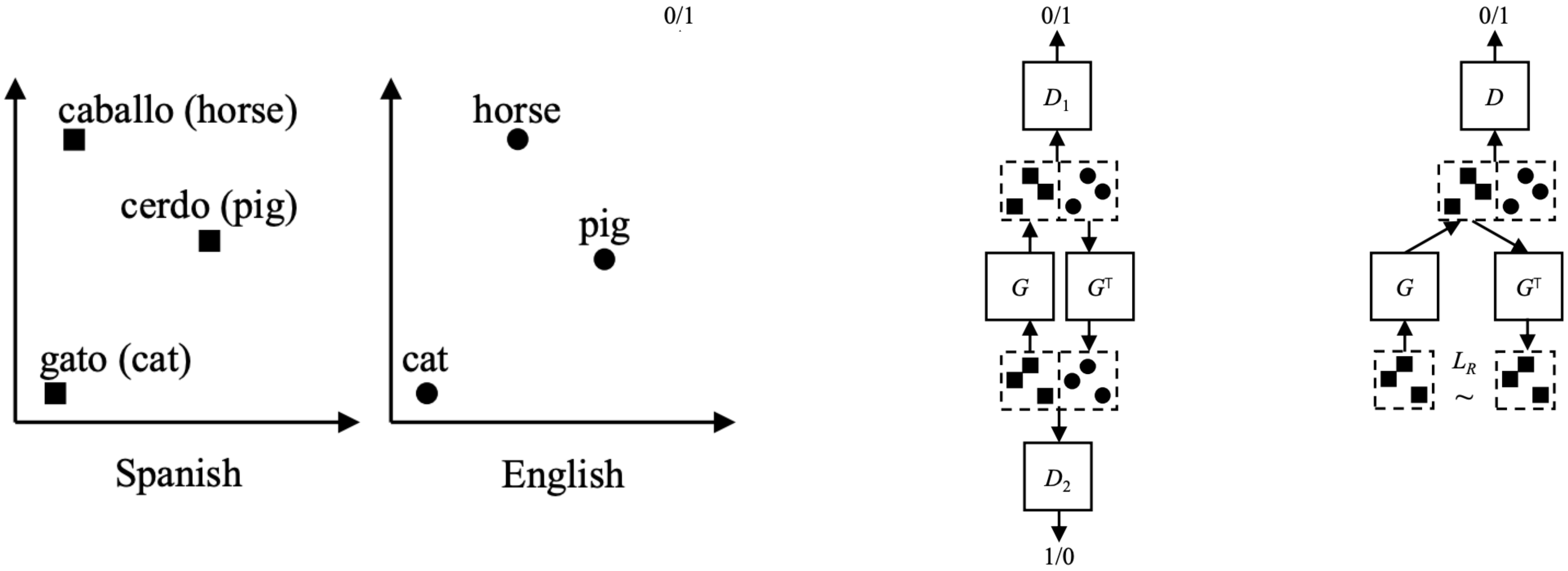
- Polysemy is not easy to project.
  - "Work" as a paid job or a research paper



[**Wang** et al., IJCAI-2016]

Unsupervised BWE

- Generative adversarial network (GAN) makes unsupervised BWE possible.
- The hypothesis is that different languages have similar word distribution.



[Zhang et al., ACL-2017]

BWE Performance

- No significant difference between supervised and unsupervised BWE

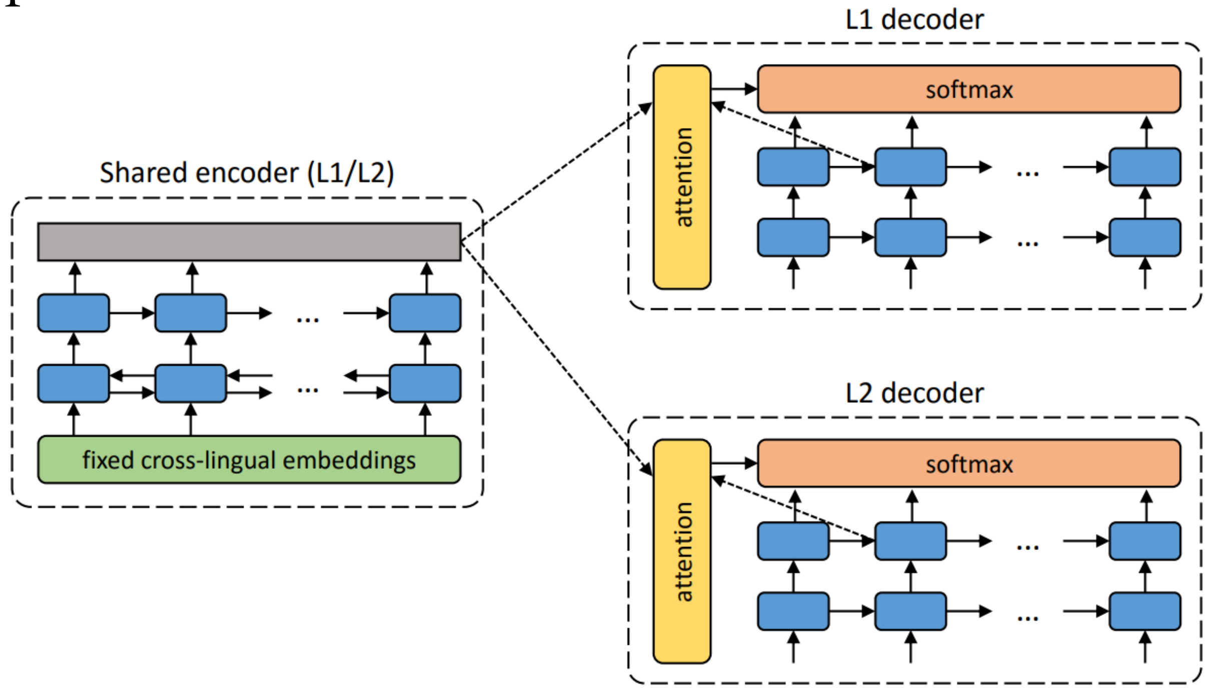
	en-de	en-fr	en-es	en-it	en-pt	de-fr	de-es	de-it	de-pt	fr-es	fr-it	fr-pt	es-it	es-pt	it-pt
Supervised methods with cross-lingual supervision															
Sup-BWE-Direct	73.5	81.1	81.4	77.3	79.9	73.3	67.7	69.5	59.1	82.6	83.2	78.1	83.5	87.3	81.0
Unsupervised methods without cross-lingual supervision															
BWE-Pivot	74.0	82.3	81.7	77.0	80.7	71.9	66.1	68.0	57.4	81.1	79.7	74.7	81.9	85.0	78.9
BWE-Direct	74.0	82.3	81.7	77.0	80.7	73.0	65.7	66.5	58.5	83.1	83.0	77.9	83.3	87.3	80.5
MAT+MPSR	**74.8**	**82.4**	**82.5**	**78.8**	**81.5**	**76.7**	**69.6**	**72.0**	**63.2**	**83.9**	**83.5**	**79.3**	**84.5**	**87.8**	**82.3**

	de-en	fr-en	es-en	it-en	pt-en	fr-de	es-de	it-de	pt-de	es-fr	it-fr	pt-fr	it-es	pt-es	pt-it
Supervised methods with cross-lingual supervision															
Sup-BWE-Direct	72.4	**82.4**	82.9	76.9	**80.3**	69.5	68.3	67.5	63.7	85.8	87.1	84.3	87.3	91.5	81.1
Unsupervised methods without cross-lingual supervision															
BWE-Pivot	72.2	82.1	83.3	**77.7**	80.1	68.1	67.9	66.1	63.1	84.7	86.5	82.6	85.8	91.3	79.2
BWE-Direct	72.2	82.1	83.3	**77.7**	80.1	69.7	68.8	62.5	60.5	86	87.6	83.9	87.7	92.1	80.6
MAT+MPSR	**72.9**	81.8	**83.7**	77.4	79.9	**71.2**	**69.0**	**69.5**	**65.7**	**86.9**	**88.1**	**86.3**	**88.2**	**92.7**	**82.6**

[Chen et al., EMNLP-2018]

What's Next?

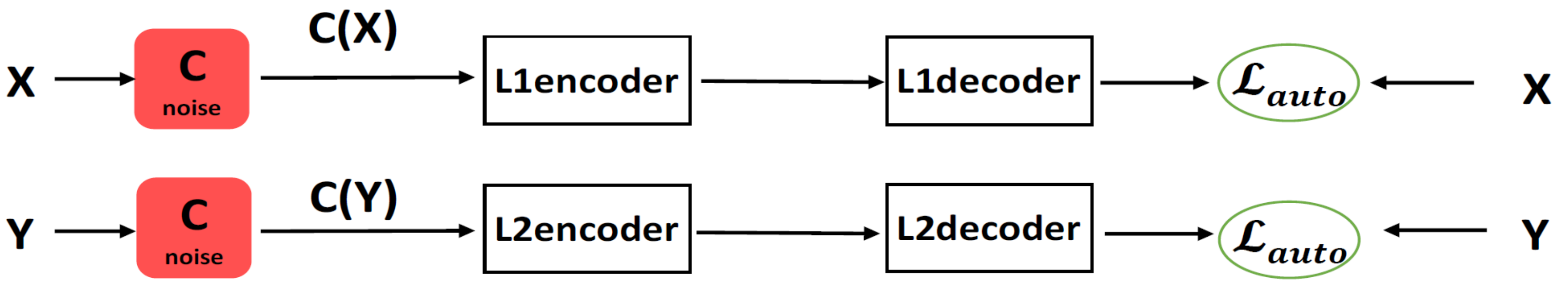
- Now we have word translation. How to conduct sentence translation?
- Initialization
  - Unsupervised bilingual word embedding
  - Cross-lingual language model
- Sharing latent representations



[Artetxe et al. ICLR-2018]

Unsupervised NMT

- Denoising: optimizes probability of reconstruction from a noised version *C(X)* in the encoder to the original sentence *(X)* in the decoder.

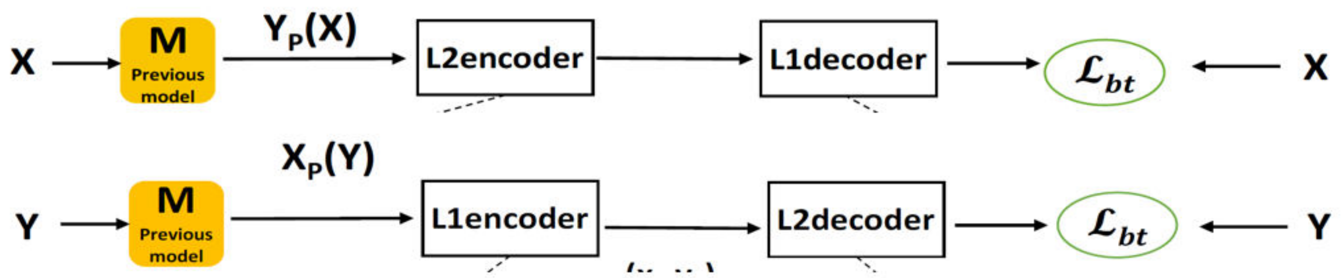


$$\mathcal{L}_D = \sum_{i=1}^{|X^1|} -logP_{L_1 \to L_1}(X_i^1 | C(X_i^1)) + \sum_{i=1}^{|X^2|} -logP_{L_2 \to L_2}(X_i^2 | C(X_i^2)),$$

Unsupervised NMT

- Back-translation
  - Optimizes the probability of encoding (pseudo parallel) translated sentence *M(X)* from L2 and recovering the original sentence ***X*** with the L1 decoder.

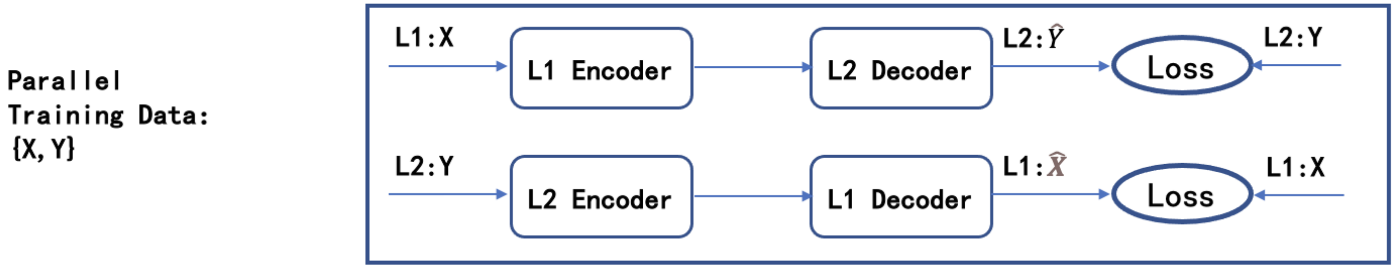
$$\mathcal{L}_B = \sum_{i=1}^{|X^1|} -logP_{L_2 \to L_1}(X_i^1 | M^2(X_i^1)) + \sum_{i=1}^{|X^2|} -logP_{L_1 \to L_2}(X_i^2 | M^1(X_i^2)),$$



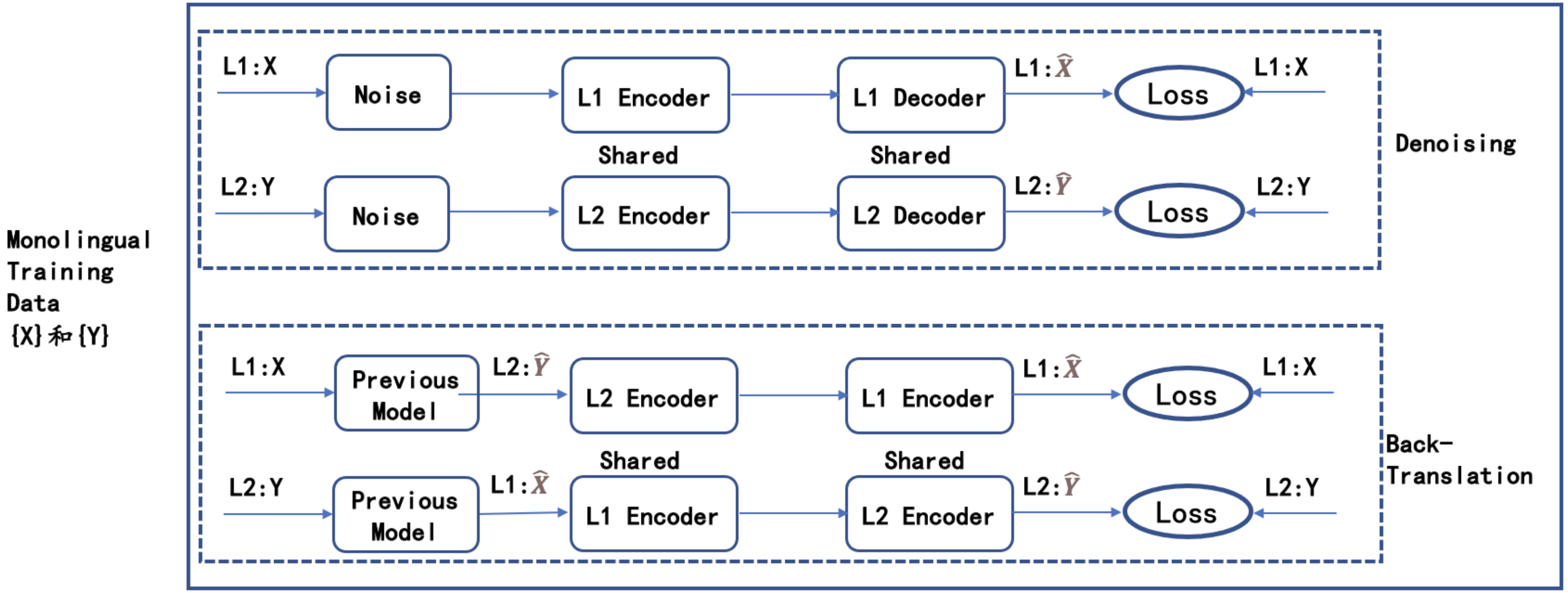
- Final Training Objective:
  - Jointly optimize the back-translation and denoising

$$\mathcal{L}_{all} = \mathcal{L}_D + \mathcal{L}_B.$$

Entire Structure



Supervised NMT



Unsupervised NMT

Performance of UNMT

- Much worse than supervised NMT
- Why?

		FR-EN	EN-FR	DE-EN	EN-DE
Unsupervised	1. Baseline (emb. nearest neighbor)	9.98	6.25	7.07	4.39
	2. Proposed (denoising)	7.28	5.33	3.64	2.40
	3. Proposed (+ backtranslation)	15.56	15.13	10.21	6.55
	4. Proposed (+ BPE)	15.56	14.36	10.16	6.89
Semi-supervised	5. Proposed (full) + 10k parallel	18.57	17.34	11.47	7.86
	6. Proposed (full) + 100k parallel	21.81	21.74	15.24	10.95
Supervised	7. Comparable NMT (10k parallel)	1.88	1.66	1.33	0.82
	8. Comparable NMT (100k parallel)	10.40	9.19	8.11	5.29
	9. Comparable NMT (full parallel)	20.48	19.89	15.04	11.05
	10. GNMT (Wu et al., 2016)	-	38.95	-	24.61

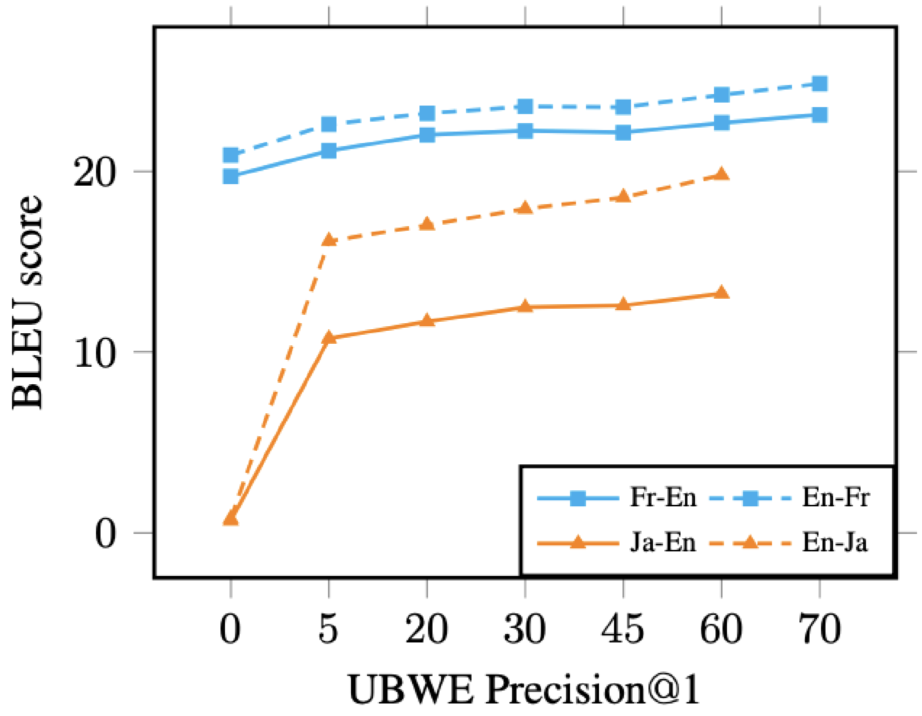
[Artetxe et al. ICLR-2018]

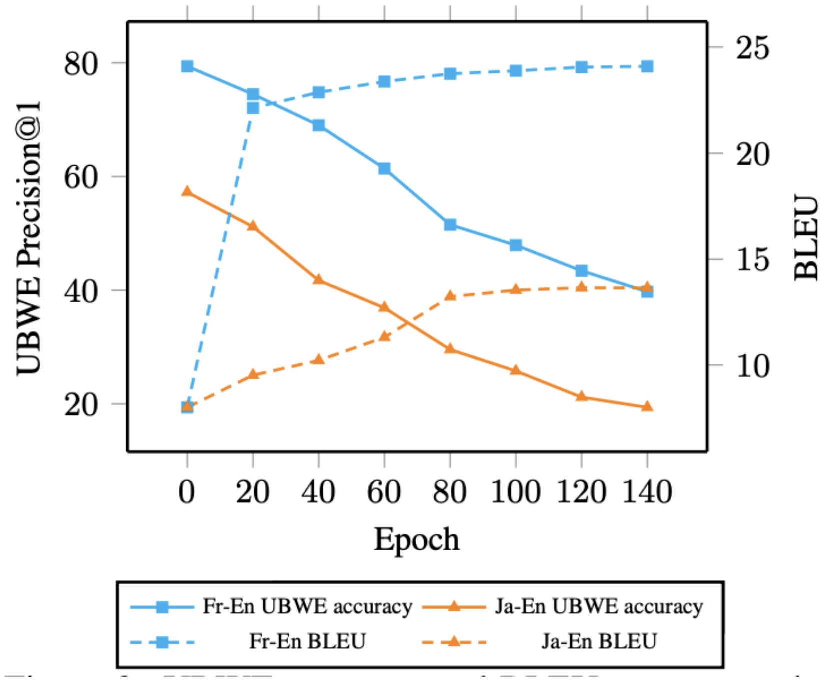
Key: Cross-Lingual Representation

- How to improve UNMT?
  - The back-translation and denoising is difficult to improve.
  - The key point is to improve the quality of cross-lingual representation.
- Method
  - Improve the pre-training of cross-lingual representation (the next chapter).
  - Improve cross-lingual representation during UNMT training.

Better Training

- The UNMT performance is related to the quality of UBWE.
- However, the quality of UBWE significantly decreases during UNMT training.

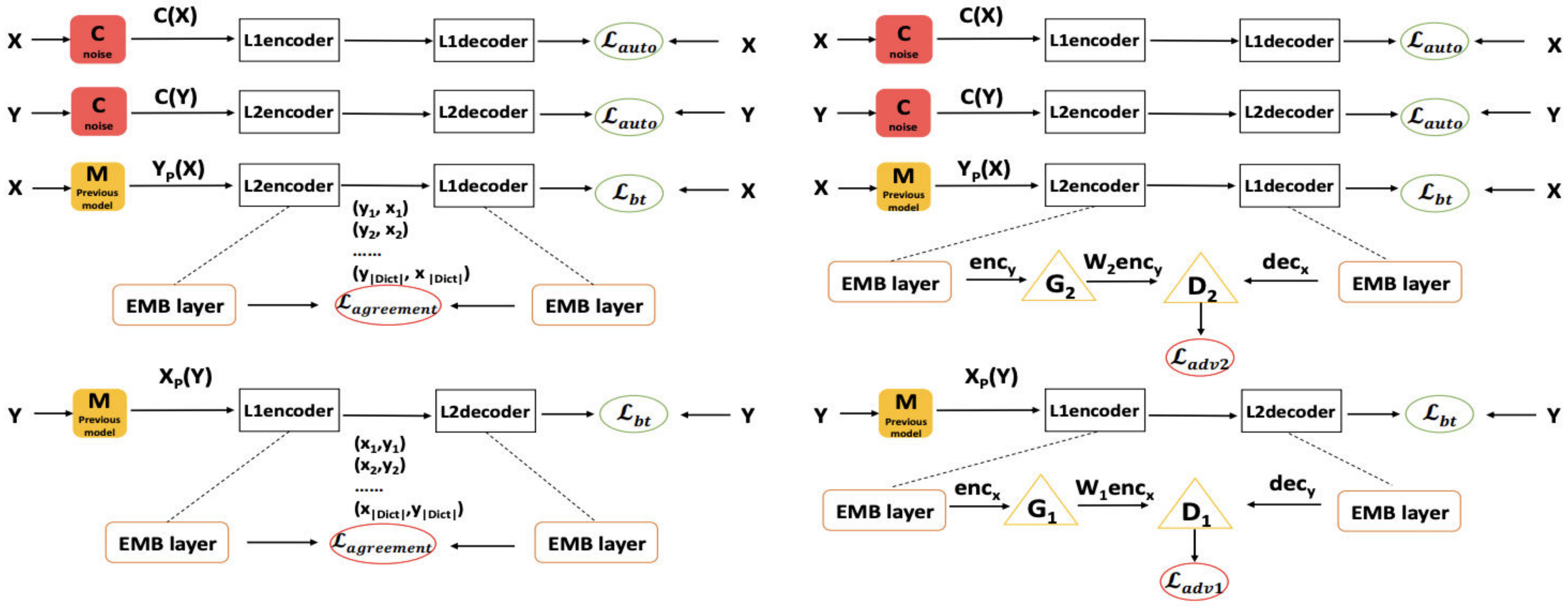




[Sun and **Wang*** et al. ACL-2019]

Joint UBWE and UNMT Training

- Our contribution
  - We propose a joint UBWE and UNMT training method.

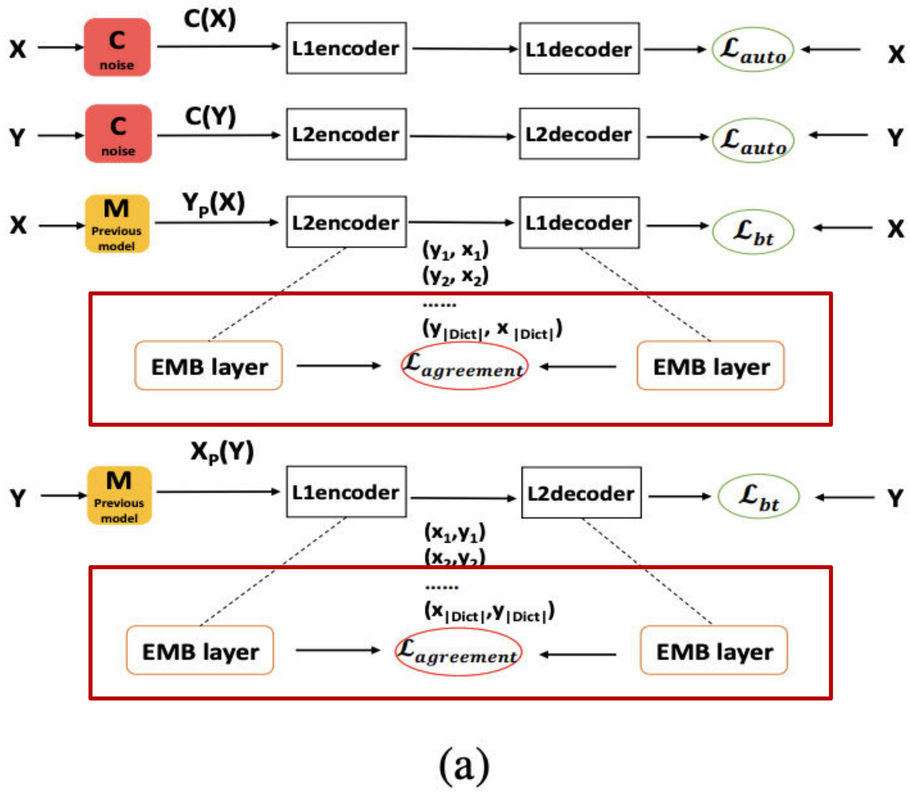


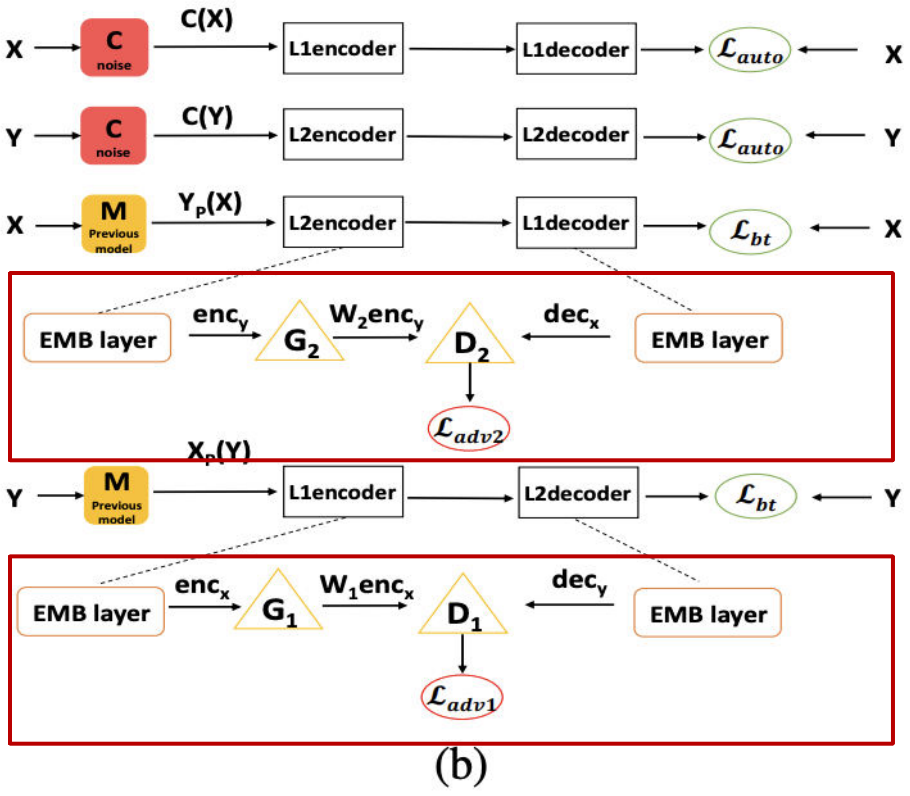
(a) (b)

$$L_{UNMT} = L_{Denoising} + L_{Back\text{-}Translation}$$

Joint UBWE and UNMT Training

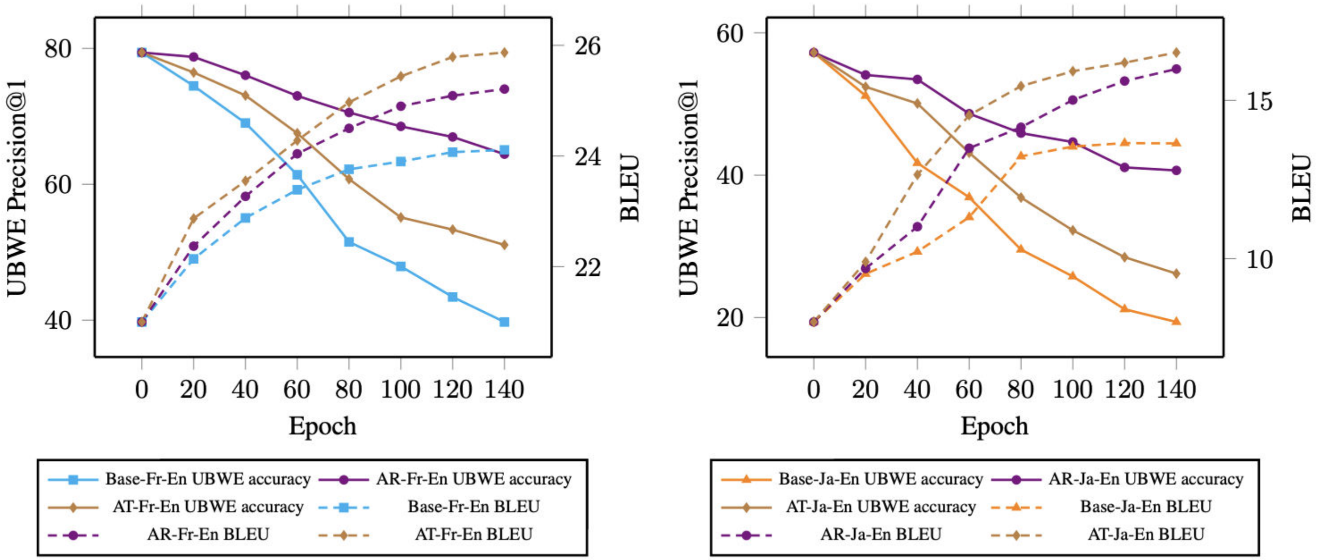
- Our contribution
  - We propose a joint UBWE and UNMT training method.





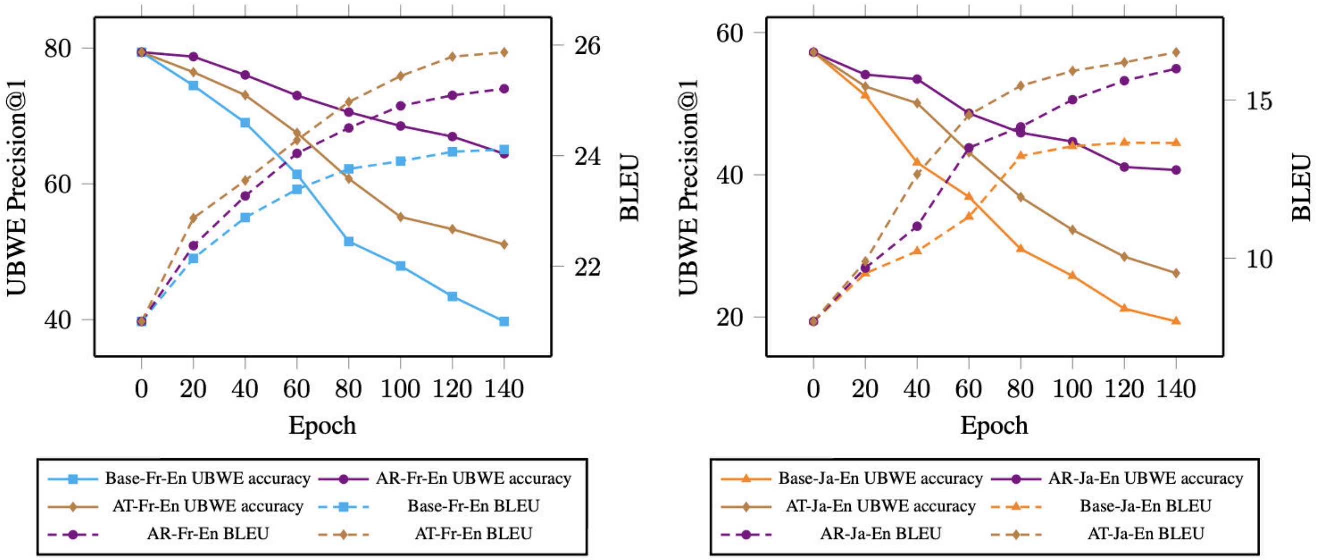
$$L_{UNMT} = L_{Denoising} + L_{Back\text{-}Translation} + L_{Agreement}$$

Performance: Unsupervised Translation



(a) Fr-En (b) Ja-En

Performance: Unsupervised Translation



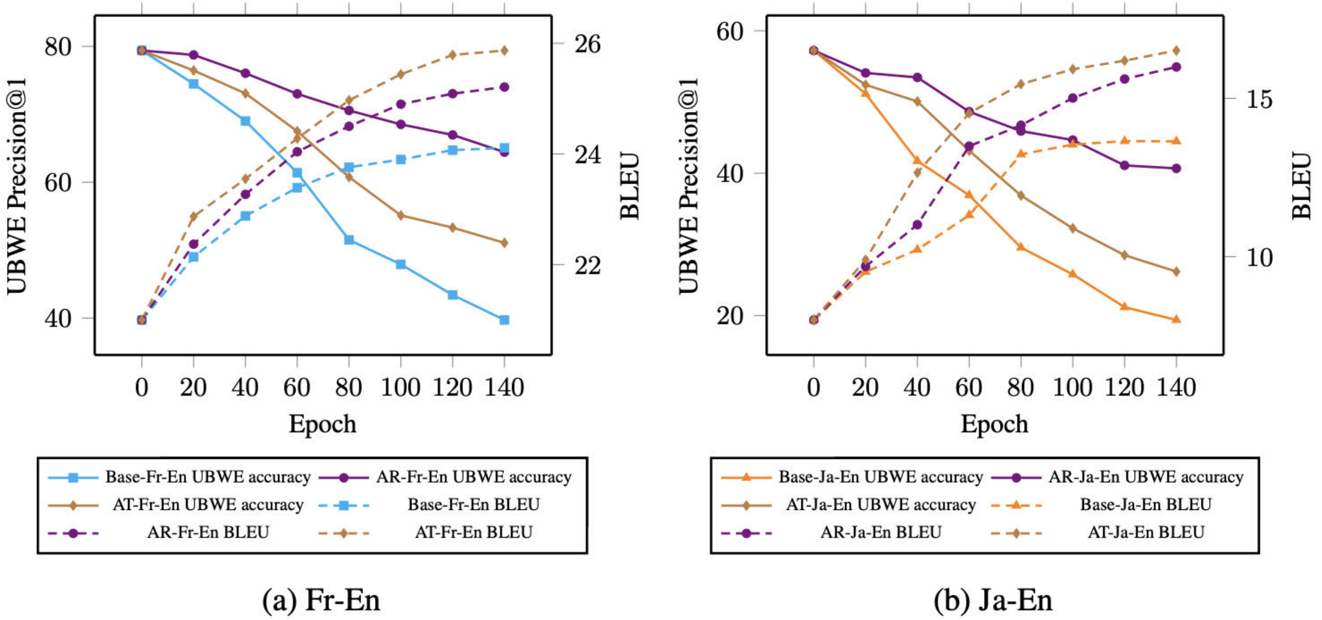
(a) Fr-En

(b) Ja-En

Method	Fr-En	En-Fr	De-En	En-De	Ja-En	En-Ja
Artetxe *et al.* [16]	15.56	15.13	n/a	n/a	n/a	n/a
Lample *et al.* [17]	14.31	15.05	13.33	9.64	n/a	n/a
Yang *et al.* [36]	15.58	16.97	14.62	10.86	n/a	n/a
Lample *et al.* [19]	24.20	25.10	21.00	17.20	n/a	n/a
UNMT-BWE Baseline	24.50	25.37	21.23	17.06	14.09	21.63
+ UBWE agreement regularization	25.21++	27.86++	22.38++	18.04++	16.36++	23.01++
+ UBWE adversarial training	**25.87++**	**28.38++**	**22.67++**	**18.29++**	**17.22++**	**23.64++**

(Sun and **Wang*** et al. ACL-2019)

Performance: Unsupervised Translation



Distant language pair

Method	Fr-En	En-Fr	De-En	En-De	Ja-En	En-Ja
Artetxe *et al.* [16]	15.56	15.13	n/a	n/a	n/a	n/a
Lample *et al.* [17]	14.31	15.05	13.33	9.64	n/a	n/a
Yang *et al.* [36]	15.58	16.97	14.62	10.86	n/a	n/a
Lample *et al.* [19]	24.20	25.10	21.00	17.20	n/a	n/a
UNMT-BWE Baseline	24.50	25.37	21.23	17.06	14.09	21.63
+ UBWE agreement regularization	25.21++	27.86++	22.38++	18.04++	16.36++	23.01++
+ UBWE adversarial training	**25.87++**	**28.38++**	**22.67++**	**18.29++**	**17.22++**	**23.64++**

(Sun and **Wang*** et al. ACL-2019)

What Is the Performance Now?

ACL 2019
FOURTH CONFERENCE ON
MACHINE TRANSLATION (WMT19)

August 1-2, 2019
Florence, Italy

Shared Task: Machine Translation of News

[HOME] [SCHEDULE] [PAPERS] [RESULTS]
TRANSLATION TASKS: [NEWS] [BIOMEDICAL] [ROBUSTNESS] [SIMILAR]
EVALUATION TASKS: [METRICS] [QUALITY ESTIMATION]
OTHER TASKS: [AUTOMATIC POST-EDITING] [PARALLEL CORPUS FILTERING]

Account

System	Submitter	System Notes	Constraint	Run Notes	BLEU	BLEU-cased	TER	BEER 2.0	CharactTER
NICT (Details)	Nedved NICT	Pre-trained cross-lingual LM + UNMT + USMT + pseudo SMT + pseudo NMT + fine-tuning + ensemble + USMT reranking + fixed quotes	yes		20.5	**20.1**	0.726	0.519	0.624
NICT (Details)	Nedvěd NICT	repeated submission due to web lag	yes		20.5	20.1	0.726	0.519	0.624
NEU&KingSoft (Details)	NiuTrans *Northeastern University*	Pre-training of a cross-lingual language model + Unsupervised SMT startup + Ensemble of 2 Transformer-big models + iterative back-translation + denoising auto-encoding + fix quotes	yes		19.2	18.9	0.731	0.509	0.633
Unsupervised.de-cs (Details)	StillKeepTry *Nanjing University of Science and Technology*	+ fix quotes, + iterative back-translation, + Unsupervised SMT data fine-tuning, + fix quotes, + beam10,	yes	Ensemble 2 model, + Rerank, + fine tune more weight-domain data in source side,	18.0	17.8	0.752	0.486	0.670
lmu-unsup-nmt-de-cs (Details)	dario *LMU Munich*	Cross-lingual LM pretraining + unsupervised NMT with denoising auto-encoding and on-the-fly backtranslation + fine-tuned with unsupervised SMT backtranslated data	yes	fixed quotes	17.4	17.0	0.754	0.488	0.758
NICT (Details)	Nedvěd NICT	repeated submission due to web lag	yes		16.9	16.5	0.763	0.494	0.655
Unsupervised.de-cs (Details)	StillKeepTry *Nanjing University of Science and Technology*	+ fix quotes, + iterative back-translation, + Unsupervised SMT data fine-tuning, + fix quotes, + beam10,	yes	Single Model	16.3	16.1	0.771	0.475	0.686
NICT (Details)	Nedved NICT	single UNMT model	yes		15.9	15.5	0.774	0.482	0.673
CUNI-Unsupervised (Details)	kvapili *Charles University*	Unsupervised phrased based model + Iterative back translation + NMT trained on synthetic parallel data with reordering (Transformer)	yes		15.3	15.0	0.784	0.489	0.672
CUNI-Unsupervised-combined (Details)	kvapili *Charles University*	Sentences with named entities tranlated by CUNI-Unsupervised-NER, sentences without named entities tranlated by CUNI-Unsupervised	yes		14.9	14.6	0.785	0.488	0.674

What Is the Performance Now?

- Our system is the best in WMT-2019 and WMT-2020, the most important MT shared task in the world.
- Our system is comparable to the online commercial systems (in gray) which (may) uses the parallel data.

German→Czech

Ave.	Ave. z	System
63.9	0.426	online-Y
62.7	0.386	online-B
61.4	0.367	NICT
59.8	0.319	online-G
55.7	0.179	NEU-KingSoft
54.4	0.134	online-A
47.8	−0.099	lmu-unsup-nmt
46.6	−0.165	CUNI-Unsupervised-NER-post
41.7	−0.328	Unsupervised-6929
39.1	−0.405	Unsupervised-6935
28.4	−0.807	CAiRE

[Benjamin and **Wang*** et al. WMT-2019]

Very Low Resource Supervised MT

- If we added some parallel data to UNMT.

System Name	DE-HSB	HSB-DE	citation
SJTU-NICT	60.7	58.5	(Li et al., 2020)
Helsinki-NLP	57.9	59.6	(Scherrer et al., 2020)
NRC-CNRC	57.3	58.9	(Knowles et al., 2020)
LMU-supervised-ensemble	56.5	57.6	(Libovický et al., 2020)
CUNI-Transfer	55.5	56.9	(Kvapilíková et al., 2020)
Brown-NLP-b	46.2	45.7	(Berckmann and Hiziroglu, 2020)
IITBHU-NLPRL-DE-HSB	45.9	47.9	(Baruah et al., 2020)
Adobe-AMPS	45.2	47.6	(Singh, 2020)
UdS-DFKI	40.9		(Dutta et al., 2020)
HierarchicalTransformer	38.2	40.1	

Table 4: Ten primary systems submitted to the Very Low Resource Task, sorted by DE-HSB BLEU score.

[Li, Zhao, and Wang et al. WMT-2020]

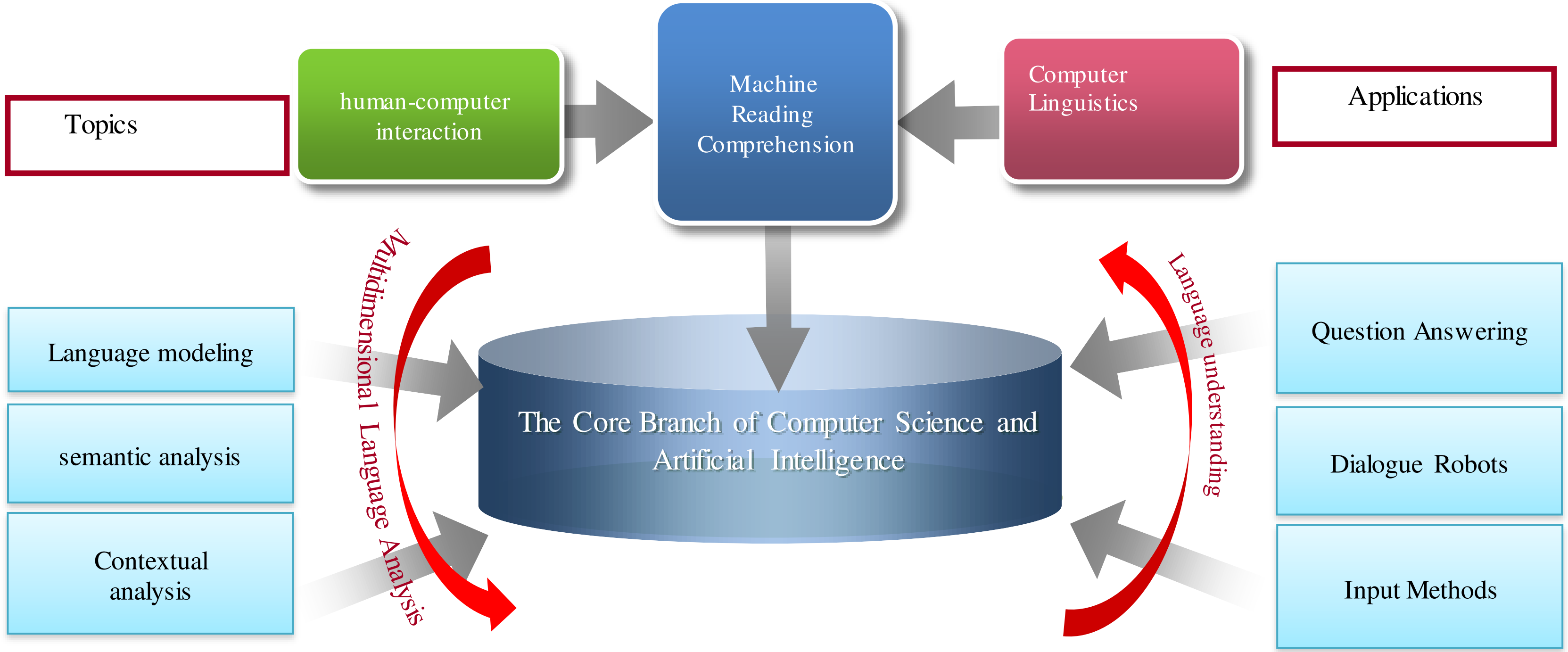
Menu

Outline

- ✓ The Evolution of Pre-trained Language Model
  - Motivation
  - The Path to Pre-trained Language Model
- To Learn Pre-trained Language Model
  - Encoder Design
  - Training Objective： A Unified Perspective of ADE
    - Discriminative
    - Generative
    - Both
  - Tokenization and Masking Unit
- Application of Pre-trained Language Models
  - Multi-task Learning: LIMIT-BERT
  - Extension of Pre-training and Fine-tuning Framework

Machine Reading Comprehension

- MRC: Give the accurate **answer** for a **question** according to a **passage**.

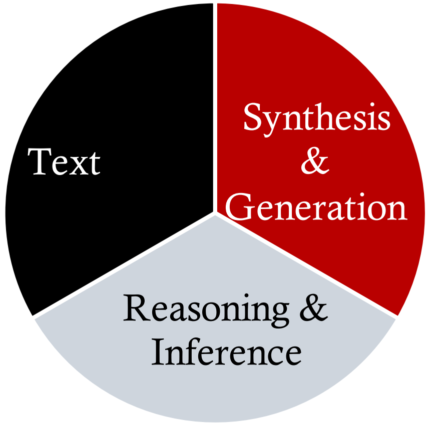


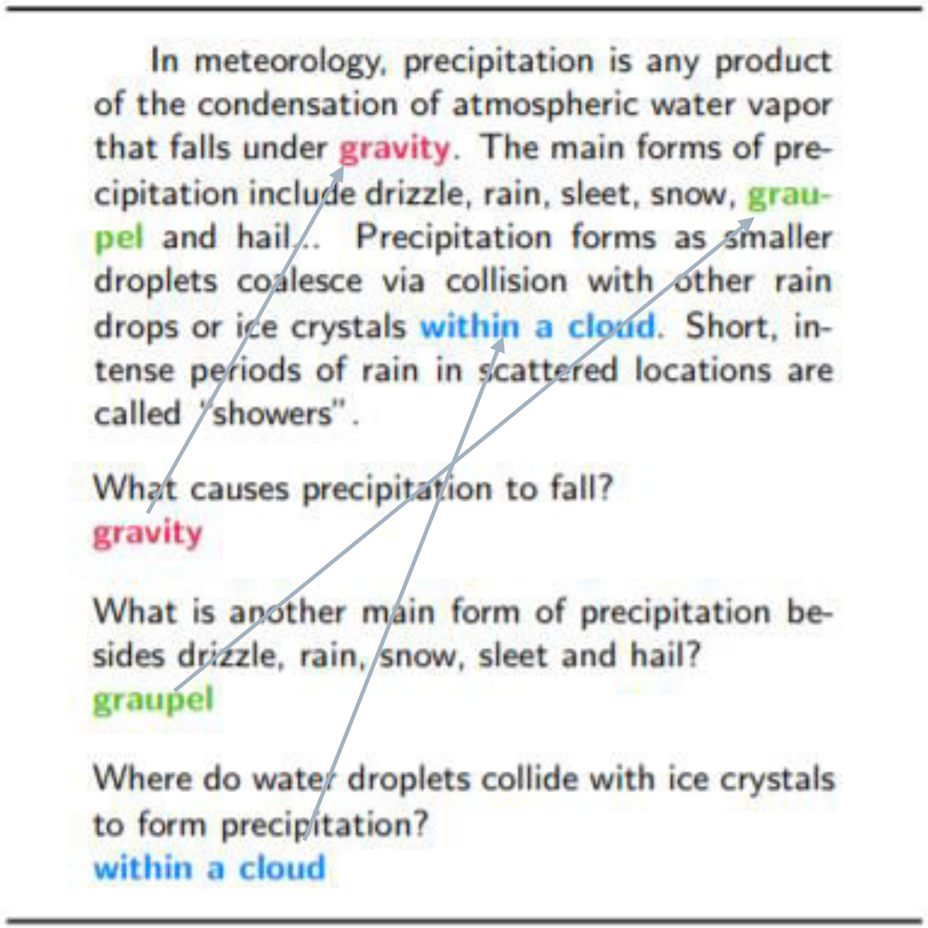
- Types
  - Cloze-style
  - Multi-choice
  - Span-based

> • MRC Survey ：
> Zhuosheng Zhang, Hai Zhao, Rui Wang.2020. Machine Reading Comprehension: The Role of Contextualized Language Models and Beyond. arXiv:2005.06249.

MRC Task: Extractive

- Extractive MRC： SQuAD
  - given passage and question, find the accurate answer
  - Answer – a span of the passage





In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

(Sentence/Contextual) Encoder as a Standard Network Block

- Word embeddings have changed NLP
- However, sentence is the least unit that delivers complete meaning as human use language
- Deep learning for NLP quickly found it is a frequent requirement on using a network component encoding a sentence input.
  - So that we have the Encoder for encoding the complete sentence-level Context
- Encoder differs from sliding window input that it covers a full sentence.
- It especially matters when we have to handle passages in MRC tasks, where passage always consists of a lot of sentences (not words).
  - When the model faces passages, sentence becomes the basic unit
  - Usually building blocks for an encoder: RNN, especially LSTM

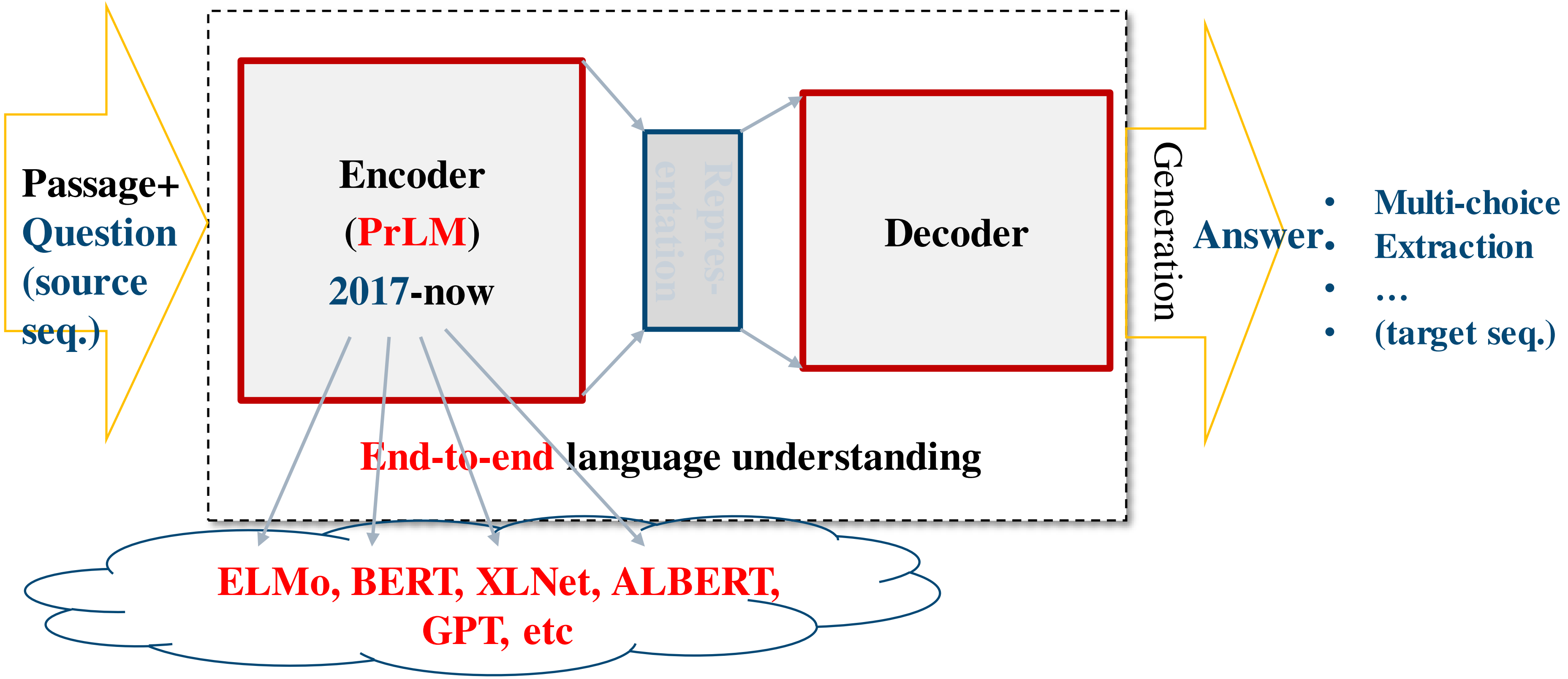
NLP and NLU Modeling

NLU = MRC + NLI

➔ NLU ≈ MRC (Passage + Question ➔ Answer)

➔ NLP (MT) ≈ source language ➔ target sentence

Pre-trained Language Model (PrLM) ⟺ Language Representation



From Language Models to Language Representation

- MRC and other application NLP need a full sentence encoder,
  - Deep contextual information is required in MRC
  - Word and sentence should be represented as embeddings.
- Model can be trained in a style of *n*-gram language model
- So that there comes the language representation (or, pre-trained contextualized language model) which includes
  - *n*-gram language model (training object), plus
  - Embedding (representation form), plus
  - Contextual encoder (model architecture)

→ The representation for each word depends on the entire context in which it is used, **dynamic embedding**.

	Repr. form	Context	Training obj.
n-gram LM	One-hot	Sliding-window	*n*-gram LM(MLE)
Word2vec/GloVe...	Embedding		
Contextualized LR (LM)		sentence	*n*-gram LM(MLE) & extension

PrLM：Terms

- Pre-trained Models ╳
  - Hard to distinguish non-language models
- Pre-trained Language Models √
  - Hard to distinguish non contextualized methods, such as word2vec/GloVe
- Pre-trained Language Representation Models √
- Pre-trained Contextualized Language Models √ √
- Pre-trained Contextualized Language Representation Models √ √
- (pre-trained contextualized language representation model)

Working mode

Essential characteristics of language model

Embedding form

Outline

- The Evolution of Pre-trained Language Model
  - Motivation
  - ✓ The Path to Pre-trained Language Model
- To Learn Pre-trained Language Model
  - Encoder Design
  - Training Objective： A Unified Perspective of ADE
    - Discriminative
    - Generative
    - Both
  - Tokenization and Masking Unit
- Application of Pre-trained Language Models
  - Multi-task Learning: LIMIT-BERT
  - Extension of Pre-training and Fine-tuning Framework

n-gram Language Model (LM)

- An *n*-gram Language model is a probability distribution over word (*n*-gram) sequences
  - P("*And nothing but the truth*") ≈ 0.001
  - P("*And nuts sing on the roof*") ≈ 0
- How to compute P("And nothing but the truth"):
  - Decompose probability

  P("*And nothing but the truth*") = P("*And*") × P("*nothing*|*and*") × P("*but*|*and nothing*") × P("*the*|*and nothing but*") × P("*truth*|*and nothing but the*")

  - Estimate probabilities: Get real text, and start counting!

  P("*the* | *nothing but*") ≈ C("*nothing but the*") / C("*nothing but*")

- *n*-gram LM can be regarded with a training objective of predicting *uni*gram from (*n*-1)-gram
  - Called autoregressive
- *n*-gram LM is with one-hot representation.

Neural Language Model

- **Neural networks** use continuous representations or **embeddings** of words to make their predictions.
- Alleviate the **curse of dimensionality**: as language models are trained on larger and larger texts, the number of unique words increases.
- Learn a **probability distribution**: $P(W_t|\text{context})\ \forall t \in V$
- The context might be a fixed-size window of previous words, so

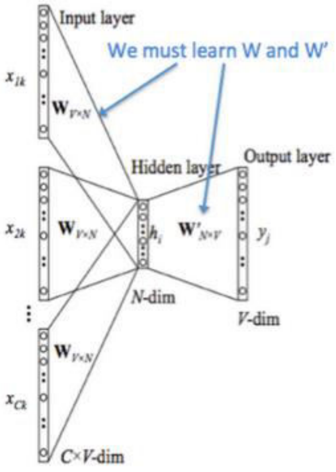
$$P(W_t|\text{context}) = P(W_t|W_{t-k}, \ldots, W_{t-1})$$

- To train a model, minimize the negative log-probability (MLE, the same training objective as n-gram LM):

$$-\sum \log P(W_{t+j}|W_t)\ \text{as objective function.}$$

NNLM, word2vec, GloVe …

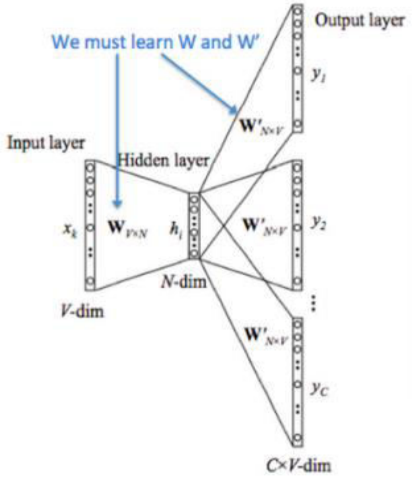
Distributional representations - Word2Vec



CBOW: "***predicting the word given its context***"

- generate one hot vectors $(x^{(c-m)}, \cdots, x^{(c-1)}, x^{(c+1)}, \cdots, x^{(c+m)})$ for the input context of size m
- get the embedded word vectors for the context $(v_i = \mathcal{V}x^{(i)})$
- average these vectors to get $\hat{v} = \frac{v_{c-m}+v_{c-m+1}+\cdots+v_{c+m}}{2m}$
- generate a score vector $z = \mathcal{U}\hat{v}$
- turn the scores into probabilities: $\hat{y} = softmax(z)$
- we desire the generated probabilities $\hat{y}$ to match the true probabilities y, which happens to be the one hot vector of the actual word

Advantage: several times faster to train than the skip-gram, slightly better accuracy for the frequent words



Skip-gram: "***predicting the context given a word***"

- generate one hot input vector x
- get embedded word vectors for the context $v_c = \mathcal{V}x$
- not averaging, just set $\hat{v} = v_c$
- generate $2m$ score vectors: $u_{c-m}, \cdots, u_{c-1}, u_{c+1}, \cdots, u_{c+m}$ using $u = \mathcal{U}v_c$
- turn each of the scores into probabilities: $y = softmax(u)$
- we desire the generated probability vector to match the true probabilities which is $y^{(c-m)}, \cdots, y^{(c-1)}, y^{(c+1)}, \cdots, y^{(c+m)}$, the one hot vectors of the actual output

Advantage: works well with small amount of the training data, represents well even rare words or phrases

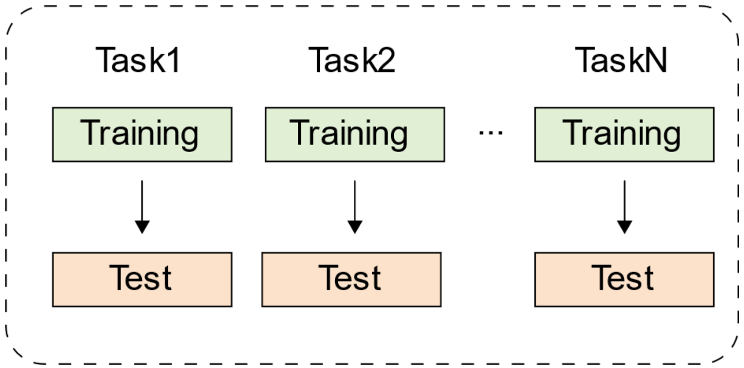
PrLM and MRC

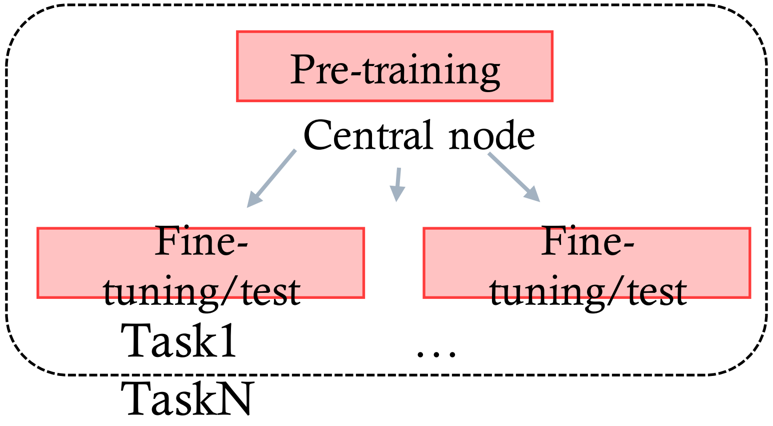
	NLI		MRC			
	SNLI	GLUE	SQuAD1.1	SQuAD2.0	RACE	CoQA
ELMo	√√		√√			
GPT						√
BERT		√√	√√	√√		
RoBERTa		√√	√√	√	√	
XLNet		√√	√√	√√	√√	
ALBERT		√√	√√	√√	√√	

Complementarily Developing for

- MRC Boosts the development of language models
- Pre-trained Language Models stimulates MRC

Pre-trained Language Model: New Paradigm in Machine Learning





Past

Develop the individual model for each task and finish both the training and test.

Now

The central node completes the large-scale pre-training of the general language model. Other users borrow the existing pre-trained model as the standard module for further fine-tuning.

Individual training ➡ Centralized pre-training + individual fine-tuning

Extreme case: gpt3 directly makes generation prediction after pre-training, eliminating fine-tuning

Outline

- The Evolution of Pre-trained Language Model
  - Motivation
  - The Path to Pre-trained Language Model
- To Learn Pre-trained Language Model
  - Encoder Design
  - Training Objective： A Unified Perspective of ADE
    - Discriminative
    - Generative
    - Both
  - Tokenization and Masking Unit
- Application of Pre-trained Language Models
  - Multi-task Learning: LIMIT-BERT
  - Extension of Pre-training and Fine-tuning Framework

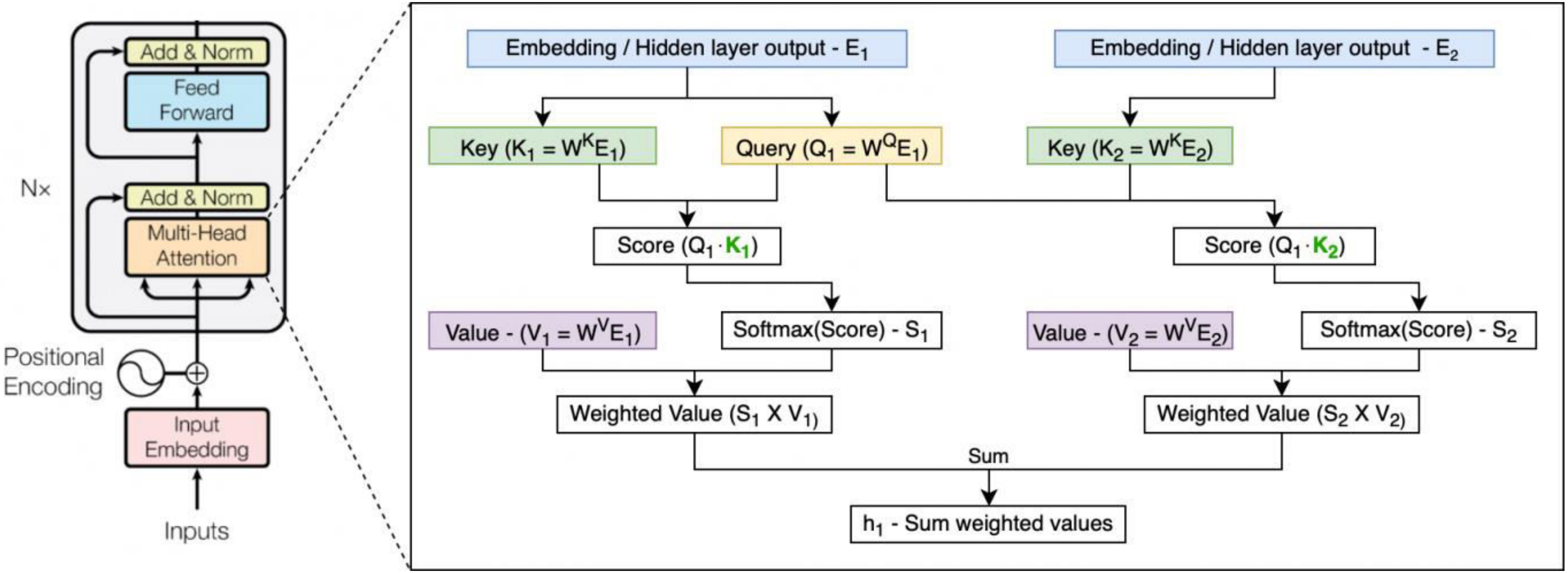
Elements of PrLMs

- Encoder Architecture
  - RNN/Transformer/…
- Training Objective
  - (Autoregressive / denoising) tasks
- Sampling (training) methods

Encoder Architecture for PrLMs

- RNN (LSTM)
  - Capture the dependency between words. However, RNN is often difficult to train because of gradient computation and low computing speed.
  - The ability of learning long dependency is limited (experience shows that LSTM can only model 200 context words on average).
- Transformer √
  - Apply self-attention mechanism (SAN) for global processing.
  - Learn Three weight matrixes (query, key and value) at one time to capture the dependency between the parts of the input sequence.
  - Multi layer network: each layer is composed of multi attention mechanism and feedforward network.
  - SAN can not directly capture the important position information in the sequence, so it adds position encoding to the input, and uses sine function to generate position vector for each position.
- Transformer-XL √ from two improvements on :
  - Recurrence Mechanism
  - Relative Positional Encoding

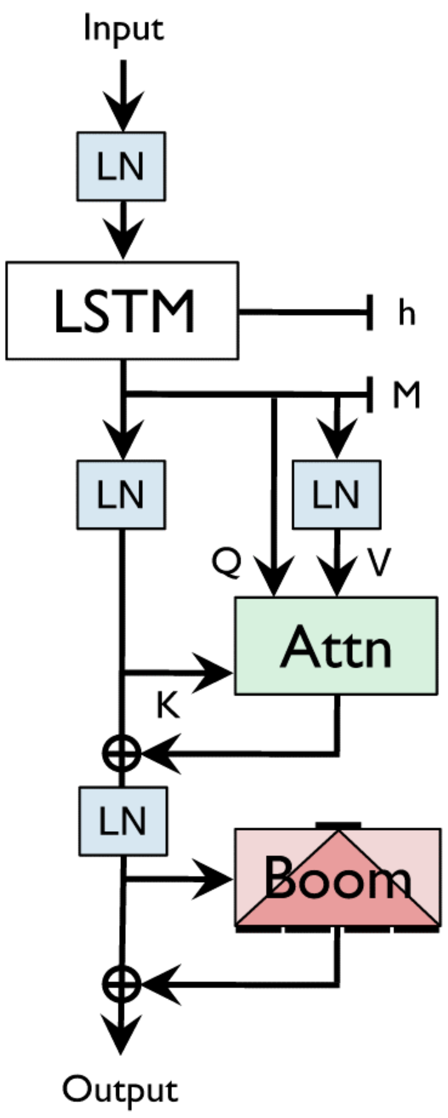
Transformer



SHA-LSTM

- Stephen Merity.2019. Single Headed Attention RNN: Stop Thinking With Your Head. https://arxiv.org/pdf/1911.11423
- https://github.com/smerity/

- It simplifies LSTM architecture and makes it more efficient
- The 24-hour training of single GPU achieves comparable BPC performance to that of transformer on envik8



Outline

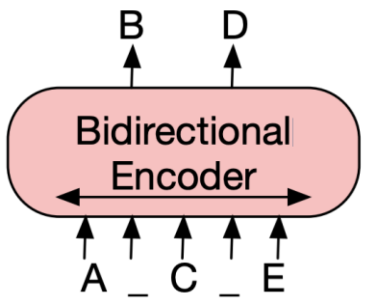
- The Evolution of Pre-trained Language Model
  - Motivation
  - The Path to Pre-trained Language Model
- To Learn Pre-trained Language Model
  - Encoder Design
  - Training Objective： A Unified Perspective of ADE
    - Discriminative
    - Generative
    - Both
  - Tokenization and Masking Unit
- Application of Pre-trained Language Models
  - Multi-task Learning: LIMIT-BERT
  - Extension of Pre-training and Fine-tuning Framework

Training Objectives

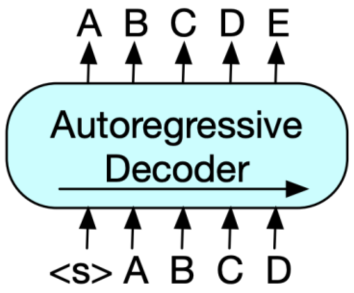
- Language model is the largest machine learning task ever
- Where does the training corpus come from?
  - The number of unmarked natural languages is almost unlimited;
  - Automatic construction / natural tagging in natural language;
  - ➔The biggest machine learning task
- The PrLM is an automatic denoising encoder

Training Objectives

- Two ways to be autoregressive:
- Discriminative vs. Generative
  - Discriminative: restore corrupted language on Encoder
  - Generative: predict completed language on decoder



(a) Discriminative



(b) Generative

Training Objectives (Ennoising) Unified PrLM

- Artificially changing different level units of natural language text
- ➔ Edit distance operation
  - delete
  - add
  - Exchange
  - replace

	word level	sentence level
delete	Masking	NSP
replace		
add		
Exchang e	XLNet ?	SOP

- Two levels of language units :
  - Word level
  - Sentence level
- Total 4 ╳ 2 ╳ 2 = 16 specific training objectives

- two types of training objectives :
  - Direct prediction (others)
  - Discriminant prediction (ELECTRA)

Outline

- The Evolution of Pre-trained Language Model
  - Motivation
  - The Path to Pre-trained Language Model
- To Learn Pre-trained Language Model
  - Encoder Design
  - Training Objective： A Unified Perspective of ADE
    - Discriminative
    - Generative
    - Both
  - Tokenization and Masking Unit
- Application of Pre-trained Language Models
  - Multi-task Learning: LIMIT-BERT
  - Extension of Pre-training and Fine-tuning Framework

BERT Training Objectives

- Task #1: Masked LM
  - replace the chosen words with [MASK] then predict it
  - Not always replace the word with [MASK]
- Task #2: Next Sentence Prediction
  - [CLS] sentence A [SEP] sentence B [SEP]
  - 50% of the time B is the actual next sentence that follows A, and 50% of the time it is a random sentence from the corpus

BERT - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT 2019.

- 80% of the time: Replace the word with the [MASK] token, e.g., `my dog is hairy` → `my dog is [MASK]`
- 10% of the time: Replace the word with a random word, e.g., `my dog is hairy` → `my dog is apple`
- 10% of the time: Keep the word unchanged, e.g., `my dog is hairy` → `my dog is hairy`. The purpose of this is to bias the representation towards the actual observed word.

Input = `[CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]`

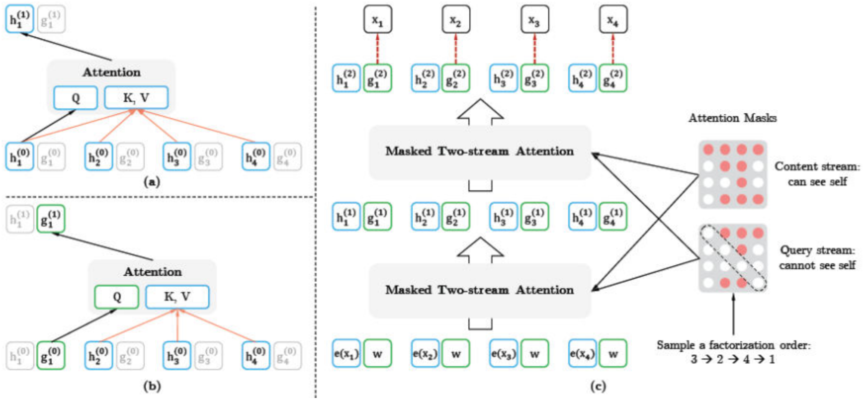
Label = `IsNext`

Input = `[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds [SEP]`

Label = `NotNext`

XLNet Training Objectives： Word Permutation

- Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding, NeurIPS 2019.
- Objective: maximize the factorization order of the permutation language model, bi-direction training
  - Using autoregressive mechanism to overcome the shortcomings of masked LM
  - In the sentence, words are rearranged and reordered, and then further language model prediction is made



Architecture: Two-Stream Self-Attention for target representation

- Computation：512 TPU v3， 500K steps, batch size = 2048, 2.5 days

Training corpus：
- 13G: BooksCorpus + English Wikipedia
- 16G: Giga5
- 19G: ClueWeb 2012-B
- 78G: Common Crawl

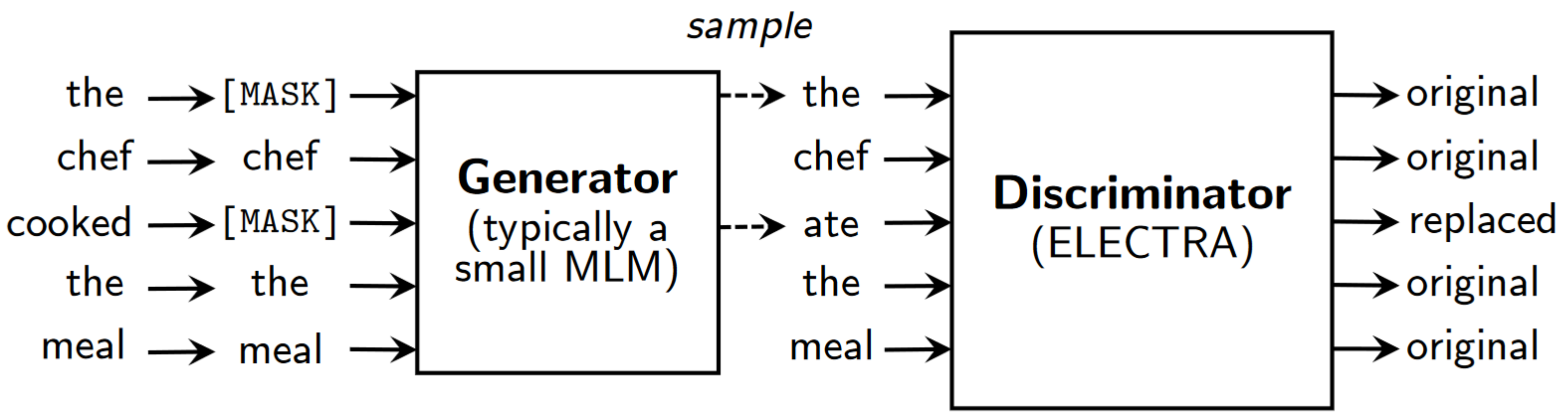
ALBERT Training Objectives: Sentence Permutation

- Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut. ALBERT: A Lite BERT for Self-supervised Learning of Language Representation. *ICLR* 2020.

- Three improvements compared with the original BERT :
  - Adjust the dimension of input embedding (E) and hidden layer vector (H) to H > > E instead of E = H of original BERT
  - Use parameter sharing among the intermediate layers, including all forward networks and attention weights (greatly reducing model size)
  - Modify the sentence training objective (next sentence prediction) of BERT to sentence order prediction

Discrimination Rather than Direct Prediction: ELECTRA

- Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. *ICLR* 2020.



- Adversarial generative training (GAN)
- Replaced token detection

ELECTRA Performance

Model	Train FLOPs	CoLA	SST	MRPC	STS	QQP	MNLI	QNLI	RTE	WNLI	Avg.*	Score
BERT	1.9e20 (0.06x)	60.5	94.9	85.4	86.5	89.3	86.7	92.7	70.1	65.1	79.8	80.5
RoBERTa	3.2e21 (1.02x)	67.8	96.7	89.8	91.9	90.2	90.8	95.4	88.2	89.0	88.1	88.1
ALBERT	3.1e22 (10x)	69.1	**97.1**	**91.2**	92.0	90.5	**91.3**	–	89.2	91.8	89.0	–
XLNet	3.9e21 (1.26x)	70.2	**97.1**	90.5	**92.6**	90.4	90.9	–	88.5	**92.5**	89.1	–
ELECTRA	3.1e21 (1x)	**71.7**	**97.1**	90.7	92.5	**90.8**	**91.3**	**95.8**	**89.8**	**92.5**	**89.5**	**89.4**

GLUE

Model	Train FLOPs	Params	SQuAD 1.1 dev		SQuAD 2.0 dev		SQuAD 2.0 test	
			EM	F1	EM	F1	EM	F1
BERT-Base	6.4e19 (0.09x)	110M	80.8	88.5	–	–	–	–
BERT	1.9e20 (0.27x)	335M	84.1	90.9	79.0	81.8	80.0	83.0
SpanBERT	7.1e20 (1x)	335M	88.8	94.6	85.7	88.7	85.7	88.7
XLNet-Base	6.6e19 (0.09x)	117M	81.3	–	78.5	–	–	–
XLNet	3.9e21 (5.4x)	360M	**89.7**	**95.1**	87.9	**90.6**	87.9	90.7
RoBERTa-100K	6.4e20 (0.90x)	356M	–	94.0	–	87.7	–	–
RoBERTa-500K	3.2e21 (4.5x)	356M	88.9	94.6	86.5	89.4	86.8	89.8
ALBERT	3.1e22 (44x)	235M	89.3	94.8	87.4	90.2	88.1	90.9
BERT (ours)	7.1e20 (1x)	335M	88.0	93.7	84.7	87.5	–	–
ELECTRA-Base	6.4e19 (0.09x)	110M	84.5	90.8	80.5	83.3	–	–
ELECTRA-400K	7.1e20 (1x)	335M	88.7	94.2	86.9	89.6	–	–
ELECTRA-1.75M	3.1e21 (4.4x)	335M	**89.7**	94.9	**88.0**	**90.6**	**88.7**	**91.4**

MRC

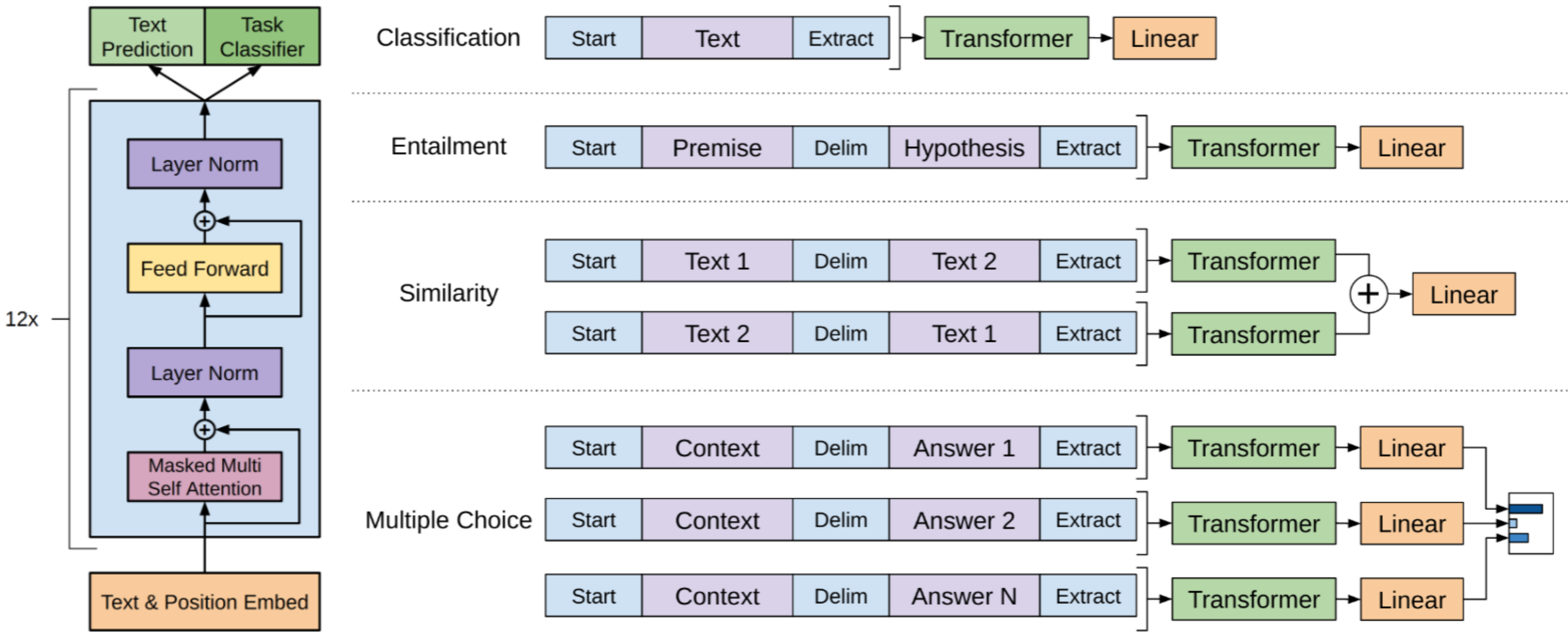
Outline

- The Evolution of Pre-trained Language Model
  - Motivation
  - The Path to Pre-trained Language Model
- To Learn Pre-trained Language Model
  - Encoder Design
  - Training Objective： A Unified Perspective of ADE
    - Discriminative
    - Generative
    - Both
  - Tokenization and Masking Unit
- Application of Pre-trained Language Models
  - Multi-task Learning: LIMIT-BERT
  - Extension of Pre-training and Fine-tuning Framework

GPT-1

GPT-1: Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever. 2018.
Improving Language Understanding by Generative Pre-Training.

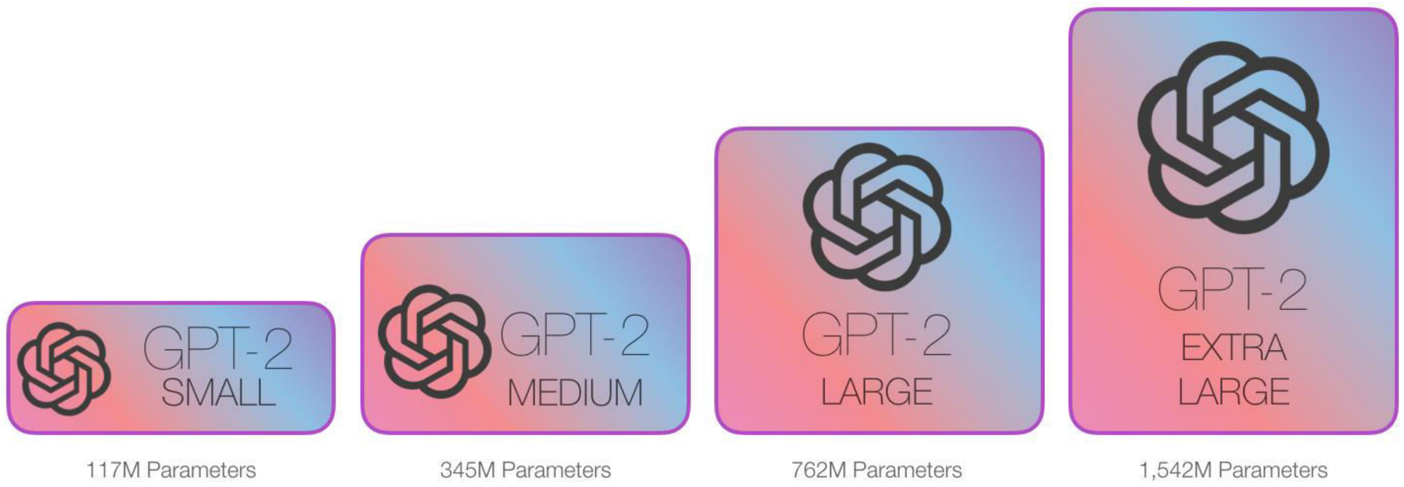
- Generative pre-training of language models on Books Corpus
- Discriminative fine-tuning on specific tasks
- Use Transformer instead of LSTM as Encoder
  - GPT-1 : 12 layers
  - GPT-2 : 48 layers
  - GPT-3 : 96 layers



GPT-2

GPT-2: Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners

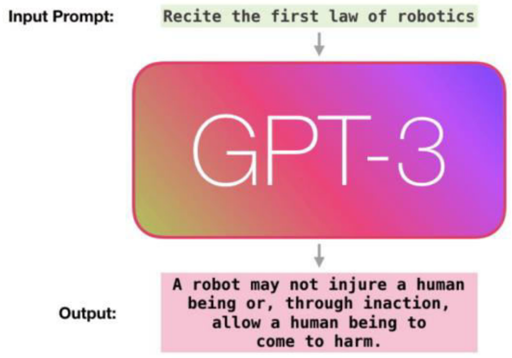
- Follow GPT single directional Transformer
- Abandon fine-tuning process
- More data (8 million webpages， 40G data)
- More parameters（12 layers->48 layers， hidden dimensions1600， about 1.5 billion parameters）



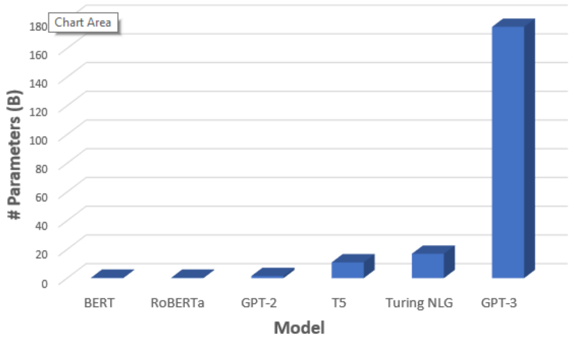
GPT-3

GPT-3: Brown, Tom B., et al. 2019. GPT-3: Language Models are Few-Shot Learners

- Increase parameters to 175 billion
- Use 45TB data
- Solve tasks with less domain data and no fine tuning



Model Name	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{head}	Batch Size	Learning Rate
GPT-3 Small	125M	12	768	12	64	0.5M	6.0×10^{-4}
GPT-3 Medium	350M	24	1024	16	64	0.5M	3.0×10^{-4}
GPT-3 Large	760M	24	1536	16	96	0.5M	2.5×10^{-4}
GPT-3 XL	1.3B	24	2048	24	128	1M	2.0×10^{-4}
GPT-3 2.7B	2.7B	32	2560	32	80	1M	1.6×10^{-4}
GPT-3 6.7B	6.7B	32	4096	32	128	2M	1.2×10^{-4}
GPT-3 13B	13.0B	40	5140	40	128	2M	1.0×10^{-4}
GPT-3 175B or "GPT-3"	175.0B	96	12288	96	128	3.2M	0.6×10^{-4}



Neglected Training Objective: <u>**Adding**</u>

- Effective negative sampling
  - My dog is hairy.
  - → My dog is Trump hairy.
- Invalid negative sampling (positive example)
  - My dog is hairy
  - → My dog is too hairy.

Positive example dilemma of noising sampling?

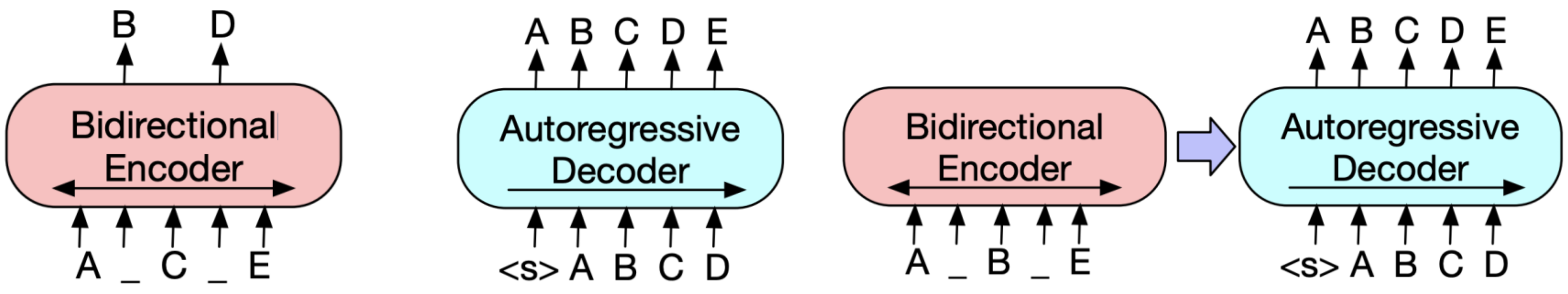
	word level	sentence level
delete	Masking	NSP
replace		
add	?	?
exchange	XLNet ?	SOP

Outline

- The Evolution of Pre-trained Language Model
  - Motivation
  - The Path to Pre-trained Language Model
- To Learn Pre-trained Language Model
  - Encoder Design
  - Training Objective： A Unified Perspective of ADE
    - Discriminative
    - Generative
    - Both
  - Tokenization and Masking Unit
- Application of Pre-trained Language Models
  - Multi-task Learning: LIMIT-BERT
  - Extension of Pre-training and Fine-tuning Framework

BART: Denoising Sequence-to-Sequence Pre-training

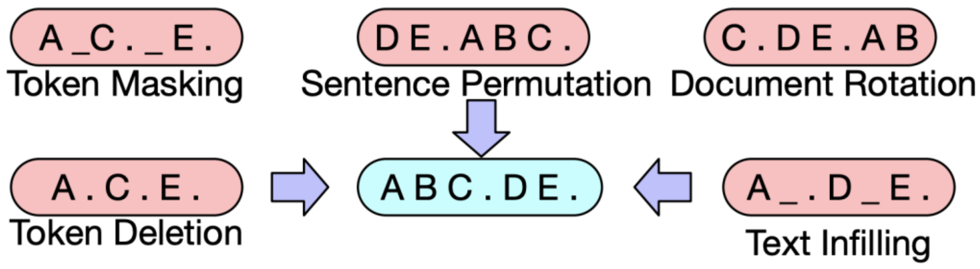
- **Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer . BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. ACL-2020.**
- Architecture: Transformer-based **Encoder-Decoder, both discriminative and generative training**
- Training criteria:
  - Corrupt text with an arbitrary noising function
  - Learn a model to reconstruct the original text.



(a) Discriminative (BERT) (b) Generative (GPT) (c) Discriminative + Generative (BART)

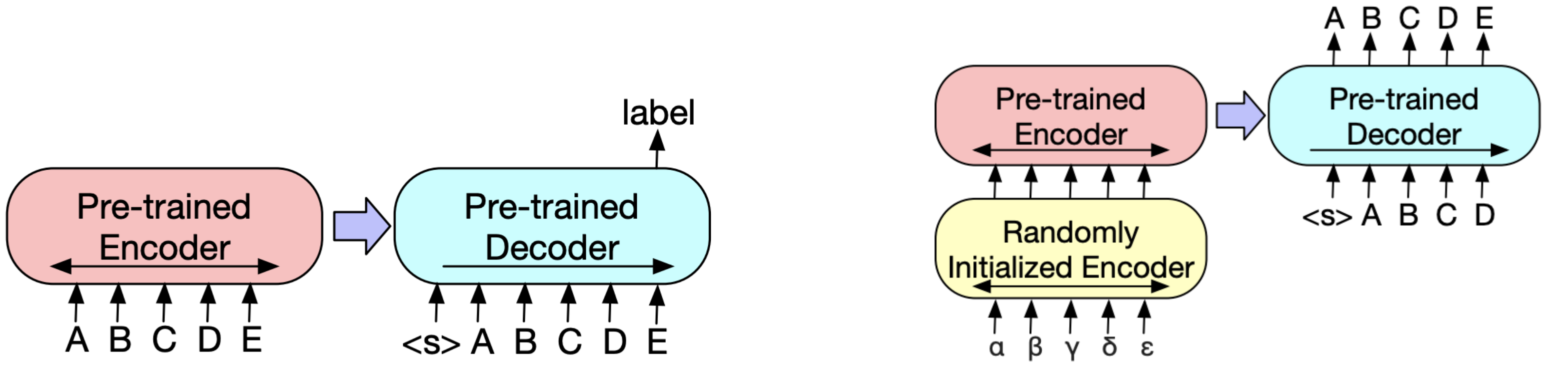
BART: Pre-training

- Corrupting and optimizing a reconstruction loss
  - Token Masking
  - Token Deletion
  - Text Infilling
  - Sentence Permutation
  - Document Rotation



BART: Fine-tuning

- Classification: the same input is fed into the encoder and decoder
- Machine translation: a small additional encoder that replaces the word embeddings in BART
  - Trains the new encoder to map foreign words into an input that BART can de-noise to English.



(a) Classification

(b) Machine Translation

BART Performance: Generation Tasks

- Performance of pre-training methods varies significantly across tasks
- Token masking is crucial
- Left-to-right pre-training improves generation

Model	**SQuAD 1.1** F1	**MNLI** Acc	**ELI5** PPL	**XSum** PPL	**ConvAI2** PPL	**CNN/DM** PPL
BERT Base (Devlin et al., 2019)	88.5	**84.3**	-	-	-	-
Masked Language Model	90.0	83.5	24.77	7.87	12.59	7.06
Masked Seq2seq	87.0	82.1	23.40	6.80	11.43	6.19
Language Model	76.7	80.1	**21.40**	7.00	11.51	6.56
Permuted Language Model	89.1	83.7	24.03	7.69	12.23	6.96
Multitask Masked Language Model	89.2	82.4	23.73	7.50	12.39	6.74
BART Base						
w/ Token Masking	90.4	84.1	25.05	7.08	11.73	6.10
w/ Token Deletion	90.4	84.1	24.61	6.90	11.46	5.87
w/ Text Infilling	**90.8**	84.0	24.26	**6.61**	**11.05**	5.83
w/ Document Rotation	77.2	75.3	53.69	17.14	19.87	10.59
w/ Sentence Shuffling	85.4	81.5	41.87	10.93	16.67	7.89
w/ Text Infilling + Sentence Shuffling	**90.8**	83.8	24.17	6.62	11.12	**5.41**

BART Performance: Discriminative Tasks

- MRC and NLI Tasks

	SQuAD 1.1 EM/F1	SQuAD 2.0 EM/F1	MNLI m/mm	SST Acc	QQP Acc	QNLI Acc	STS-B Acc	RTE Acc	MRPC Acc	CoLA Mcc
BERT	84.1/90.9	79.0/81.8	86.6/-	93.2	91.3	92.3	90.0	70.4	88.0	60.6
UniLM	-/-	80.5/83.4	87.0/85.9	94.5	-	92.7	-	70.9	-	61.1
XLNet	**89.0**/94.5	86.1/88.8	89.8/-	95.6	91.8	93.9	91.8	83.8	89.2	63.6
RoBERTa	88.9/**94.6**	**86.5/89.4**	**90.2/90.2**	96.4	92.2	94.7	**92.4**	86.6	**90.9**	**68.0**
BART	88.8/**94.6**	86.1/89.2	89.9/90.1	**96.6**	**92.5**	**94.9**	91.2	**87.0**	90.4	62.8

- Summarization

	CNN/DailyMail			XSum		
	R1	R2	RL	R1	R2	RL
Lead-3	40.42	17.62	36.67	16.30	1.60	11.95
PTGEN (See et al., 2017)	36.44	15.66	33.42	29.70	9.21	23.24
PTGEN+COV (See et al., 2017)	39.53	17.28	36.38	28.10	8.02	21.72
UniLM	43.33	20.21	40.51	-	-	-
BERTSUMABS (Liu & Lapata, 2019)	41.72	19.39	38.76	38.76	16.33	31.15
BERTSUMEXTABS (Liu & Lapata, 2019)	42.13	19.60	39.18	38.81	16.50	31.27
BART	**44.16**	**21.28**	**40.90**	**45.14**	**22.27**	**37.25**

- Dialogue

	ConvAI2	
	Valid F1	Valid PPL
Seq2Seq + Attention	16.02	35.07
Best System	19.09	17.51
BART	**20.72**	**11.85**

- Translation

	RO-EN
Baseline	36.80
Fixed BART	36.29
Tuned BART	**37.96**

Outline

- The Evolution of Pre-trained Language Model
  - Motivation
  - The Path to Pre-trained Language Model
- To Learn Pre-trained Language Model
  - Encoder Design
  - Training Objective： A Unified Perspective of ADE
    - Discriminative
    - Generative
    - Both
  - Tokenization and Masking Unit
- Application of Pre-trained Language Models
  - Multi-task Learning: LIMIT-BERT
  - Extension of Pre-training and Fine-tuning Framework

Tokenization and Masking Units

- Embedding representation unit
  - character √ ELMo
  - subword √ BERT …
  - word ╳

- Masking unit
  - Subword
  - Span
  - Knowledge item
  - Statistical unit

ELMo

- **ELMo** - **E**mbeddings from **L**anguage **Mo**dels
  - Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. 2018. Deep contextualized word representations. NAACL 2018.
- **Contextual**: The representation for each word depends on the entire context in which it is used, **dynamic embedding**.
- **Big corpus: 1.8 Billion,** 1 Billion Word Benchmark and 800M tokens of news crawl data from WMT 2011.
- **Objective function**: minimize the negative log likelihood:

$$\mathcal{L} = -\sum_{i=1}^{n} \Big(\log p(x_i \mid x_1, \ldots, x_{i-1}; \Theta_e, \overrightarrow{\Theta}_{\mathrm{LSTM}}, \Theta_s) + \log p(x_i \mid x_{i+1}, \ldots, x_n; \Theta_e, \overleftarrow{\Theta}_{\mathrm{LSTM}}, \Theta_s) \Big)$$

ELMo: Performance

- Better word representation: *play*, GloVe vs. biLM

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {. . . }	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {. . . }	{. . . } they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

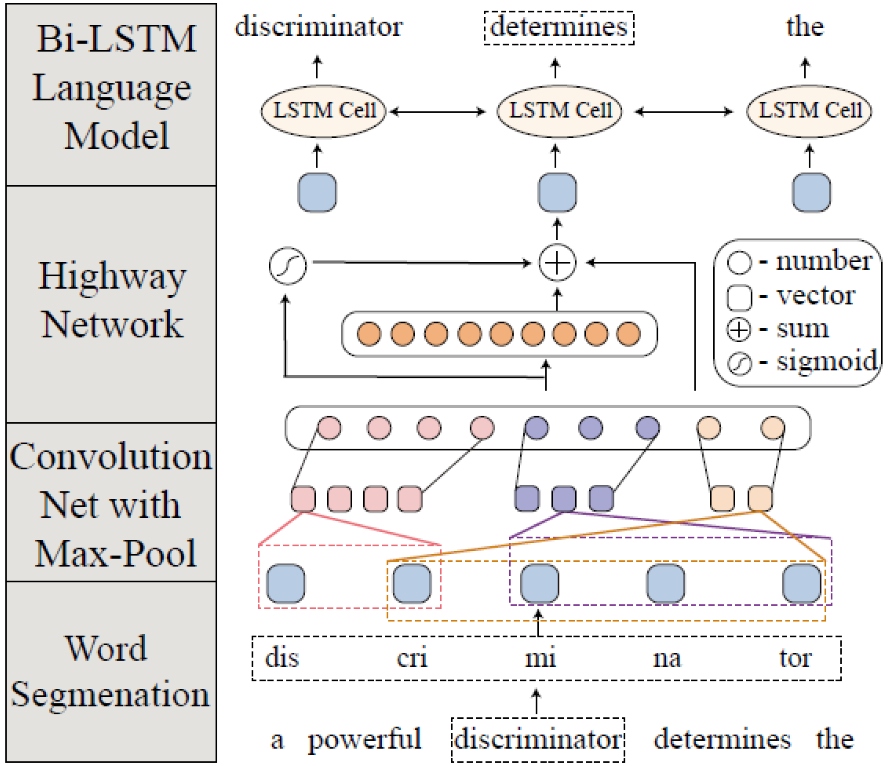
- For downstream tasks (SQuAD1.1)

TASK	PREVIOUS SOTA		OUR BASELINE	ELMO + BASELINE	INCREASE (ABSOLUTE/ RELATIVE)
SQuAD	Liu et al. (2017)	84.4	$\boxed{81.1}$	$\boxed{85.8}$	$\boxed{4.7 / 24.9\%}$
SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

Subword ELMo

- Jiangtong Li, Hai Zhao, Zuchao Li, Wei Bi, Xiaojiang Liu. 2019. Subword ELMo. arXiv:1909.08357.
- ELMo ： character embedding as model input
- SubELMo: takes subword as model input

When changing the subwords in the right Figure into characters, the model becomes ELMo. ➔



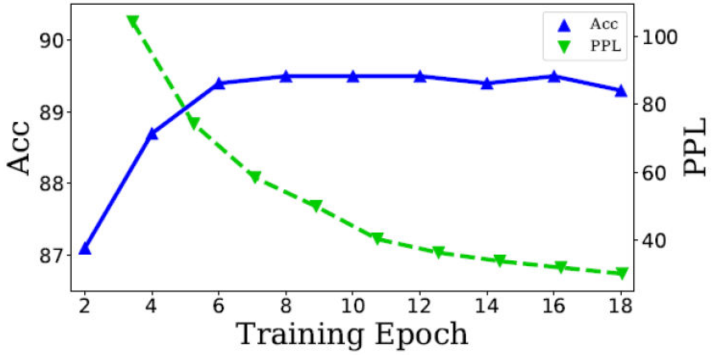
Subword ELMo – Results

Downstream Tasks:

- Syntactic Dependency Parsing (SDP)
- Semantic Role Labeling (SRL)
- Implicit Discourse Relation Recognition (IDRR)
- Textual Entailment (TE)

Tasks		SDP	SRL	IDRR	TE
SOTA(Single Model)		96.35 (2018)	90.4 (2019)	48.22 (2018)	91.1 (2019)
Our Baseline		95.83 (2017)	89.6 (2018)	47.03 (2018)	88.0 (2016)
ELMo	Char(86)	96.45	90.0	48.22	88.7
ESuLMo	BPE(500)	**96.65**$^{+}$	**90.5**$^{+}$	48.99^{+}	**89.5**$^{+}$
	BPE(1000)	96.62^{+}	90.4^{+}	**49.07**$^{+}$	89.4^{+}
	BPE(2000)	96.54	90.4^{+}	48.88^{+}	89.2^{+}
	ULM(500)	96.55	90.2^{+}	48.73^{+}	89.1^{+}
	ULM(1000)	96.51	90.0	48.32	88.9
	ULM(2000)	96.44	90.0	48.35	88.7

Training curve



Model			PPL	#Params
BIG G-LSTM-2		(2017)	36	-
BIG LSTM	Char	(2016)	30.0	1.8B
ELMo[1]	Char	(2018)	29.3(39.9)	1.94B
ESuLMo	Sub	BPE,500	**27.6(40.3)**	1.95B
		BPE,1000	28.1(42.9)	1.96B
		BPE,2000	28.6(44.4)	1.96B
		ULM,500	28.9(43.8)	1.95B
		ULM,1000	30.7(44.1)	1.96B
		ULM,2000	31.5(50.4)	1.96B

Word disambiguation

Model	F1 score
WordNet 1st Sense Baseline	65.9
Raganato et al. (2017)	69.9
Iacobacci et al. (2016)	**70.1**
ELMo	69.0
ESuLMo	69.6

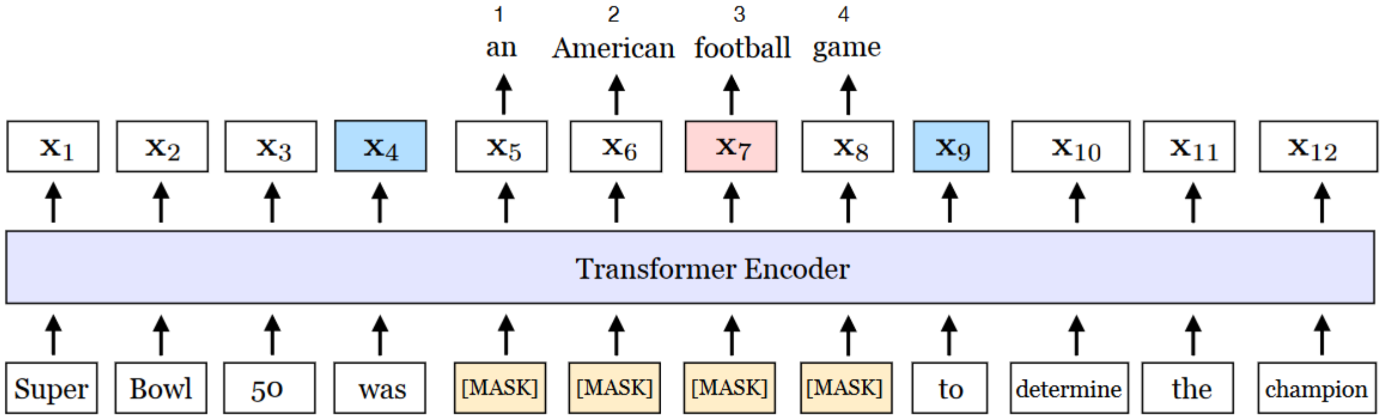
Best number of subwords

BERT_{WWM} and SpanBERT

- BERT_{WWM} : whole word masking

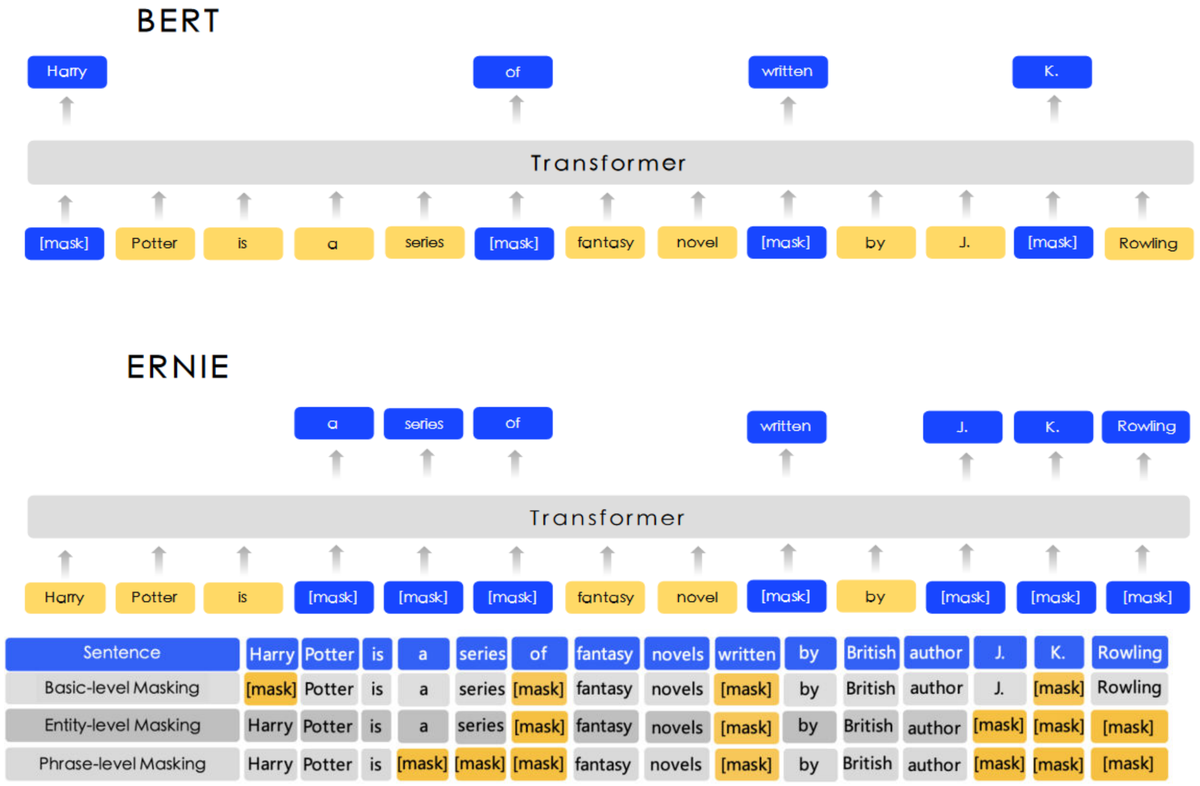
- Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, Omer Levy. 2020. SpanBERT: Improving Pre-training by Representing and Predicting Spans. TACL.
- Random masking continuous text fragments
- span boundary objective

$$\begin{aligned}\mathcal{L}(\text{football}) &= \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football}) \\ &= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)\end{aligned}$$



Masking knowledge item： ERNIE

- Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, Hua Wu. 2019. ERNIE: Enhanced Representation through Knowledge Integration. ACL 2020.
- Enhanced masking： **entity level + phrase level**



Masking Statistically Meaningful Units： BURT

- Yian Li and Hai Zhao. 2020. BURT: BERT-inspired Universal Representation from Learning Meaningful SegmenT, under review of TPAMI-2021

- Construct the same dimension embedded representation for words, sentences and phrases

- All n-gram scores were calculated according to PMI, only high-value n-gram scores were masked.

Comparison of PrLMs

<table>
<tr><th></th><th>Word-level training obj</th><th>Sentence-level training obj</th><th>Training method</th><th>Training direction</th><th>Encoder architecture</th><th>Input from</th></tr>
<tr><td>ELMo</td><td>n-gram LM</td><td></td><td rowspan="5">Predict-ion</td><td>bi-direction</td><td>RNN</td><td>Char</td></tr>
<tr><td>GPT</td><td>n-gram LM</td><td></td><td>uni-direction</td><td rowspan="3">Transformer</td><td rowspan="5">Subword</td></tr>
<tr><td>BERT</td><td rowspan="2">Masked LM</td><td>next sentence</td><td>bi-direction</td></tr>
<tr><td>ALBERT</td><td>sentence order</td><td>bi-direction</td></tr>
<tr><td>XLNet</td><td>permuted n-gram LM</td><td></td><td>bi-direction</td><td>Transformer-XL</td></tr>
<tr><td>Electra</td><td>Masked LM</td><td></td><td>Discrimi-nation</td><td>bi-direction</td><td>Transformer</td></tr>
</table>

- Training corpus size：
  - GPT 3.0, 2.0/XLNet √
- Training direction (uni->bi-directional)：
  - GPT → BERT √
- Sentence-level training objective
  - XLNet × vs. BERT/ALBERT √
- Optimization：RoBERTa/ALBERT √

- Input form：Character vs. subword
  - ELMo vs. BERT .. √
- Deep context
  - BERT vs. SemBERT √
  - (More effective for inference tasks)
- Discriminative vs generative training：
  - BERT vs. ELECTRA √

Outline

- The Evolution of Pre-trained Language Model
  - Motivation
  - The Path to Pre-trained Language Model
- To Learn Pre-trained Language Model
  - Encoder Design
  - Training Objective： A Unified Perspective of ADE
    - Discriminative
    - Generative
    - Both
  - Tokenization and Masking Unit
- Application of Pre-trained Language Models
  - Multi-task Learning: LIMIT-BERT
  - Extension of Pre-training and Fine-tuning Framework

Use of PrLMs

I. Directly use the output embedding

- Conventional language processing tasks, such as syntax and semantic analysis tasks

II. Fine-tuning

- The PrLM itself is integrated into the system as a module and continues to train according to the target task
- Typical examples are machine reading comprehension task MRC

III. Multi-task

- LIMIT-BERT

IV. New paradigm

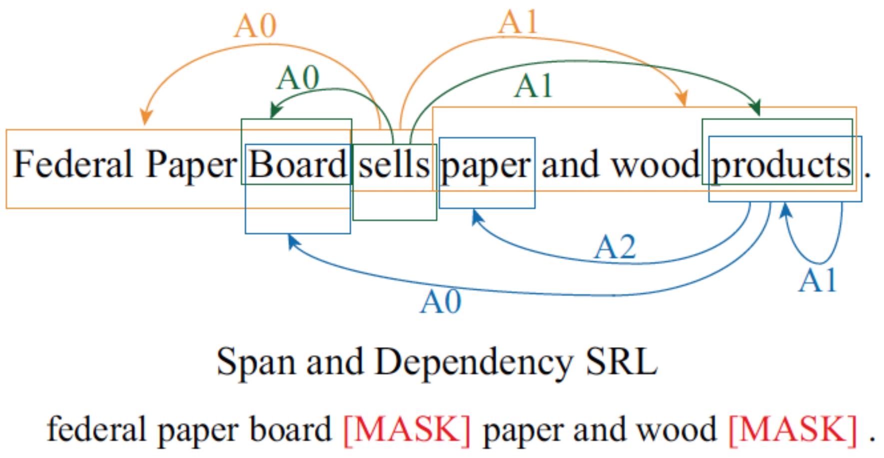
Not just pre-training + fine-tuning?

- Zuchao Li, Hai Zhao, Kevin Parnow. 2020. Global Greedy Dependency Parsing, AAAI-2020.
- https://arxiv.org/abs/1911.08673v3

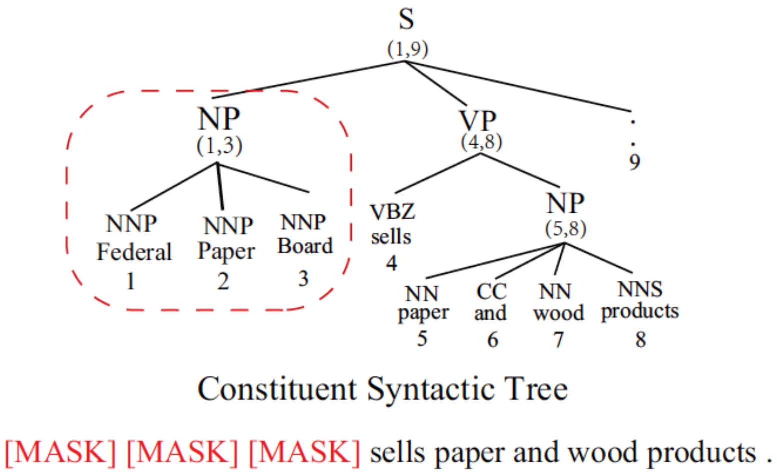
Using fine-tuning in linguistic tasks

LIMIT-BERT

- Junru Zhou, Zhuosheng Zhang, Hai Zhao*, and Shuailiang Zhang. 2020. LIMIT-BERT : Linguistics Informed Multi-Task BERT. EMNLP 2020. ACL Findings.
- Multi task learning: it combines multi task training and semi supervised training to improve the modeling performance of language model from the perspective of computational linguistics.
- Mask strategy: a mask strategy based on syntactic and semantic role annotation is proposed

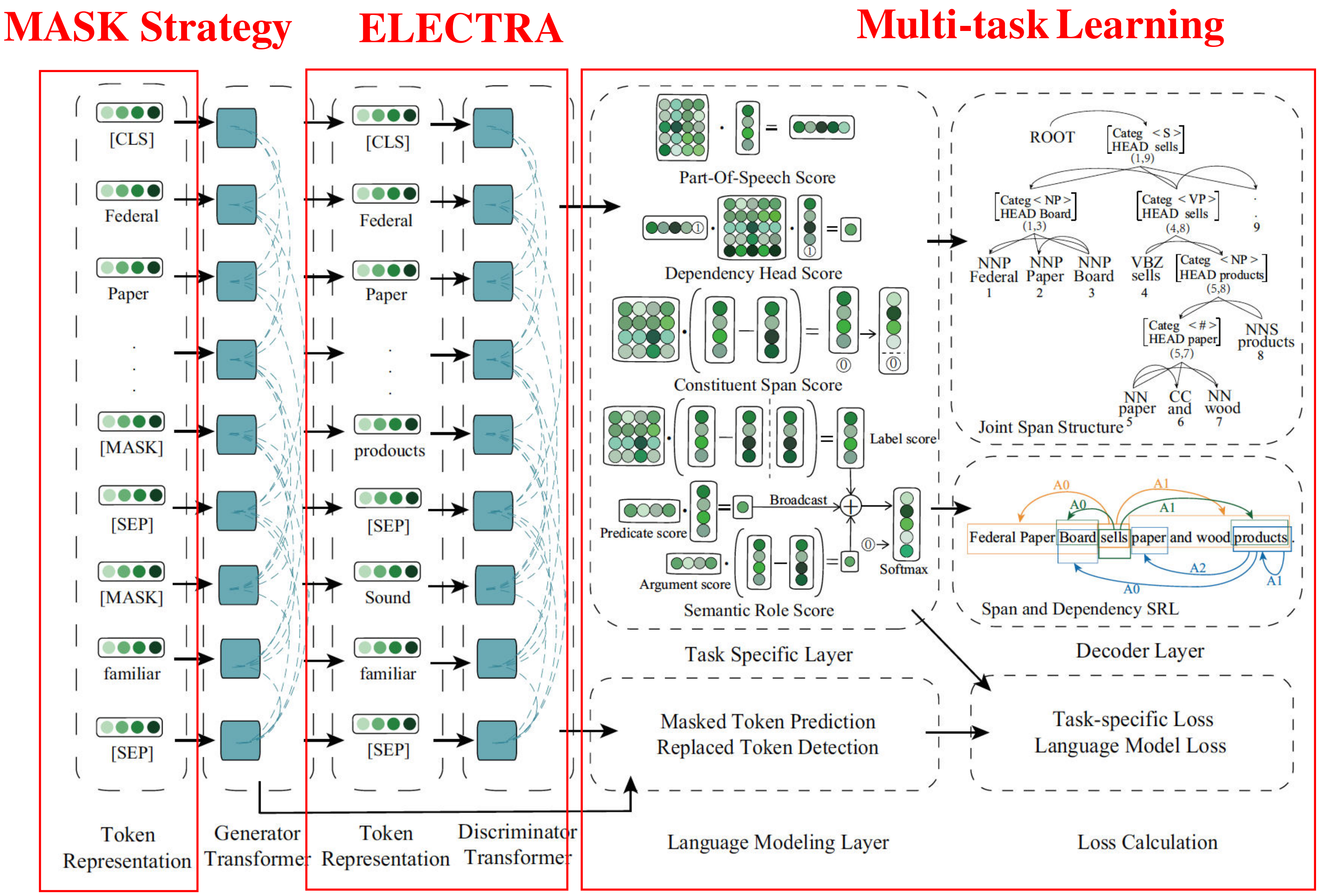


(a) Semantic Phrase Masking.



(b) Syntactic Phrase Masking.

LIMIT-BERT Framework



LIMIT-BERT Performance

	UAS	LAS
Dozat and Manning (2017)	95.74	94.08
Ma et al. (2018)	95.87	94.19
Ji et al. (2019)	95.97	94.31
Fernández-González and Gómez-Rodríguez (2019)	96.04	94.43
Liu et al. (2019a)	96.09	95.03
Zhou and Zhao (2019)(BERT)	97.00	95.43
Zhou et al. (2019)(BERT)	96.90	95.32
Zhou et al. (2019)(XLNet)	97.23	95.65
Baseline ($BERT_{WWM}$)	96.89	95.22
Our LIMIT-BERT	96.94	95.30
Our LIMIT-BERT†	97.14	95.44

Syntactic and semantic analysis tasks

	LR	LP	F1
Gaddy et al. (2018)	91.76	92.41	92.08
Kitaev and Klein (2018)(ELMo)	94.85	95.40	95.13
Kitaev et al. (2019)(BERT)	95.46	95.73	95.59
Zhou and Zhao (2019)(BERT)	95.70	95.98	95.84
Zhou et al. (2019)(BERT)	95.39	95.64	95.52
Zhou et al. (2019)(XLNet)	96.10	96.26	96.18
Baseline ($BERT_{WWM}$)	95.59	95.86	95.72
Our LIMIT-BERT	95.67	95.92	95.80
Our LIMIT-BERT†	95.72	95.96	95.84

System	WSJ			Brown		
	P	R	F_1	P	R	F_1
End-to-end Span SRL						
He et al. (2018a)	81.2	83.9	82.5	69.7	71.9	70.8
He et al. (2018a)(ELMo)	84.8	87.2	86.0	73.9	78.4	76.1
Li et al. (2019)(ELMo)	85.2	87.5	86.3	74.7	78.1	76.4
Strubell et al. (2018)(ELMo)	87.13	86.67	86.90	79.02	77.49	78.25
Zhou et al. (2019)(BERT)	86.46	88.23	87.34	77.26	80.20	78.70
Zhou et al. (2019)(XLNet)	87.48	89.51	88.48	80.46	84.15	82.26
Baseline ($BERT_{WWM}$)	86.48	88.59	87.52	79.4	82.68	81.01
Ours LIMIT-BERT	86.62	89.12	87.85	79.58	83.05	81.28
Ours LIMIT-BERT†	87.16	88.51	87.83	79.20	80.29	79.74
End-to-end Dependency SRL						
Li et al. (2019)	-	-	85.1	-	-	-
He et al. (2018b)	83.9	82.7	83.3	-	-	-
Cai et al. (2018)	84.7	85.2	85.0	-	-	72.5
Li et al. (2019)(ELMo)	84.5	86.1	85.3	74.6	73.8	74.2
Zhou et al. (2019)(BERT)	86.77	89.14	87.94	79.71	82.40	81.03
Zhou et al. (2019)(XLNet)	86.35	90.16	88.21	80.90	85.38	83.08
Baseline ($BERT_{WWM}$)	85.13	89.21	87.12	79.05	83.95	81.43
Ours LIMIT-BERT	85.84	90.01	87.87	79.50	84.85	82.09
Ours LIMIT-BERT†	85.73	89.34	87.50	79.60	82.81	81.17

GLUE

Model	CoLA (mc)	SST-2 (acc)	MRPC (F1/acc)	STS-B (pc/sc)	QQP (acc/F1)	MNLI m/mm(acc)	QNLI (acc)	RTE (acc)	Score -
Dev set results for Comparison									
BERT	60.6	93.2	-/88.0	-/90.0	91.3/-	-/86.6	92.3	70.4	84.0
MT-DNN	63.5	94.3	91.0/87.5	90.7/90.6	91.9/89.2	87.1/86.7	92.9	83.4	-
ELECTRA	69.3	96.0	-/90.6	-/92.1	92.4/-	-/90.5	94.5	86.8	89.0
Baseline ($BERT_{WWM}$)	63.6	93.6	90.8/87.0	90.5/90.2	91.7/88.8	87.4/87.2	93.9	77.3	85.6
LIMIT-BERT	64.0	94.0	94.0/91.7	91.5/91.3	91.6/88.6	87.4/87.3	93.5	85.2	87.3
Test set results for models with standard single-task finetuning									
BiLSTM+ELMo+Attn	36.0	90.4	84.9/77.9	75.1/73.3	64.8/84.7	76.4/76.1	-	56.8	70.5
BERT	60.5	94.9	89.3/85.4	87.6/86.5	72.1/89.3	86.7/85.9	92.7	70.1	80.5
MT-DNN	62.5	95.6	91.1/88.2	89.5/88.8	72.7/89.6	86.7/86.0	93.1	81.4	82.7
SemBERT	62.3	94.6	91.2/88.3	87.8/86.7	72.8/89.8	87.6/86.3	94.6	84.5	82.9
LIMIT-BERT	62.5	94.5	90.9/88.0	90.3/89.7	71.9/89.5	87.1/86.2	94.0	83.0	83.3

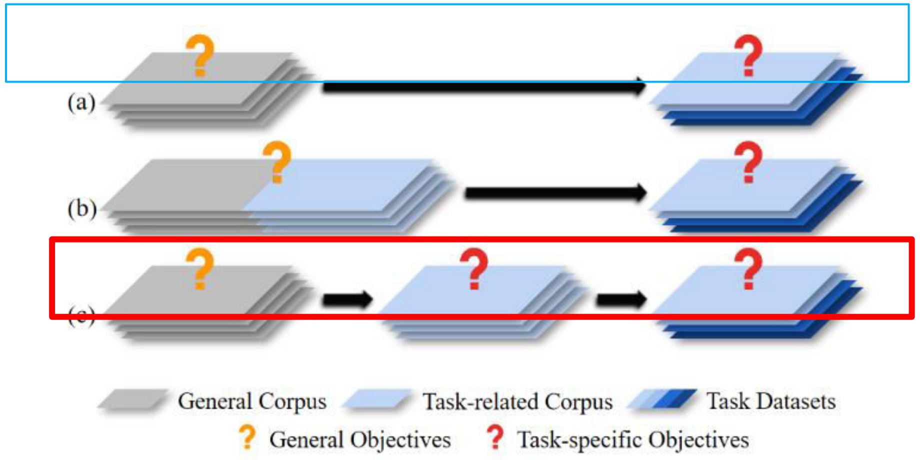
SNLI

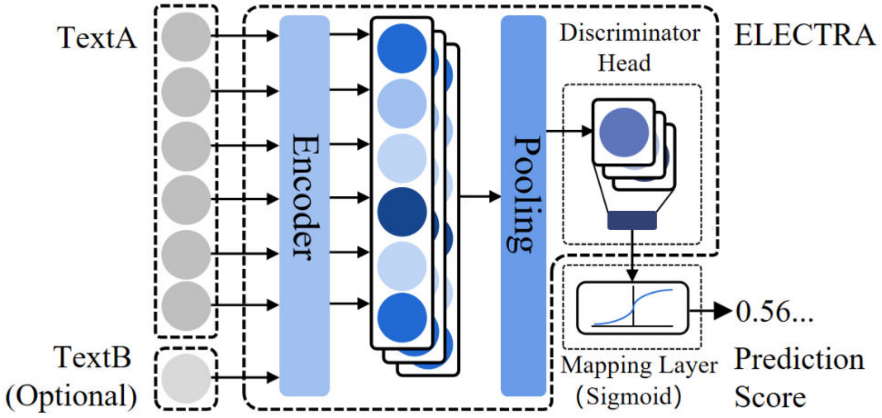
Model	**Dev**	**Test**
DRCN (Kim et al., 2018)	-	90.1
SJRC (Zhang et al., 2019)	-	91.3
MT-DNN (Liu et al., 2019b)	92.2	91.6
SemBERT (Zhang et al., 2020a)	92.3	91.6
Baseline ($BERT_{WWM}$)	91.7	91.4
LIMIT-BERT	92.3	91.7

Outline

- The Evolution of Pre-trained Language Model
  - Motivation
  - The Path to Pre-trained Language Model
- To Learn Pre-trained Language Model
  - Encoder Design
  - Training Objective： A Unified Perspective of ADE
    - Discriminative
    - Generative
    - Both
  - Tokenization and Masking Unit
- Application of Pre-trained Language Models
  - Multi-task Learning: LIMIT-BERT
  - Extension of Pre-training and Fine-tuning Framework

Post-training





- Junlong Li, Zhuosheng Zhang, Hai Zhao, Xi Zhou, Xiang Zhou. 2020. Task-specific Objectives of Pre-trained Language Models for Dialogue Adaptation. arXiv: 2009.04984. ACL-21 review
- Dialogue-Adaptive Pre-training Objective (DAPO)
  - Based on multi-turn dialogue corpus
  - Dialogue quality assessment as a pre-training objective
    - Specificity, Diversity, Readability, Coherence
  - Rich task scenarios, such as dialogue reading comprehension, dialogue selection, dialogue quality evaluation, etc.

New Paradigm, New Performance

(take dialogue as an example)

Model	MuTual			MuTual^{plus}			Model	DREAM	
	R@1	R@2	MRR	R@1	R@2	MRR		Dev	Test
In Paper (Cui et al. 2020)							*In LeaderBoard*		
Dual LSTM	0.266	0.528	0.538	0.266	0.528	0.538	BERT	66.0	66.8
SMN	0.274	0.524	0.575	0.274	0.524	0.575	XLNet	-	72.0
DAM	0.239	0.463	0.575	0.239	0.463	0.575	RoBERTa	85.4	85.0
BERT	0.657	0.867	0.803	0.657	0.867	0.803	MMM	88.0	88.9
RoBERTa	0.695	0.878	0.824	0.695	0.878	0.824	ALBERT	89.2	88.5
BERT-MC	0.661	0.871	0.806	0.661	0.871	0.806	DUMA	89.3	90.4
RoBERTa-MC	0.693	0.887	0.825	0.693	0.887	0.825	DUMA+Multi-Task Learning	**91.9**	**91.8**
Our Implementation									
ELECTRA	0.887	0.969	0.938	0.826	0.949	0.903	ELECTRA	87.4	87.4
ELECTRA-DAPO	**0.907**	**0.976**	**0.949**	**0.827**	**0.962**	**0.907**	ELECTRA-DAPO	88.0	87.7

Table 2: Results on MuTual, MuTual^{plus}, and DREAM datasets. Scores in bold are the current state-of-the-art. The results of MuTual and MuTual^{plus} are for dev set since there is no answer label provided in the test set, we will report the test results after obtaining the numbers from the leaderboard holder.

Model	DailyDialog				PERSONA-CHAT			
	Dev		Test		Dev		Test	
	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman
Our Re-running								
BLEU	0.32	0.14†	0.31	0.25	0.35	0.31	0.36	0.35
ROUGE	0.34	0.22	0.33	0.26	0.36	0.40	0.32	0.43
METEOR	0.37	0.33	0.33	0.27	0.37	0.48	0.34	0.49
BERTScore	0.38	0.31	0.37	0.39	0.40	0.49	0.41	0.42
ADEM	0.28	0.28	0.42	0.45	0.26	0.24	0.25	0.28
RUBER	0.18†	0.15†	0.36	0.30	0.33	0.34	0.38	0.35
RoBERTa-eval	0.68	**0.71**	0.62	0.63	0.72	**0.75**	**0.76**	**0.77**
Our Implementation								
ELECTRA	0.47	0.50	0.45	0.46	0.44	0.46	0.52	0.52
ELECTRA-DAPO	**0.73**	**0.71**	**0.71**	**0.72**	**0.74**	0.70	0.71	0.74

Table 3: Pearson and Spearman correlation with human judgements of *overall quality* on DailyDialog and PERSONA-CHAT datasets. All values that are not statistically significant (p-value > 0.05) are marked by †. Scores in bold are the current state-of-the-art. Following (Zhao, Lala, and Kawahara 2020), we divide the two datasets into train/dev/test set randomly with the ratio 0.8:0.1:0.1, and re-run baselines.

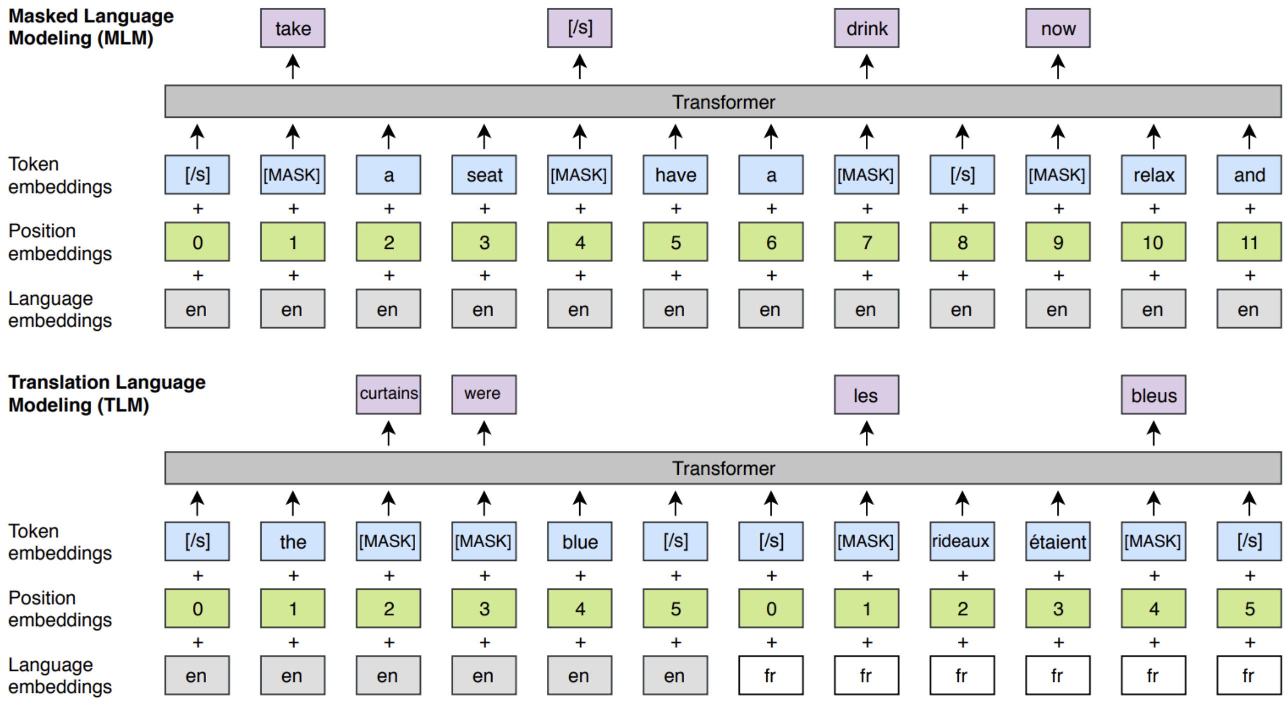
- High-precision Q&A and response selection
- Response quality closer to human level

Menu

- About Us
- Towards Unsupervised Neural Machine Translation (UNMT)
  - Background of Machine Translation (MT)
  - Supervision in MT
  - Unsupervised MT
- Advances in UNMT
  - Pre-trained (Cross-lingual) Language Model
  - Multilingual UNMT
- Challenges in UNMT
  - Reproductive Baselines
  - UNMT & Supervised NMT
  - Distance Language Pairs

Cross-Lingual LM Pre-training

- Large-scale masked cross-lingual language model.

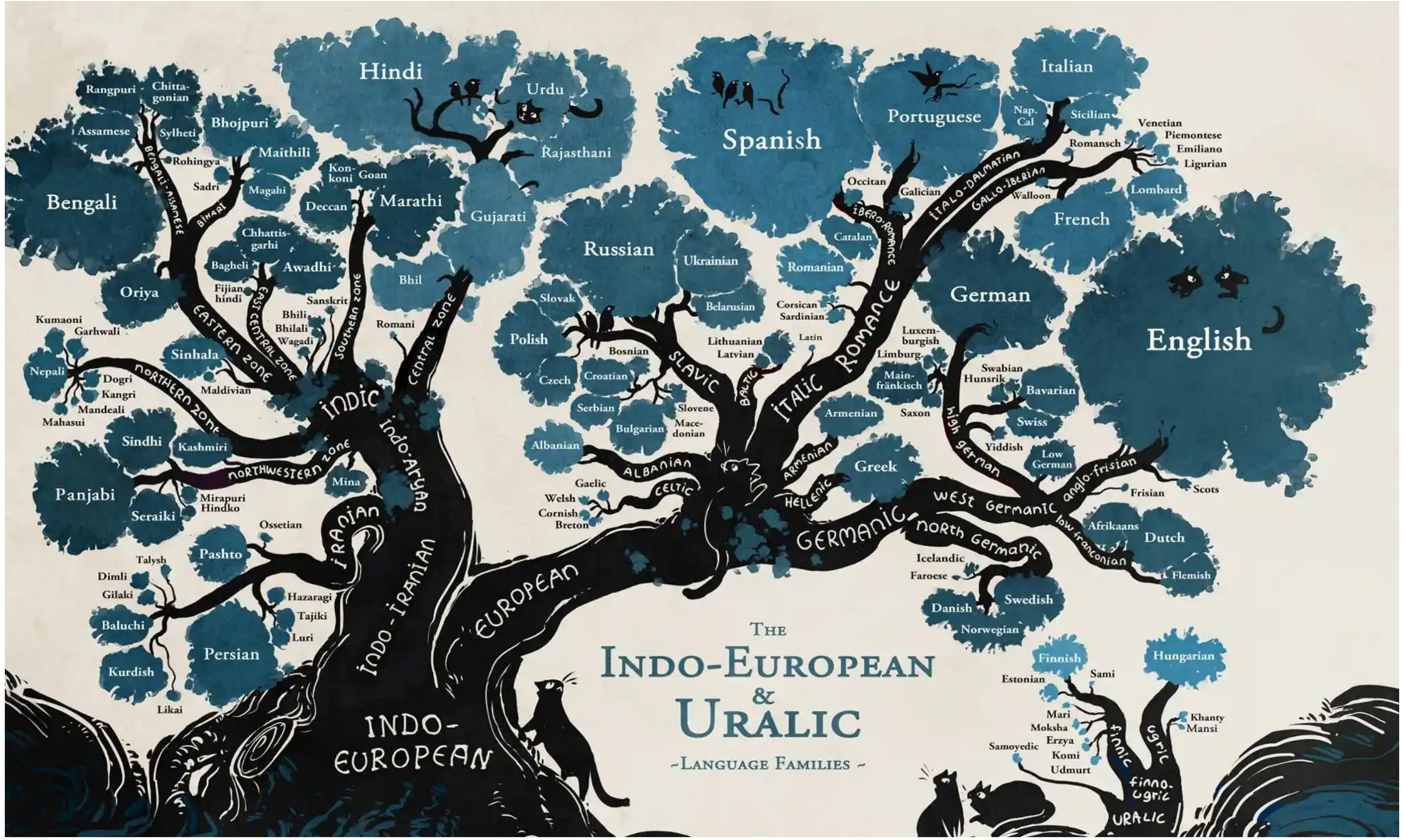


		en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lample et al. (2018b)							
NMT		25.1	24.2	17.2	21.0	21.2	19.4
PBSMT		28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT		27.6	27.7	20.2	25.2	25.1	23.9
Our results for different encoder and decoder initializations							
EMB	EMB	29.4	29.4	21.3	27.3	27.5	26.6
-	-	13.0	15.8	6.7	15.3	18.9	18.3
-	CLM	25.3	26.4	19.2	26.0	25.7	24.6
-	MLM	29.2	29.1	21.6	28.6	28.2	27.3
CLM	-	28.7	28.2	24.4	30.3	29.2	28.0
CLM	CLM	30.4	30.0	22.7	30.5	29.0	27.8
CLM	MLM	32.3	31.6	24.3	32.5	31.6	29.8
MLM	-	31.6	32.1	**27.0**	33.2	31.8	30.5
MLM	CLM	**33.4**	32.3	24.9	32.9	31.7	30.4
MLM	MLM	**33.4**	**33.3**	26.4	**34.3**	**33.3**	**31.8**

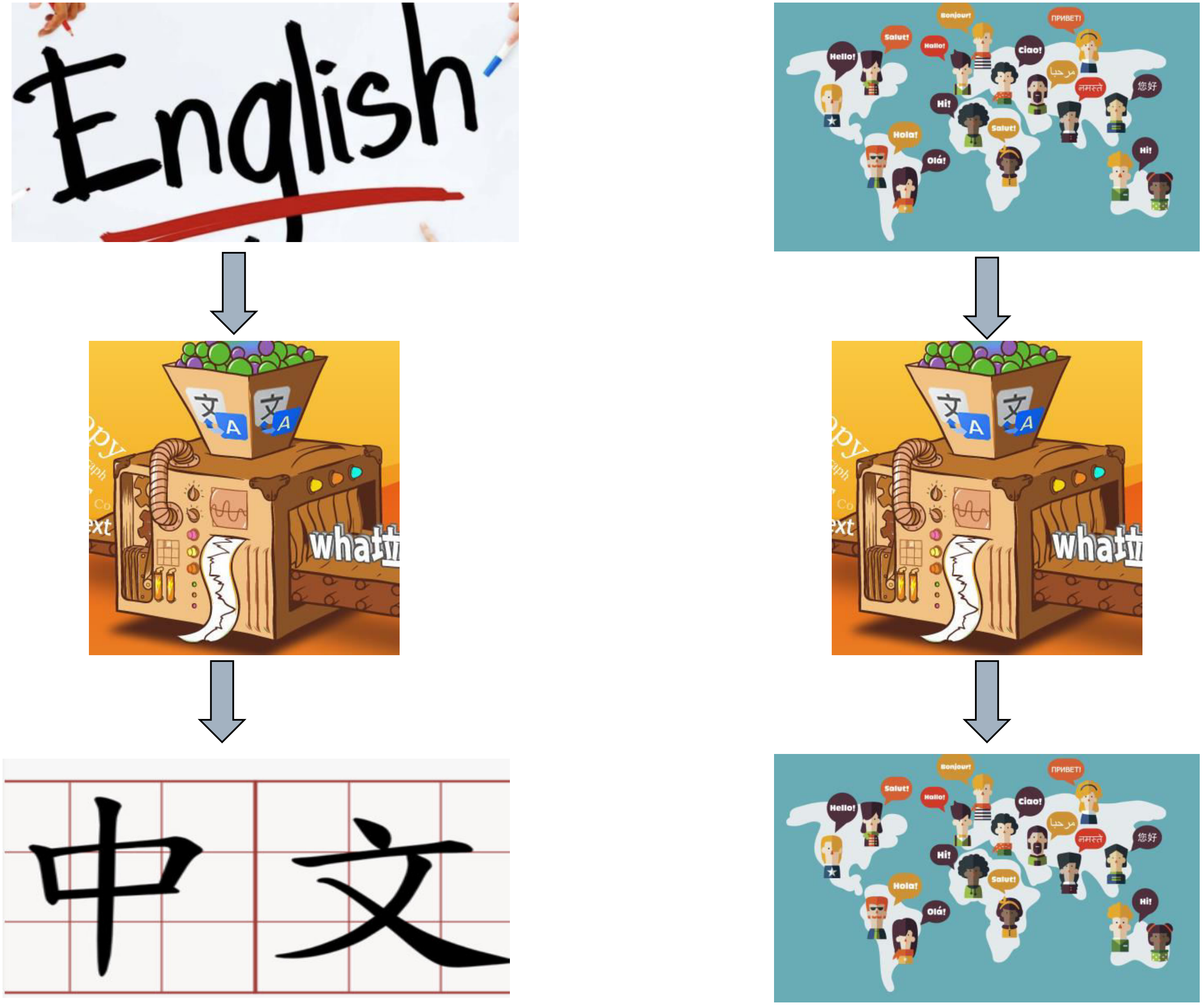
[Lample et al. NeurIPS-2019]

Multi-Lingual Unsupervised Translation

- Challenge
  - There are many language families and groups in the world.
  - The language within certain language families can help each other.

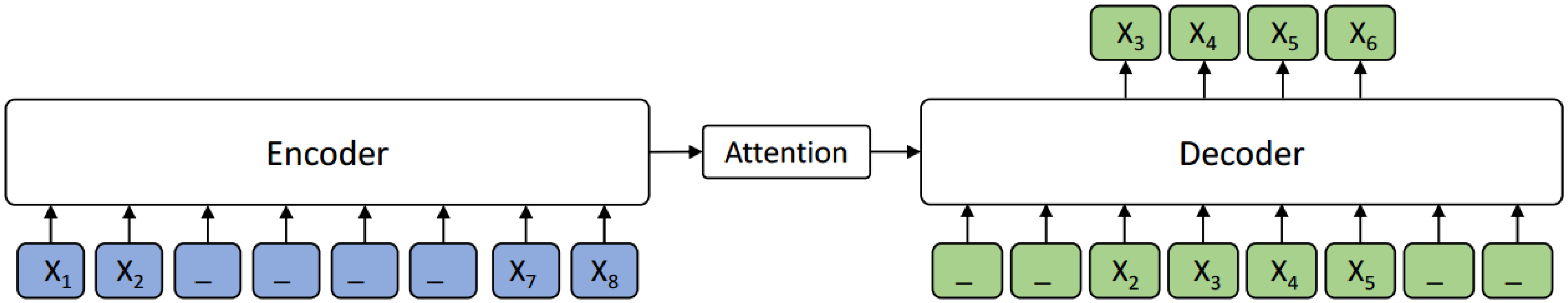


Bilingual & Multi-Lingual Translation



Multi-Lingual Pre-Trained Language Model

- MASS



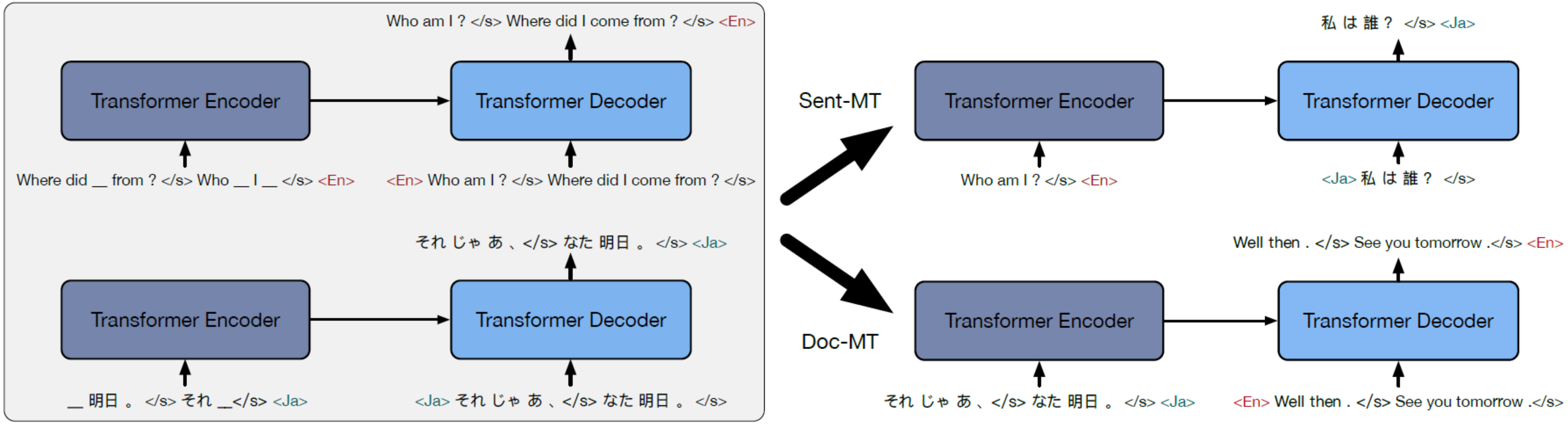
Performance

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

Table 2. The BLEU score comparisons between MASS and the previous works on unsupervised NMT. Results on en-fr and fr-en pairs are reported on *newstest2014* and the others are on *newstest2016*. Since XLM uses different combinations of MLM and CLM in the encoder and decoder, we report the highest BLEU score for XLM on each language pair.

Multi-Lingual Pre-Trained Language Model

- mBART



Multilingual Denoising Pre-Training (mBART)

Fine-tuning on Machine Translation

Performance

Code	Language	Tokens/M	Size/GB
En	English	55608	300.8
Ru	Russian	23408	278.0
Vi	Vietnamese	24757	137.3
Ja	Japanese	530 (*)	69.3
De	German	10297	66.6
Ro	Romanian	10354	61.4
Fr	French	9780	56.8
Fi	Finnish	6730	54.3
Ko	Korean	5644	54.2
Es	Spanish	9374	53.3
Zh	Chinese (Sim)	259 (*)	46.9
It	Italian	4983	30.2
Nl	Dutch	5025	29.3
Ar	Arabic	2869	28.0
Tr	Turkish	2736	20.9
Hi	Hindi	1715	20.2
Cs	Czech	2498	16.3
Lt	Lithuanian	1835	13.7
Lv	Latvian	1198	8.8
Kk	Kazakh	476	6.4
Et	Estonian	843	6.1
Ne	Nepali	237	3.8
Si	Sinhala	243	3.6
Gu	Gujarati	140	1.9
My	Burmese	56	1.6

Table 1: **Languages and Statistics of the CC25 Corpus.** A list of 25 languages ranked with monolingual corpus size. Throughout this paper, we replace the language names with their ISO codes for simplicity. (*) Chinese and Japanese corpus are not segmented, so the tokens counts here are sentences counts

	En-De		En-Ne		En-Si	
	←	→	←	→	←	→
Random	21.0	17.2	0.0	0.0	0.0	0.0
XLM (2019)	34.3	26.4	0.5	0.1	0.1	0.1
MASS (2019)	**35.2**	28.3	–	–	–	–
mBART	34.0	**29.8**	**10.0**	**4.4**	**8.2**	**3.9**

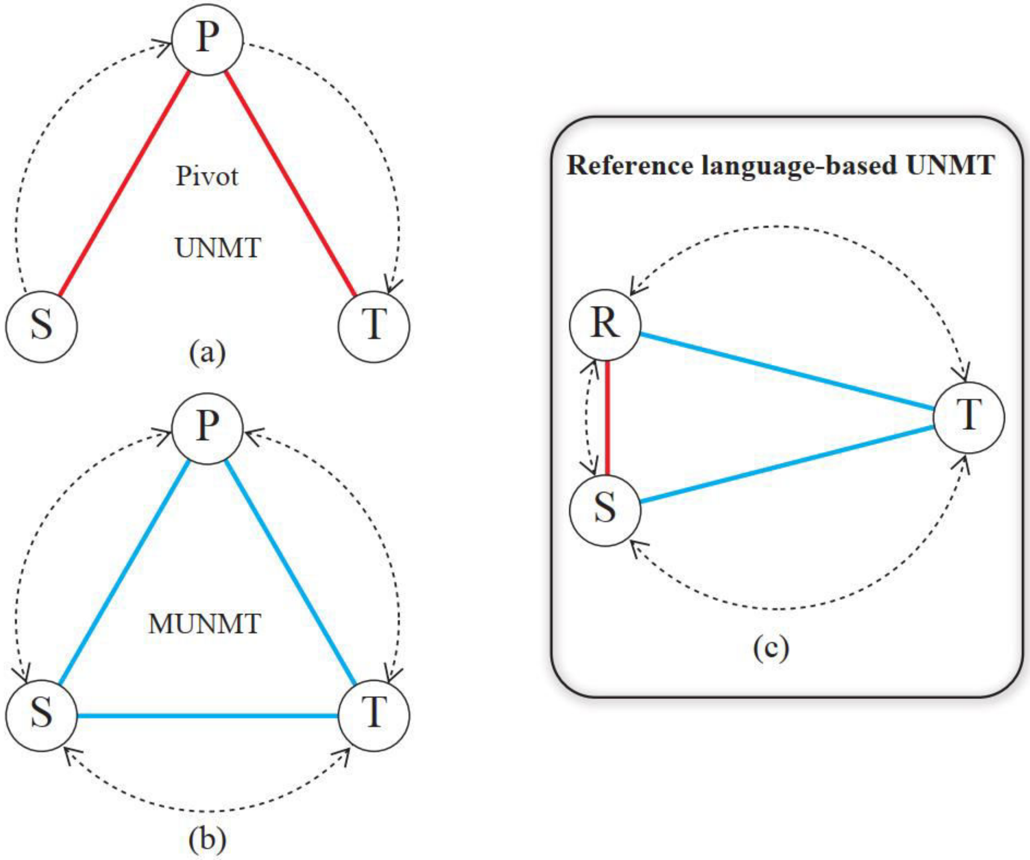
Table 7: Unsupervised MT via BT between dis-similar languages.

Testing Languages	Fine-tuning Languages: Zh	Ja	Ko	Cs	Ro	Nl	It	Ar	Hi	Ne	Si	Gu
Domain	News	TED	TED	News	News	TED	TED	TED	News	Wiki	Wiki	Wiki
Zh	23.7	8.8	**9.2**	2.8	7.8	7.0	6.8	6.2	7.2	4.2	5.9	0.0
Ja	9.9	19.1	**12.2**	0.9	4.8	6.4	5.1	5.6	4.7	4.2	6.5	0.0
Ko	5.8	**16.9**	24.6	5.7	8.5	9.5	9.1	8.7	9.6	8.8	11.1	0.0
Cs	9.3	15.1	17.2	21.6	**19.5**	17.0	16.7	16.9	13.2	15.1	16.4	0.0
Ro	16.2	18.7	17.9	**23.0**	37.8	22.3	21.6	22.6	16.4	18.5	22.1	0.0
Nl	14.4	30.4	32.3	21.2	27.0	43.3	**34.1**	31.0	24.6	23.3	27.3	0.0
It	16.9	25.8	27.8	17.1	23.4	30.2	39.8	**30.6**	20.1	18.5	23.2	0.0
Ar	5.8	**15.5**	12.8	12.7	12.0	14.7	14.7	37.6	11.6	13.0	16.7	0.0
Hi	3.2	10.1	9.9	5.8	6.7	6.1	5.0	7.6	23.5	**14.5**	13.0	0.0
Ne	2.1	6.7	6.5	5.0	4.3	3.0	2.2	5.2	**17.9**	14.5	10.8	0.0
Si	5.0	5.7	3.8	3.8	1.3	0.9	0.5	3.5	8.1	**8.9**	13.7	0.0
Gu	8.2	8.5	4.7	5.4	3.5	2.1	0.0	6.2	**13.8**	13.5	12.8	0.3

Table 11: **Unsupervised MT via Language Transfer** on X-En translations. The model fine-tuned on one language pair is directly tested on another. We use gray color to show the direct fine-tuning results, and lightgray color to show language transfer within similar language groups. We **bold** the highest transferring score for each pair.

Multilingual UNMT: Intuition

- (a) Pivot UNMT: [Leng et al., ACL-2019]
- (b) Multilingual (shared encoder-decoder) UNMT: [Sun et al., ACL-2020]
- (c) Reference language-based UNMT: [Li et al., EMNLP-2020]



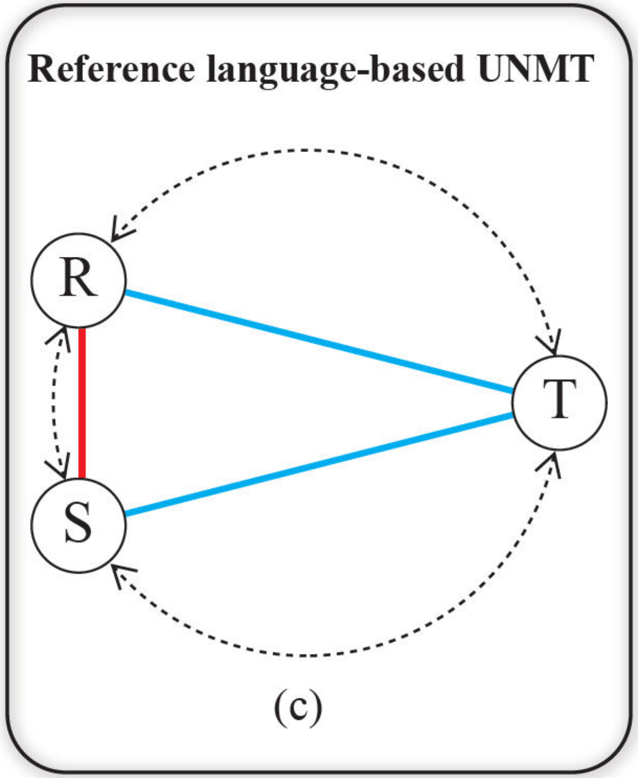
- S: Source language
- T: Target language
- P: Pivot language
- R: Reference language

Reference language-based UNMT (RUNMT)

- RUNMT: some languages are parallel and some not.
  - *Reference Language based Unsupervised Neural Machine Translation*

    Zuchao Li (SJTU), Hai Zhao (SJTU), Rui Wang, Masao Utiyama and Eiichiro Sumita

    The 2020 Conference on Empirical Methods in Natural Language Processing (**EMNLP-Findings**)

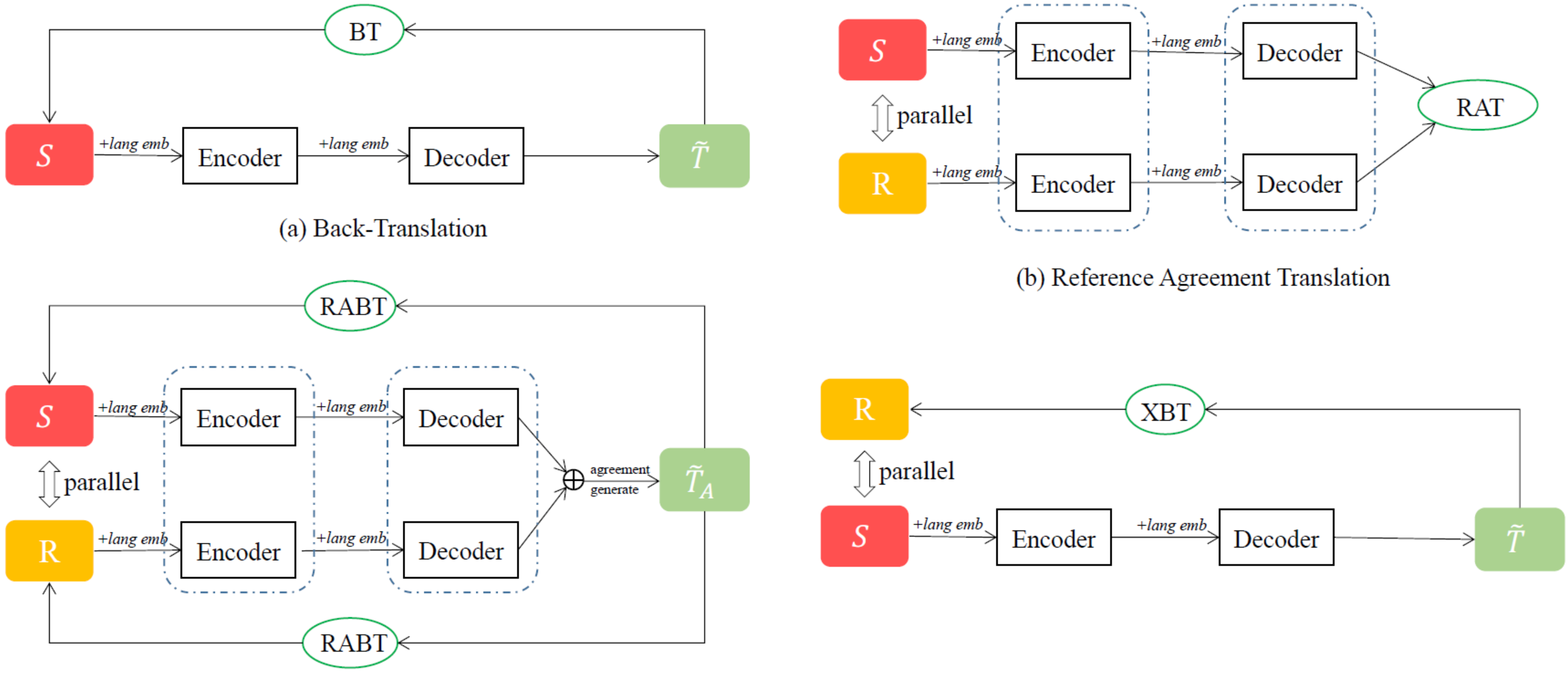


- French—English—Romanian
  - S: Source language (French)
  - T: Target language (English)
  - R: Reference language (Romanian)

Parallel corpus

Monolingual corpus

The Usage of the Reference Language



(a) Back-Translation

(b) Reference Agreement Translation

(c) Reference Agreement Back-Translation

(d) Cross-lingual Back-Translation

Main Results

Pure Unsupervised

	en-fr-ro				*en-zh-ro*				
	en→ro	*ro→en*	*fr→ro*	*ro→fr*	*en→ro*	*ro→en*	*ro→zh*	*zh→ro*	#
PBSMT + NMT	25.13	23.90	n/a	n/a	25.13	23.90	n/a	n/a	1
XLM	33.30	31.80	n/a	n/a	33.30	31.80	n/a	n/a	2
MASS	35.20	33.10	n/a	n/a	35.20	33.10	n/a	n/a	3
UNMT	34.45	32.42	25.26	27.99	34.45	32.42	8.66 [2.31]	10.92 [3.56]	4
MUNMT	34.44	32.60	25.31	27.91	33.79	31.82	8.85 [2.63]	11.55 [3.87]	5
+ RAT	35.83	33.52	25.66	28.25	34.59	32.12	9.73 [3.02]	12.44 [3.95]	6
+ RABT	36.05	33.74	25.65	28.44	35.23	32.67	10.09 [3.30]	12.95 [4.00]	7
+ XBT	36.08	33.84	25.78	28.45	34.76	32.30	10.54 [3.32]	13.66 [4.03]	8
+ALL	36.14	34.12	25.60	28.89	35.66	32.88	10.83 [3.44]	13.75 [4.24]	9
MUNMT + RNMT	36.39	33.85	25.53	28.57	35.50	33.66	10.98 [3.64]	14.42 [4.39]	10
+ RAT	36.65	34.07	25.78	28.63	36.26	34.18	11.26 [3.87]	14.77 [4.78]	11
+ RABT	36.84	34.32	25.75	29.04	36.78	34.26	11.52 [3.90]	14.79 [5.01]	12
+ XBT	37.13	34.66	26.02	29.11	36.31	34.14	11.80 [4.03]	14.86 [4.98]	13
+ALL	**37.27**	**34.85**	**26.50**	**29.45**	**37.01**	**34.55**	**11.92 [4.07]**	**15.02 [5.11]**	14

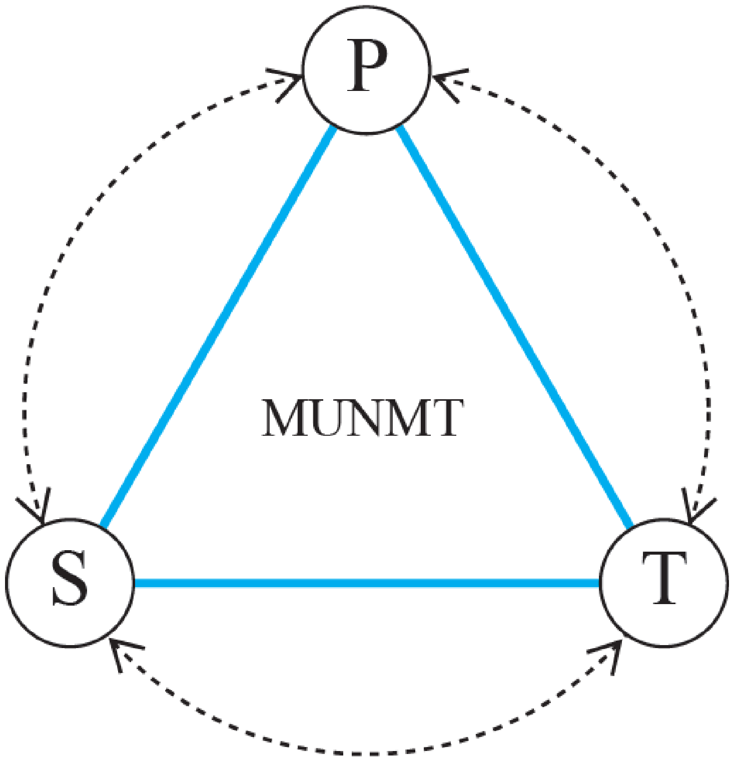
With Source-Reference Parallel Corpus

Multilingual UNMT (MUNMT)

- MUNMT is a general structure.
  - *Knowledge Distillation for Multilingual Unsupervised Neural Machine Translation*

    Haipeng Sun (HIT), Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao(HIT)

    The 58th Annual Meeting of the Association for Computational Linguistics (**ACL-2020**)

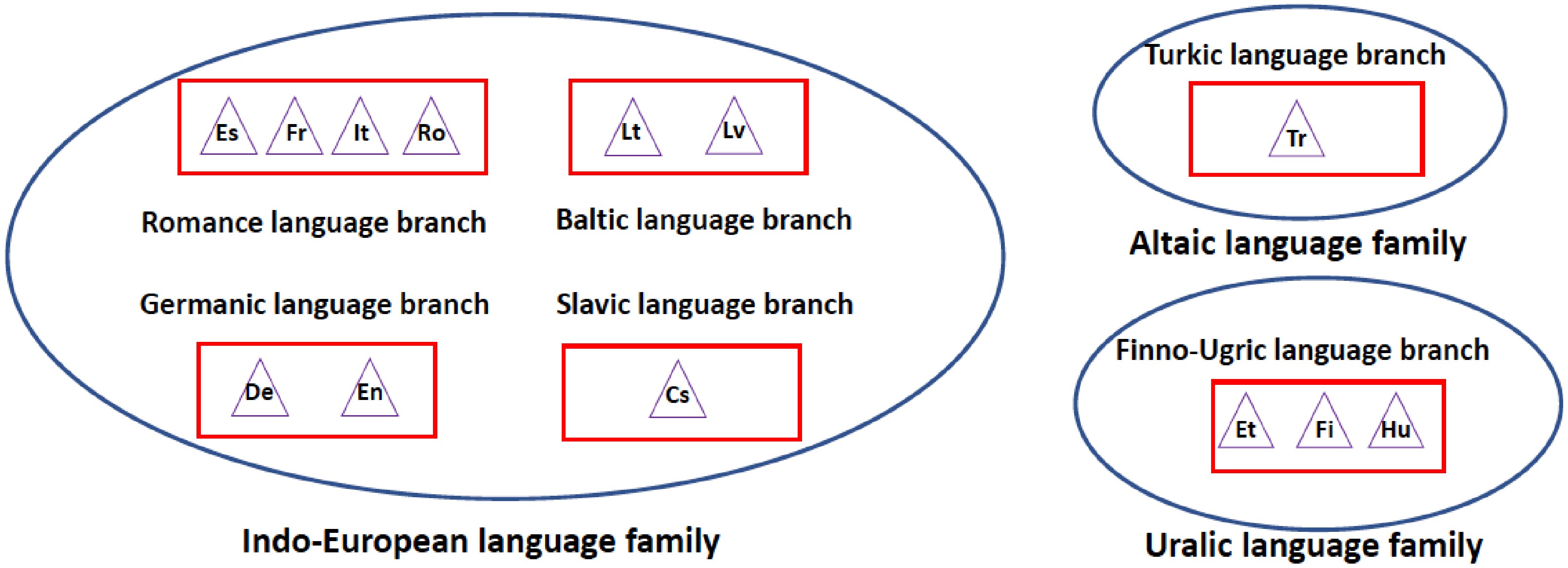


Monolingual corpus

Datasets

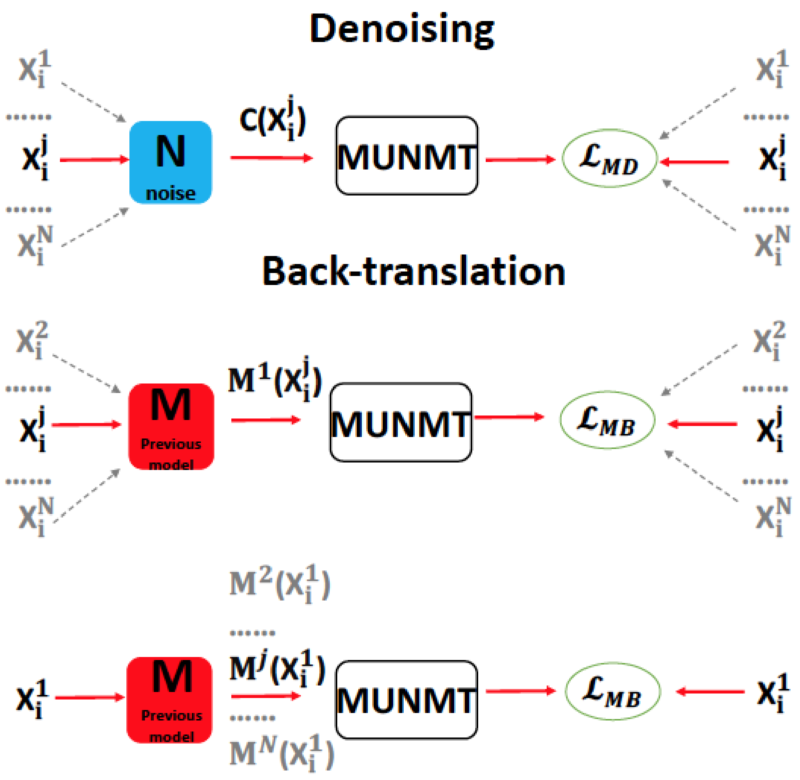
- 13 European languages from WMT monolingual news crawl datasets: Cs, De, En, Es, Et, Fi, Fr, Hu, It, Lt, Lv, Ro, and Tr.
- WMT newstest2013 for Cs-En, De-En, Es-En, and Fr-En are mutual parallel.

Language	Sentences	Words	Sub-words
Cs	50.00M	860.36M	1.16B
De	50.00M	887.37M	1.19B
En	50.00M	1.15B	1.32B
Es	36.33M	1.01B	1.19B
Et	3.00M	51.39M	101.43M
Fi	15.31M	189.39M	359.78M
Fr	50.00M	1.19B	1.38B
Hu	34.35M	708.13M	1.03B
It	30.82M	755.56M	911.51M
Lt	0.34M	6.38M	14.64M
Lv	8.60M	172.56M	281.54M
Ro	8.92M	207.07M	279.95M
Tr	9.14M	153.03M	254.70M



Multilingual Unsupervised Neural Machine Translation

- Multilingual Pretraining
  - To construct a multilingual masked language model, using a single encoder.
  - To initialize the full set of parameters of MUNMT



Multilingual Unsupervised Neural Machine Translation

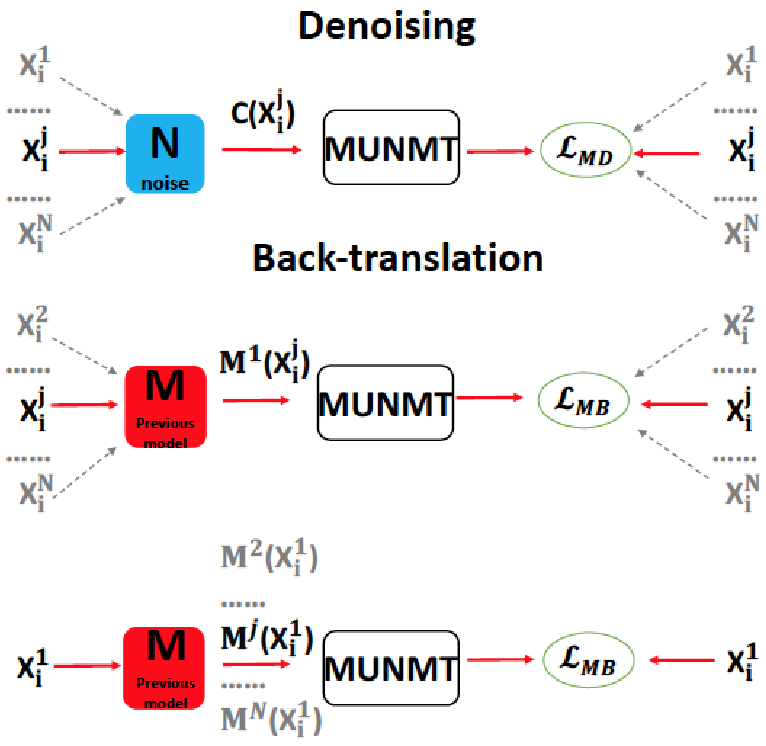
- Multilingual Pretraining
  - To construct a multilingual masked language model, using a single encoder.
  - To initialize the full set of parameters of MUNMT
- Multilingual UNMT Training
  - Denoising training

$$\mathcal{L}_{MD} = \sum_{j=1}^{N} \sum_{i=1}^{|X^j|} -log P_{L_j \to L_j}(X_i^j | C(X_i^j)),$$

  - Back-translation training

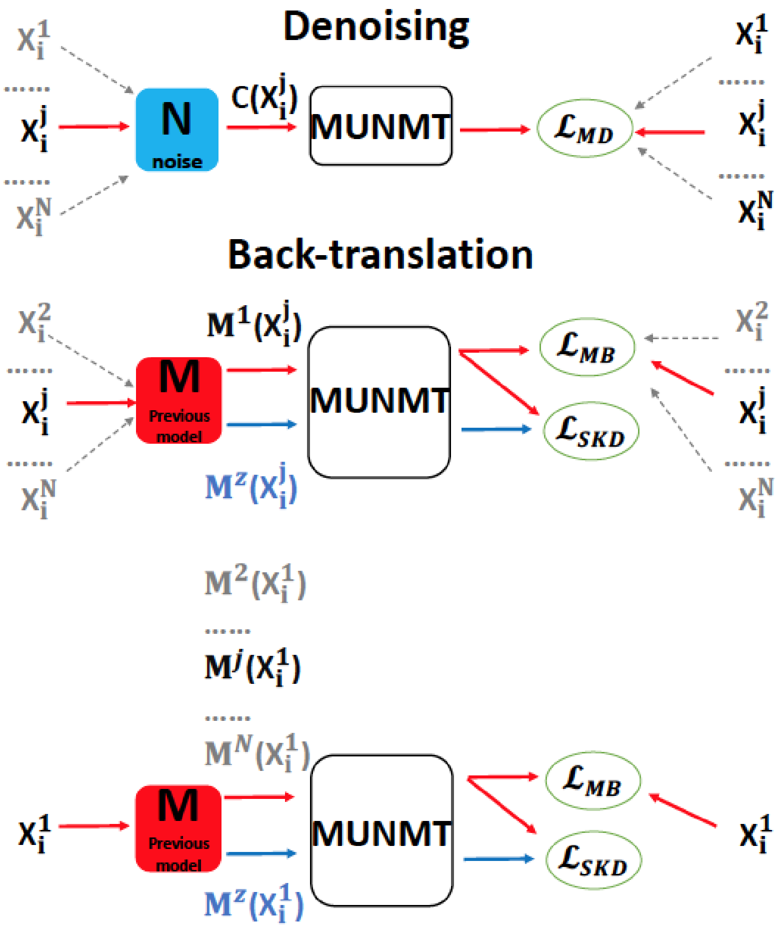
$$\mathcal{L}_{MB} = \sum_{j=2}^{N} \sum_{i=1}^{|X^1|} -log P_{L_j \to L_1}(X_i^1 | M^j(X_i^1))$$

$$+ \sum_{j=2}^{N} \sum_{i=1}^{|X^j|} -log P_{L_1 \to L_j}(X_i^j | M^1(X_i^j)),$$



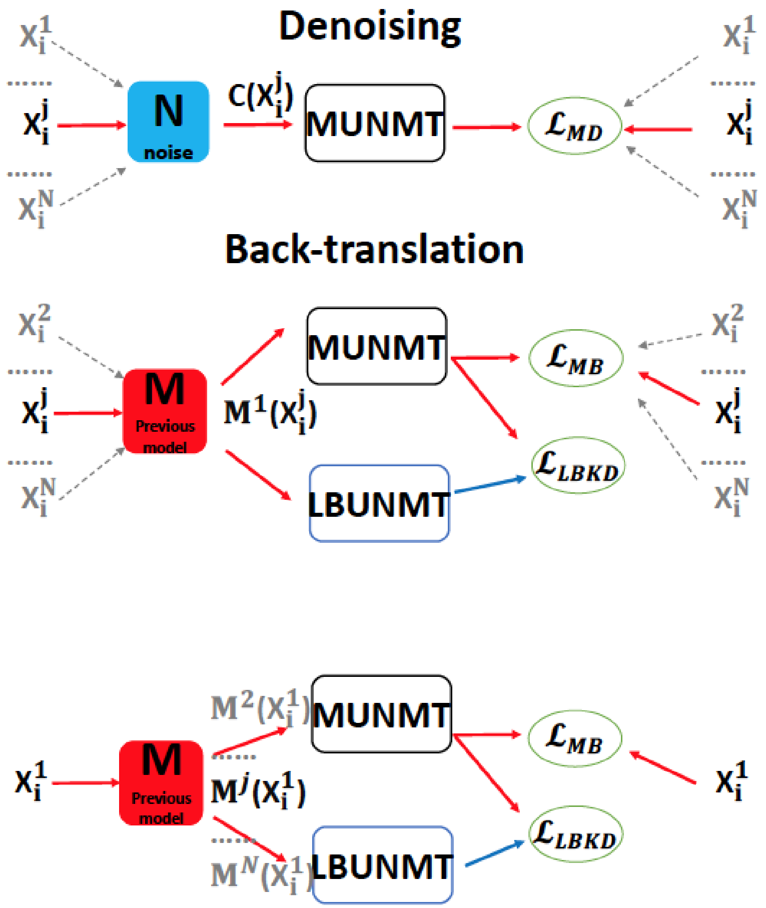
Self-knowledge Distillation

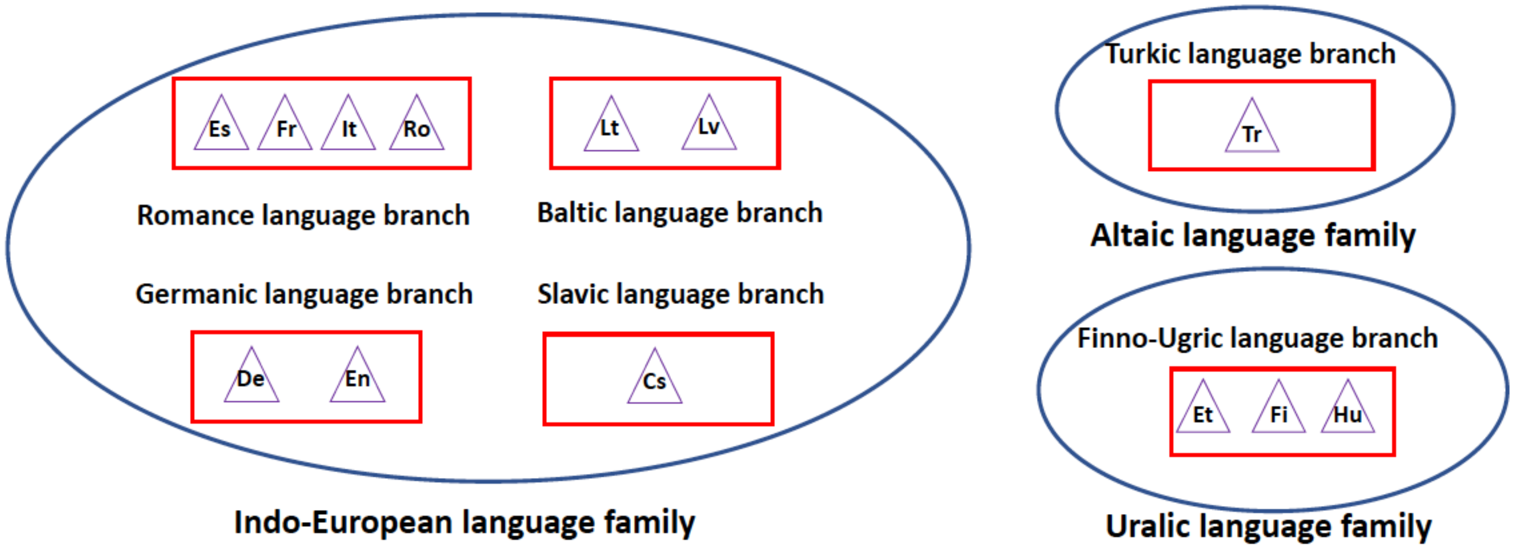
- During back-translation, only language L_j sentences are generated before training the MUNMT model in the $L_j \rightarrow L_1$ direction. However, other languages are not used during this training.
- We propose to introduce another language L_z (randomly chosen but distinct from L_1 and L_j) during this training.
- The translation from the source sentences through different paths, $L_1 \rightarrow L_j \rightarrow L_1$ and $L_1 \rightarrow L_z \rightarrow L_1$, should be similar.



Language Branch Knowledge Distillation

- LBUNMT model performed better than the single model because similar languages have a positive interaction
- The distilled information of LBUNMT is used to guide the MUNMT model during back-translation.





Main Results

Corpus	SNMT	Sen et al. (2019)	Xu et al. (2019)	SM	LBUNMT	MUNMT	SKD	LBKD
En-Cs	19.20	-	6.79	14.54	14.54	14.40	14.89	**15.47**
En-De	20.30	8.09	13.25	18.26	18.26	17.58	18.47	**19.28**
En-Es	30.40	14.82	20.43	25.14	25.40	25.05	25.61	**26.79**
En-Et	25.20	-	-	14.86	15.02	14.09	15.03	**15.62**
En-Fi	27.40	-	-	9.87	9.99	9.75	**10.70**	10.57
En-Fr	30.60	13.71	20.27	26.02	26.36	25.84	26.45	**27.78**
En-Hu	-	-	-	11.32	11.40	10.90	11.64	**12.03**
En-It	-	-	-	24.19	24.30	23.80	24.69	**25.52**
En-Lt	20.10	-	-	0.79	8.29	10.07	**11.15**	11.11
En-Lv	21.10	-	-	1.02	11.55	13.09	13.90	**14.33**
En-Ro	28.90	-	-	29.44	29.58	28.82	29.65	**31.28**
En-Tr	20.00	-	-	11.87	11.87	12.41	13.24	**13.83**
Average	-	-	-	15.61	17.21	17.15	17.95	**18.63**

Baselines:
SNMT: supervised NMT
SM: single language pair NMT
LBUNMT: UNMT in language branch
MUNMT: multi-lingual UNMT

Ours:
SKD: self-knowledge distillation
LBKD: language branch SKD

- LBUNMT performed better than SM because similar languages have a positive interaction during the training process.
- However, the performance of MUNMT is slightly worse than SM in some language pairs.

Main Results

Corpus	SNMT	Sen et al. (2019)	Xu et al. (2019)	SM	LBUNMT	MUNMT	SKD	LBKD
En-Cs	19.20	-	6.79	14.54	14.54	14.40	14.89	**15.47**
En-De	20.30	8.09	13.25	18.26	18.26	17.58	18.47	**19.28**
En-Es	30.40	14.82	20.43	25.14	25.40	25.05	25.61	**26.79**
En-Et	25.20	-	-	14.86	15.02	14.09	15.03	**15.62**
En-Fi	27.40	-	-	9.87	9.99	9.75	**10.70**	10.57
En-Fr	30.60	13.71	20.27	26.02	26.36	25.84	26.45	**27.78**
En-Hu	-	-	-	11.32	11.40	10.90	11.64	**12.03**
En-It	-	-	-	24.19	24.30	23.80	24.69	**25.52**
En-Lt	20.10	-	-	0.79	8.29	10.07	**11.15**	11.11
En-Lv	21.10	-	-	1.02	11.55	13.09	13.90	**14.33**
En-Ro	28.90	-	-	29.44	29.58	28.82	29.65	**31.28**
En-Tr	20.00	-	-	11.87	11.87	12.41	13.24	**13.83**
Average	-	-	-	15.61	17.21	17.15	17.95	**18.63**

Language	Sentences	Words	Sub-words
Cs	50.00M	860.36M	1.16B
De	50.00M	887.37M	1.19B
En	50.00M	1.15B	1.32B
Es	36.33M	1.01B	1.19B
Et	3.00M	51.39M	101.43M
Fi	15.31M	189.39M	359.78M
Fr	50.00M	1.19B	1.38B
Hu	34.35M	708.13M	1.03B
It	30.82M	755.56M	911.51M
Lt	0.34M	6.38M	14.64M
Lv	8.60M	172.56M	281.54M
Ro	8.92M	207.07M	279.95M
Tr	9.14M	153.03M	254.70M

- SM performed very poorly on low-resource language pairs such as En-Lt and En-Lv in the Baltic language branch.

Main Results

Corpus	SNMT	Sen et al. (2019)	Xu et al. (2019)	SM	LBUNMT	MUNMT	SKD	LBKD
Cs-En	27.10	-	11.56	20.62	20.62	20.09	21.05	**21.25**
De-En	28.40	11.94	16.46	21.31	21.31	21.95	22.54	**22.81**
Es-En	31.40	15.45	20.35	25.53	25.77	25.37	26.15	**26.59**
Et-En	30.90	-	-	19.48	20.30	19.60	20.95	**21.31**
Fi-En	33.00	-	-	7.62	7.68	7.19	**7.92**	7.80
Fr-En	32.20	14.47	19.87	25.86	26.02	25.41	26.07	**26.48**
Hu-En	-	-	-	14.48	14.86	14.54	15.16	**15.34**
It-En	-	-	-	24.33	24.87	24.77	25.30	**25.35**
Lt-En	36.30	-	-	1.72	11.00	14.04	15.31	**15.84**
Lv-En	21.90	-	-	0.95	12.75	14.90	**15.49**	15.33
Ro-En	35.20	-	-	28.52	29.57	28.38	29.58	**30.18**
Tr-En	28.00	-	-	12.99	12.99	15.65	16.85	**17.35**
Average	-	-	-	16.95	18.98	19.32	20.20	**20.47**

Baselines:
SNMT: supervised NMT
SM: single language pair NMT
LBUNMT: UNMT in language branch
MUNMT: multi-lingual UNMT

Ours:
SKD: self-knowledge distillation
LBKD: language branch SKD

- Our proposed knowledge distillation method outperformed the original MUNMT model by approximately 1 BLEU score.
- Regarding our two proposed methods, LBKD achieved better performance since it could obtain much more knowledge distilled from LBUNMT model.

Main Results

Corpus	SNMT	Sen et al. (2019)	Xu et al. (2019)	SM	LBUNMT	MUNMT	SKD	LBKD
Cs-En	27.10	-	11.56	20.62	20.62	20.09	21.05	**21.25**
De-En	28.40	11.94	16.46	21.31	21.31	21.95	22.54	**22.81**
Es-En	31.40	15.45	20.35	25.53	25.77	25.37	26.15	**26.59**
Et-En	30.90	-	-	19.48	20.30	19.60	20.95	**21.31**
Fi-En	33.00	-	-	7.62	7.68	7.19	**7.92**	7.80
Fr-En	32.20	14.47	19.87	25.86	26.02	25.41	26.07	**26.48**
Hu-En	-	-	-	14.48	14.86	14.54	15.16	**15.34**
It-En	-	-	-	24.33	24.87	24.77	25.30	**25.35**
Lt-En	36.30	-	-	1.72	11.00	14.04	15.31	**15.84**
Lv-En	21.90	-	-	0.95	12.75	14.90	**15.49**	15.33
Ro-En	35.20	-	-	28.52	29.57	28.38	29.58	**30.18**
Tr-En	28.00	-	-	12.99	12.99	15.65	16.85	**17.35**
Average	-	-	-	16.95	18.98	19.32	20.20	**20.47**

Baselines:
SNMT: supervised NMT
SM: single language pair NMT
LBUNMT: UNMT in language branch
MUNMT: multi-lingual UNMT

Ours:
SKD: self-knowledge distillation
LBKD: language branch SKD

- Our proposed MUNMT with knowledge distillation performed better than SM in all language pairs.
- There is a gap between the performance of our proposed MUNMT model and that of the supervised NMT systems.

Zero-shot Translation Analysis

- Zero-shot Translation: MUNMT was trained in 24 translation directions whereas 156 translation directions exist.
- Our proposed knowledge distillation methods further improved the performance of zero-shot translation.
- SKD significantly outperformed LBKD by approximately 3 BLEU scores since the third language was introduced during SKD translation training for two language pairs, achieving much more cross-lingual knowledge.

Methods	→	Cs	De	Es	Fr
Xu et al. (2019)		-	11.16	11.29	10.61
Sen et al. (2019)		-	-	-	-
MUNMT	Cs	-	11.91	15.22	14.66
LBKD		-	13.16	16.63	16.28
SKD		-	**16.96**	**20.52**	**20.14**
Xu et al. (2019)		10.52	-	13.68	9.45
Sen et al. (2019)		-	-	7.40	6.78
MUNMT	De	10.56	-	16.15	15.85
LBKD		11.53	-	17.27	16.96
SKD		**14.58**	-	**20.20**	**20.61**
Xu et al. (2019)		8.32	11.20	-	24.13
Sen et al. (2019)		-	4.78	-	13.92
MUNMT	Es	10.04	11.87	-	21.90
LBKD		10.86	12.98	-	23.05
SKD		**13.63**	**16.62**	-	**27.04**
Xu et al. (2019)		8.89	11.24	23.88	-
Sen et al. (2019)		-	4.59	13.87	-
MUNMT	Fr	9.77	11.70	22.30	-
LBKD		10.48	12.67	22.65	-
SKD		**13.04**	**16.31**	**25.92**	-

Menu

- About Us
- Towards Unsupervised Neural Machine Translation (UNMT)
  - Background of Machine Translation (MT)
  - Supervision in MT
  - Unsupervised MT
- Advances in UNMT
  - Pre-trained Cross-lingual Language Model
  - Multilingual UNMT
- Challenges in UNMT
  - Reproductive Baselines
  - UNMT & Supervised NMT
  - Distance Language Pairs

Challenges in UNMT

- In this Section, I only show the brief topic.
- I hope we can discuss the details in the Q/A session.

Reproductive Baselines

- As mentioned above, most of the baselines are not reimplemented.
- Instead, only reporting other results are not so convincing.
- We will maintain the baseline system with available codes, model, etc. at https://wangruinlp.github.io/unmt

UNMT & Supervised NMT

- Fine-tune with small parallel data can significantly improve the UNMT performance.

#	Methods	de-cs
1	Single UNMT system	15.5
2	Single USMT system	11.1
3	Single NMT system pseudo-supervised by UNMT	15.9
4	Single NMT system pseudo-supervised by USMT	15.3
5	Single Pseudo-supervised MT system	16.2
6	Ensemble Pseudo-supervised MT system	16.5
7	Re-ranking Pseudo-supervised MT system	17.0
8	Fine-tuning Pseudo-supervised MT system	18.7
9	Fine-tuning Pseudo-supervised MT system + fixed quotes	19.6
10	Fine-tuning + re-ranking Pseudo-supervised MT system + fixed quotes	20.1

Table 4: BLEU scores of UMT. #10 is our primary system submitted to the organizers.

Distant Language Pairs

- There are few shared words between distant language pairs

Languages	Similar Language Pairs		Distant Language Pair	
	De-En	Fr-En	Ja-En	Zh-En
Shared Words	37,257	43,642	454	20,662
Ratio of Shared Words	23.30%	25.40%	0.18%	4.91%
UNMT Performance（BLEU）	27.6	25.1	14.1	8.02

Distant Language Pairs

- The word orders are quite different between distant language pairs
- We analyze the word order similarity using [Chen and Wang et al., 2020]

Languages	Similar Language Pairs		Distant Language Pair	
	De-En	Fr-En	Ja-En	Zh-En
Word Order Similairy	76.3%	78.1%	53.4%	62.2%
Supervised NMT （BLEU）	40.2	35.0	30.9	26.4
Unsupervised NMT （BLEU）	27.6	25.1	14.1	8.02

Thank You!

Welcome to revisit this tutorial and contact us!

https://wangruinlp.github.io/unmt