Towards Unsupervised Machine Translation

王 瑞 (Wang, Rui)

Department of Computer Science and Engineering

Shanghai Jiao Tong University

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

Menu

- About Me
- Background of Machine Translation (MT)
- Supervision in MT
- Unsupervised MT

About Me

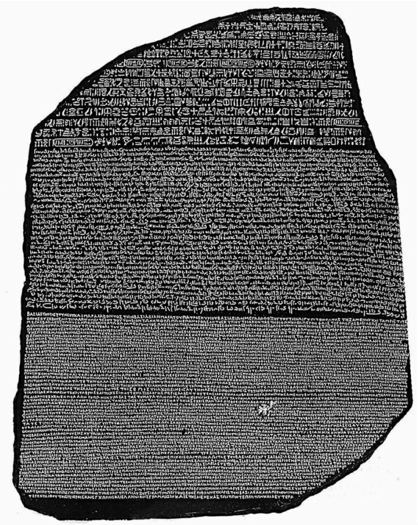
- Experience:
  - 2021-: Associate Professor & Ph.D. Advisor, Shanghai Jiao Tong University, Shanghai, China
  - 2016-2020: Postdoctoral/Tenure-Track/Tenured Researcher, National Institute of Information and Communications Technology, Kyoto, Japan
- Research Interest:
  - Machine Translation (MT)
  - Multilingual Natural Language Processing (NLP)
- Recent Research Activity
  - Area Chair: ICLR-2021 and NAACL-2021
  - Tutorial: ***EACL-2021* (this talk)** *and EMNLP-2021*
- Homepage of this tutorial:
  - https://wangruinlp.github.io/unmt

Menu

- About Me
- **Background of Machine Translation (MT)**
- Supervision in MT
- Unsupervised MT

MT: History

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



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.

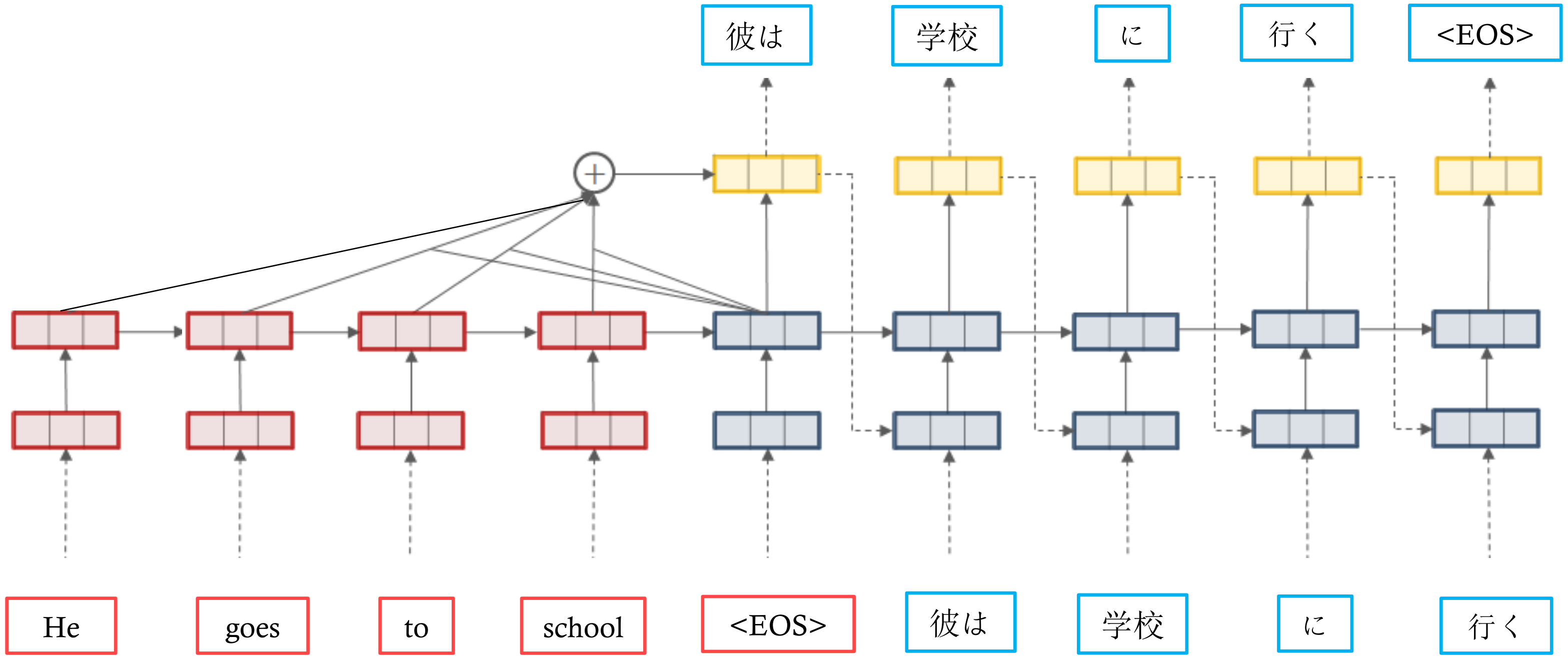
MT: from ML aspect

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

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

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

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

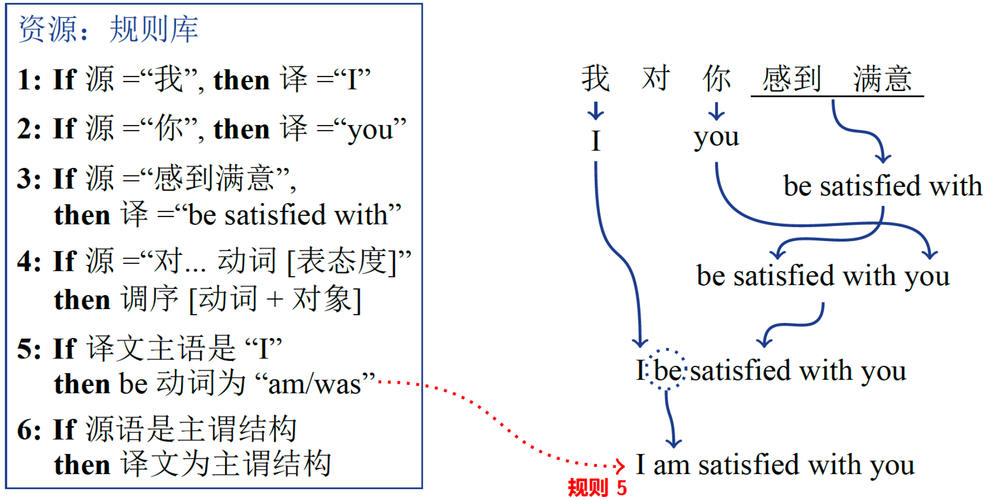


Menu

- About Me
- Background of Machine Translation (MT)
- **Supervision in MT**
- Unsupervised MT

Supervision in MT

- Rule-based MT:
  - Annotated linguistic rules

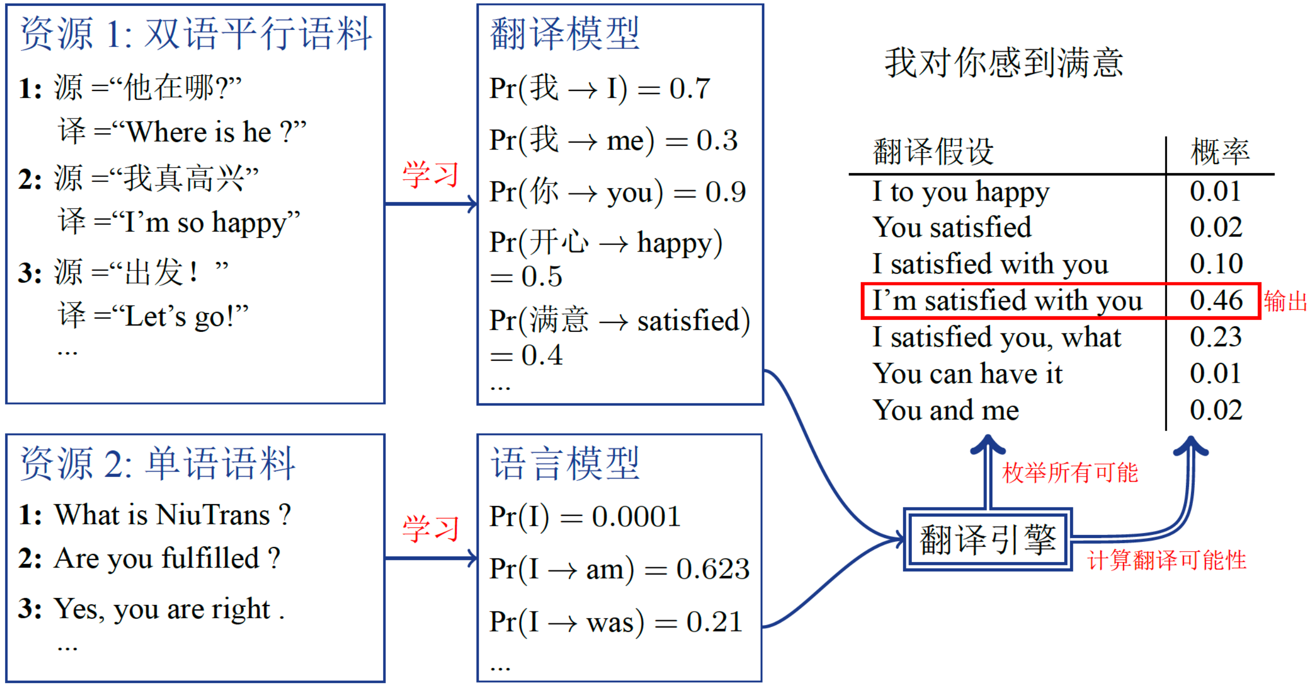


- Example-based MT:
  - Translation examples

[Examples from Xiao and Zhu, SMT-Book]

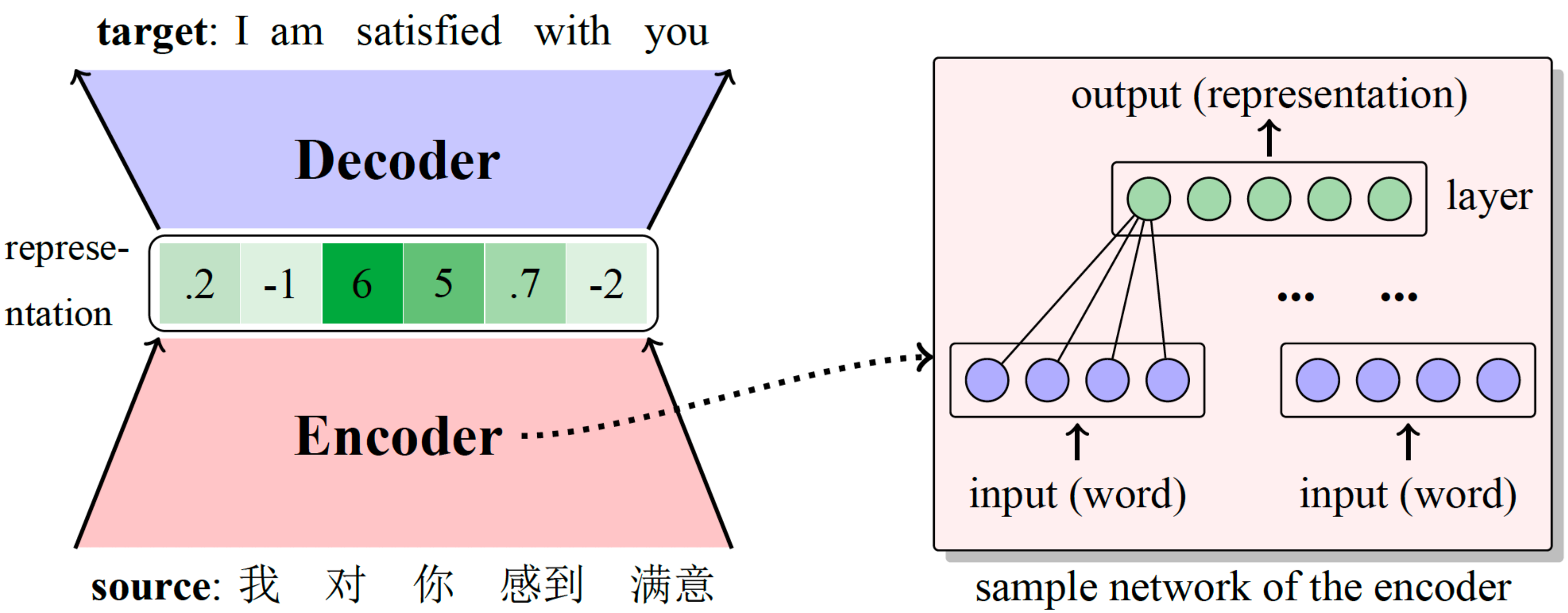
Supervision in MT

- 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.



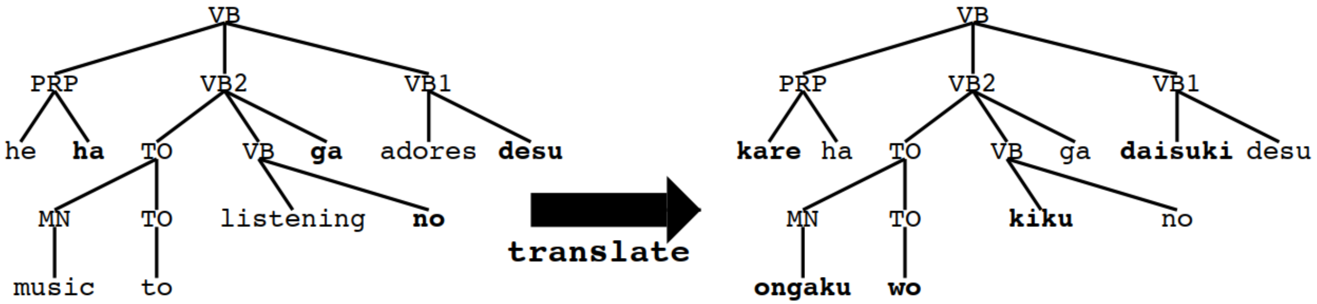
sample network of the encoder

What Is Supervision in MT

- Supervision in machine learning?
- Supervision in linguistic?
- What do you think?

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 document
  - …

Does Supervised Always Necessary?

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.

Menu

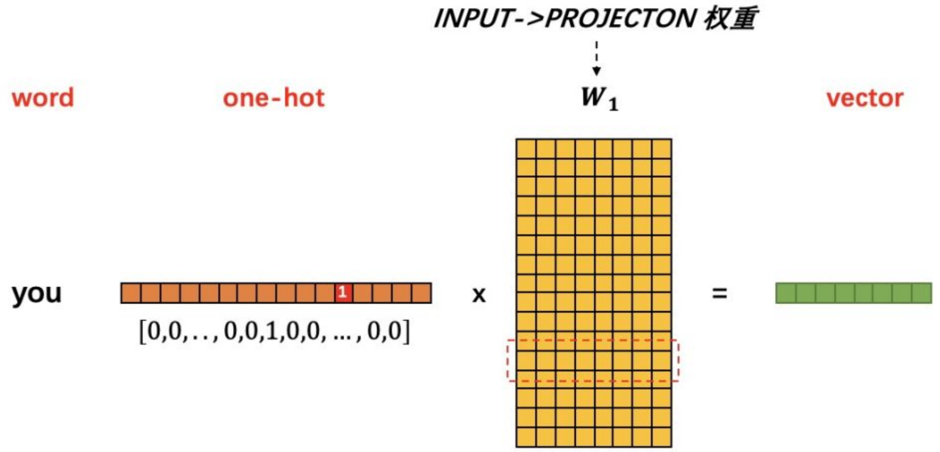
- About Me
- Background of Machine Translation (MT)
- Supervision in MT
- **Unsupervised MT**

Monolingual Word Embedding

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

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

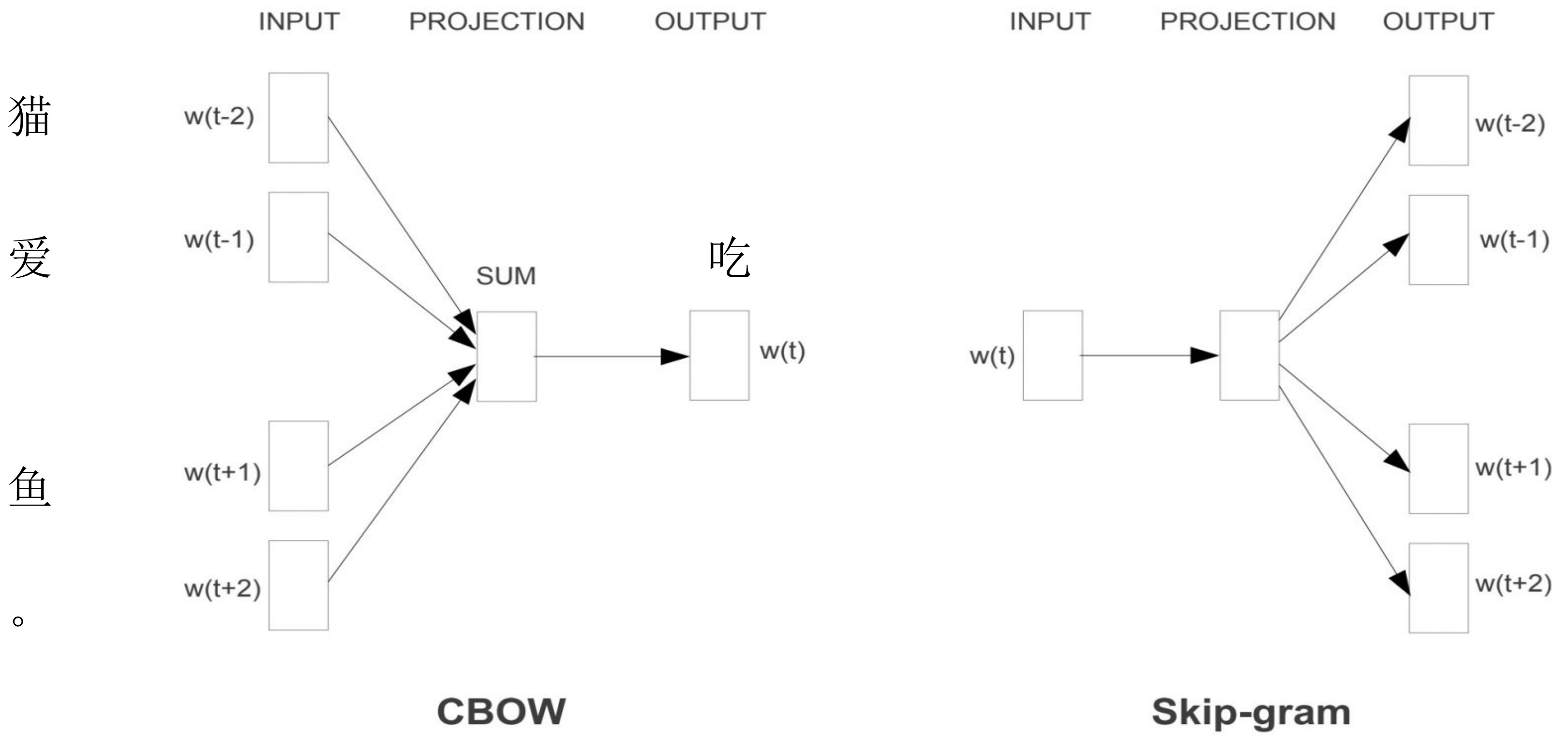
One-hot Representation



Projection

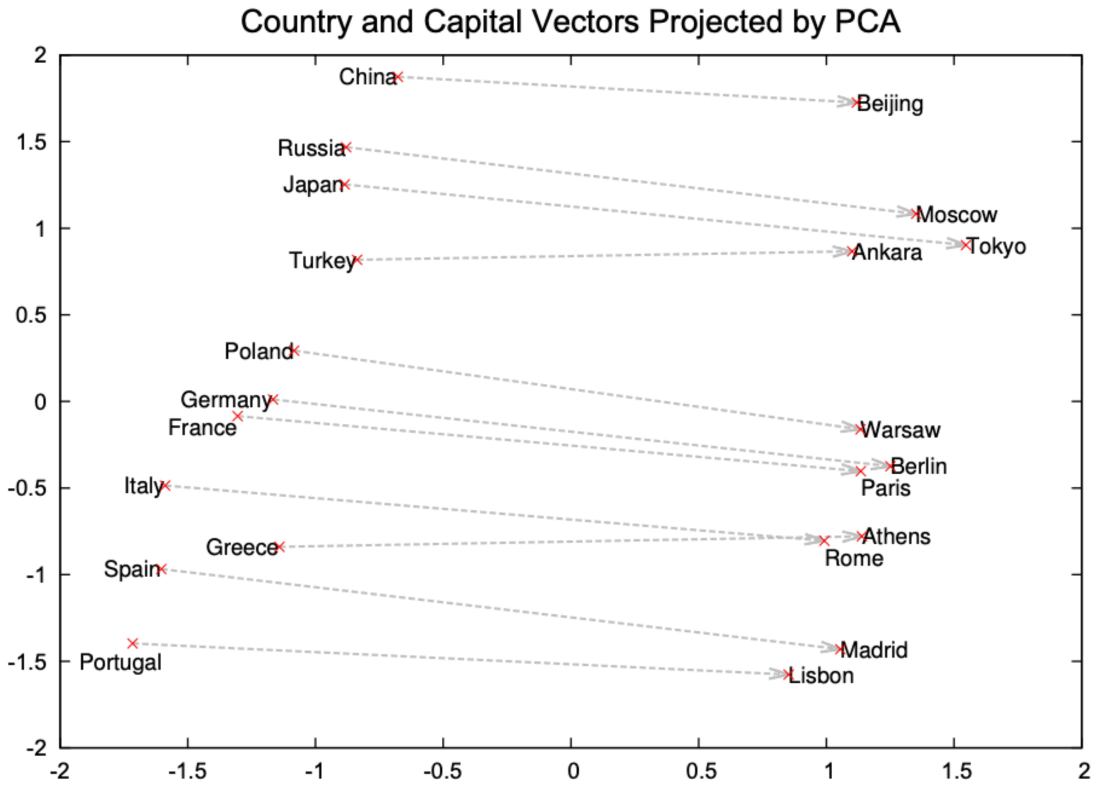
Monolingual Word Embedding

- Training Objective
- For Example:



Monolingual Word Embedding

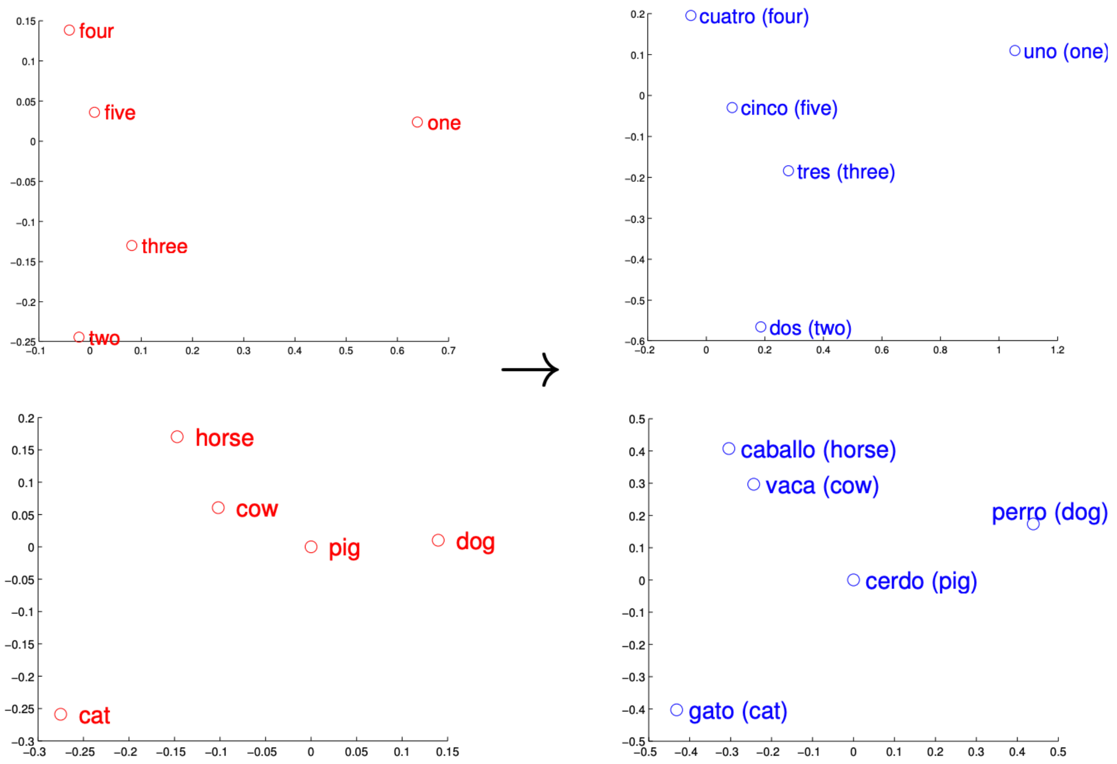
- Then, there is some interesting findings.



[Mikolov et al., NeurIPS-2013]

Bilingual Word Embedding (BWE)

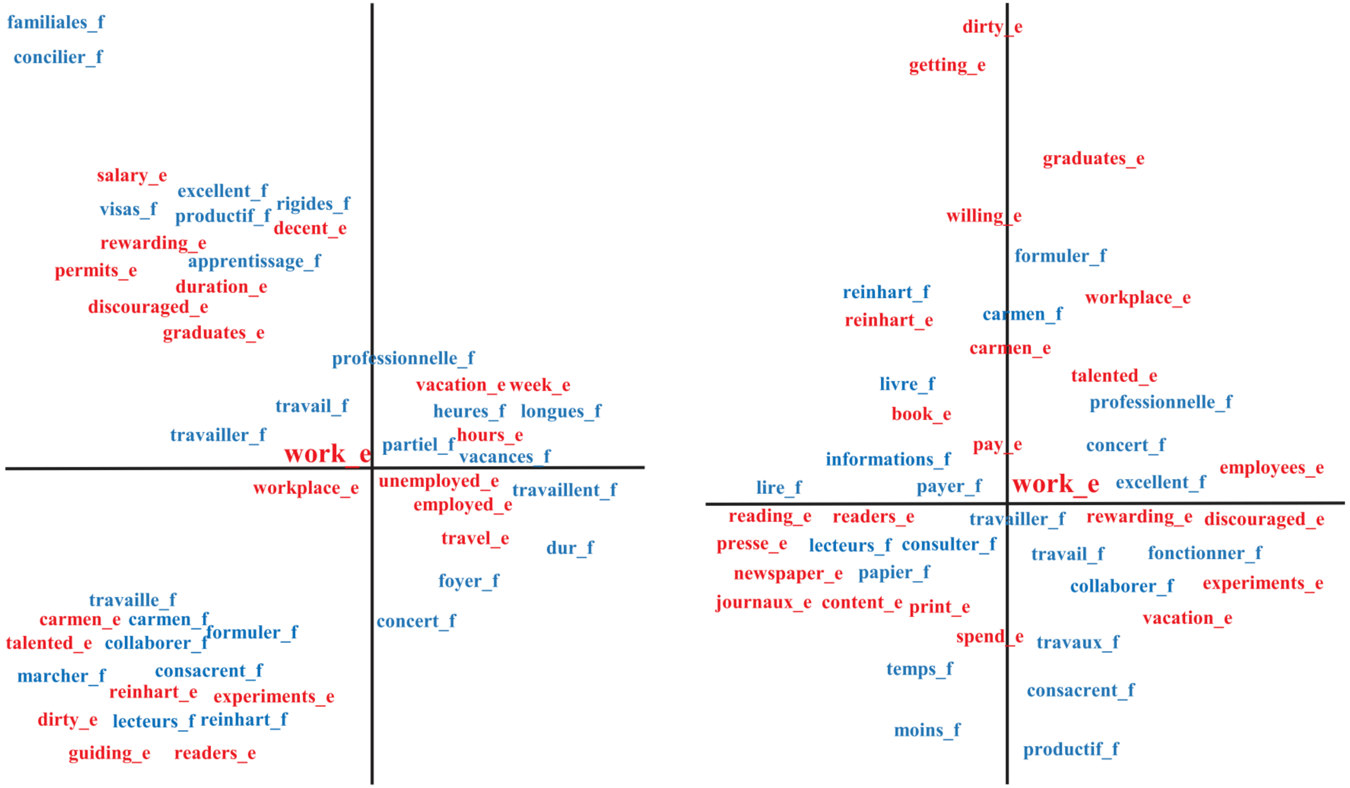
- To project one language space onto anther, researchers have to learn a translation map (matrix).
- The most typical supervision is an annotated lexicon (i.e., 5000 words).



[Mikolov et al., ArXiv-2013]

Bilingual Word Embedding (BWE)

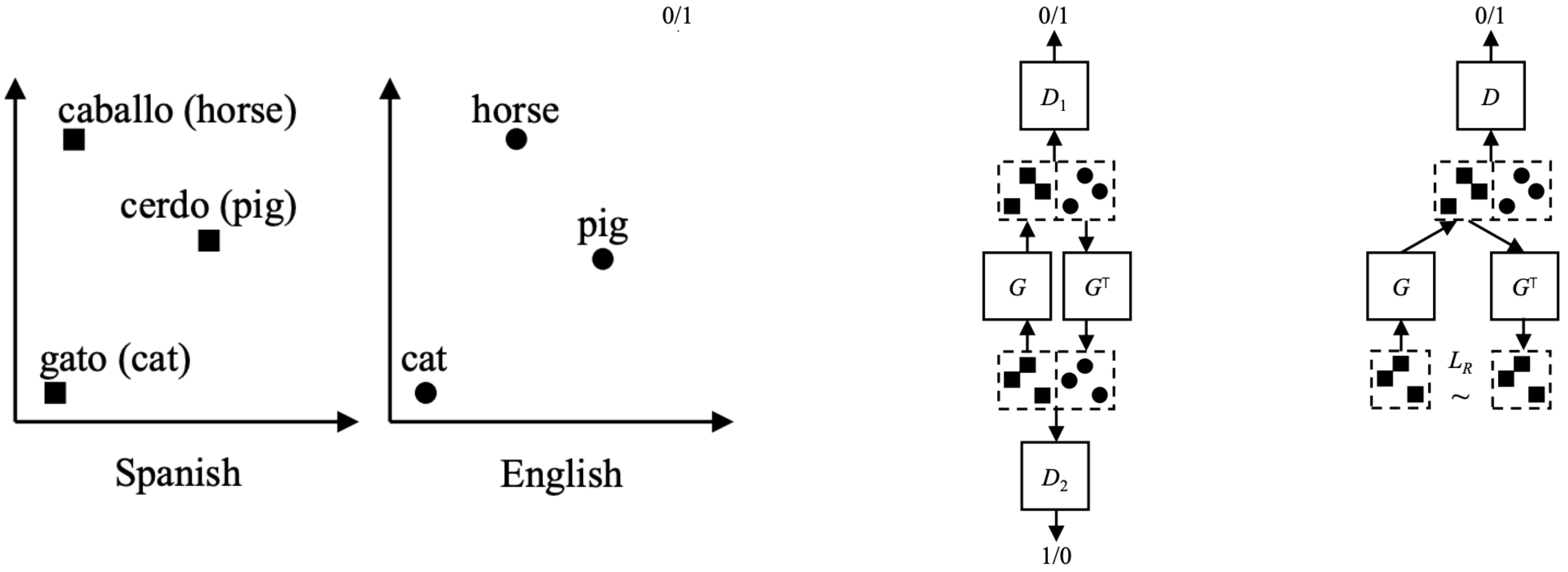
- Polysemy is not easy to project.



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

Unsupervised BWE

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



[Zhang et al., ACL-2017]

BWE Performance

- No significant difference between supervised and unsupervised BWE

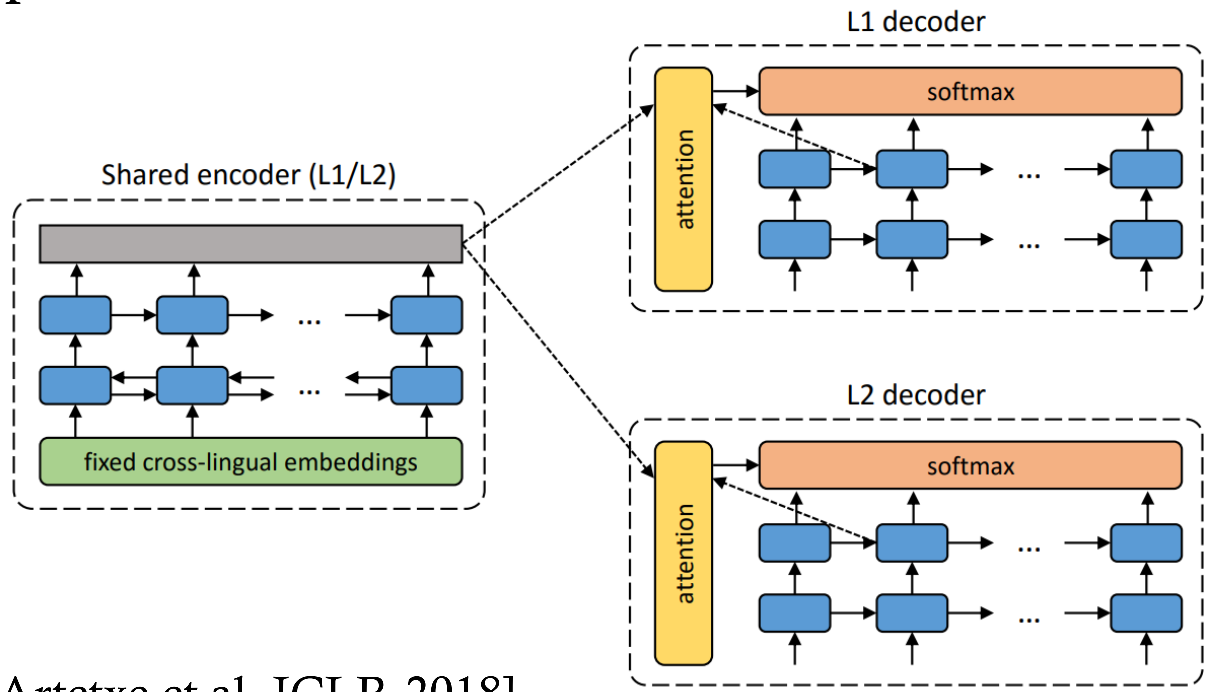
	en-de	en-fr	en-es	en-it	en-pt	de-fr	de-es	de-it	de-pt	fr-es	fr-it	fr-pt	es-it	es-pt	it-pt
Supervised methods with cross-lingual supervision															
Sup-BWE-Direct	73.5	81.1	81.4	77.3	79.9	73.3	67.7	69.5	59.1	82.6	83.2	78.1	83.5	87.3	81.0
Unsupervised methods without cross-lingual supervision															
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?

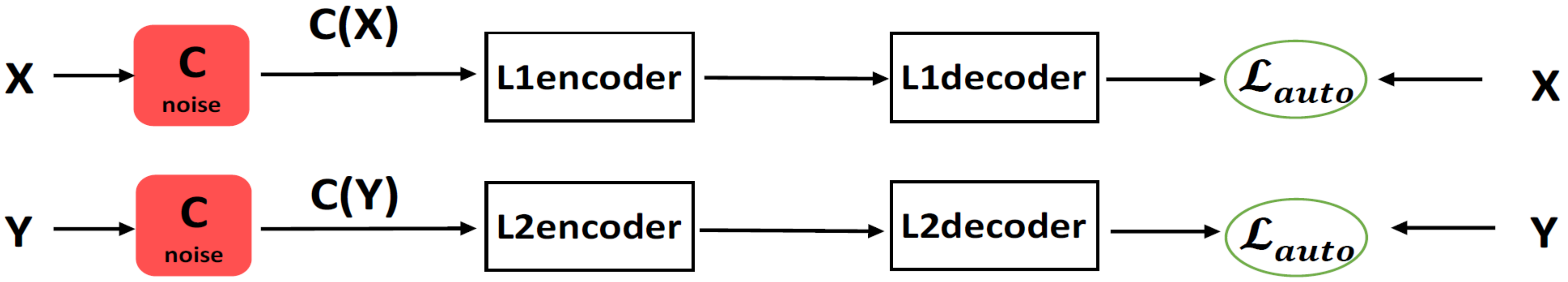
- Now we have word translation. How to conduct sentence translation?
- Initialization
  - Unsupervised bilingual word embedding
  - Cross-lingual language model
- Sharing latent representations



[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.

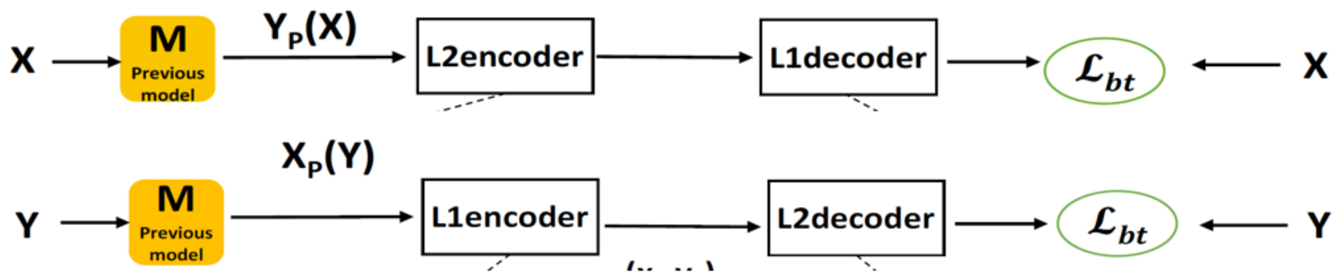


$$\mathcal{L}_D = \sum_{i=1}^{|X^1|} -log P_{L_1 \rightarrow L_1}(X_i^1 | C(X_i^1)) + \sum_{i=1}^{|X^2|} -log P_{L_2 \rightarrow L_2}(X_i^2 | C(X_i^2)),$$

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.

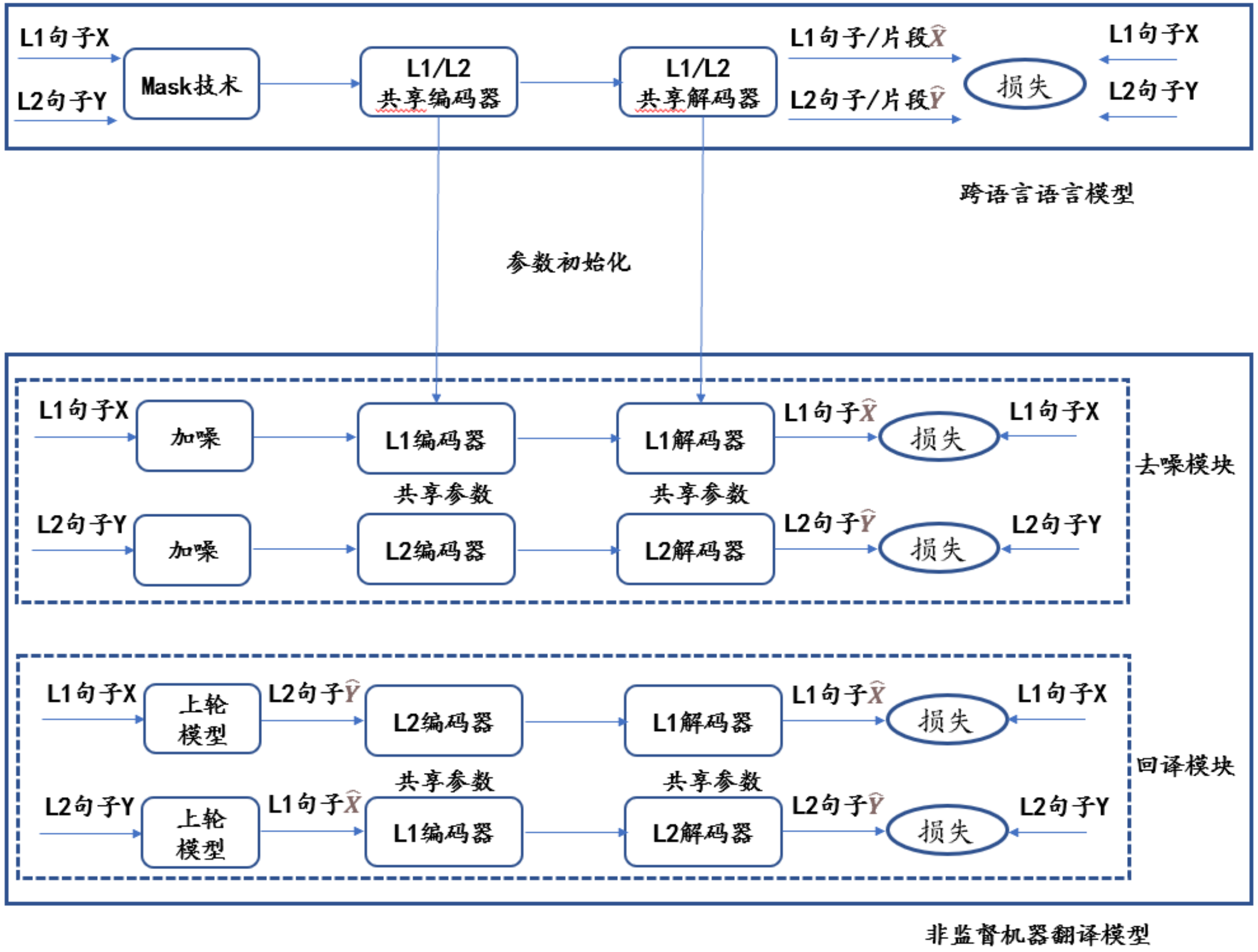
$$\mathcal{L}_B = \sum_{i=1}^{|X^1|} -log P_{L_2 \to L_1}(X_i^1 | M^2(X_i^1)) + \sum_{i=1}^{|X^2|} -log P_{L_1 \to L_2}(X_i^2 | M^1(X_i^2)),$$



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

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

Entire Structure (Sorry in Chinese)



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

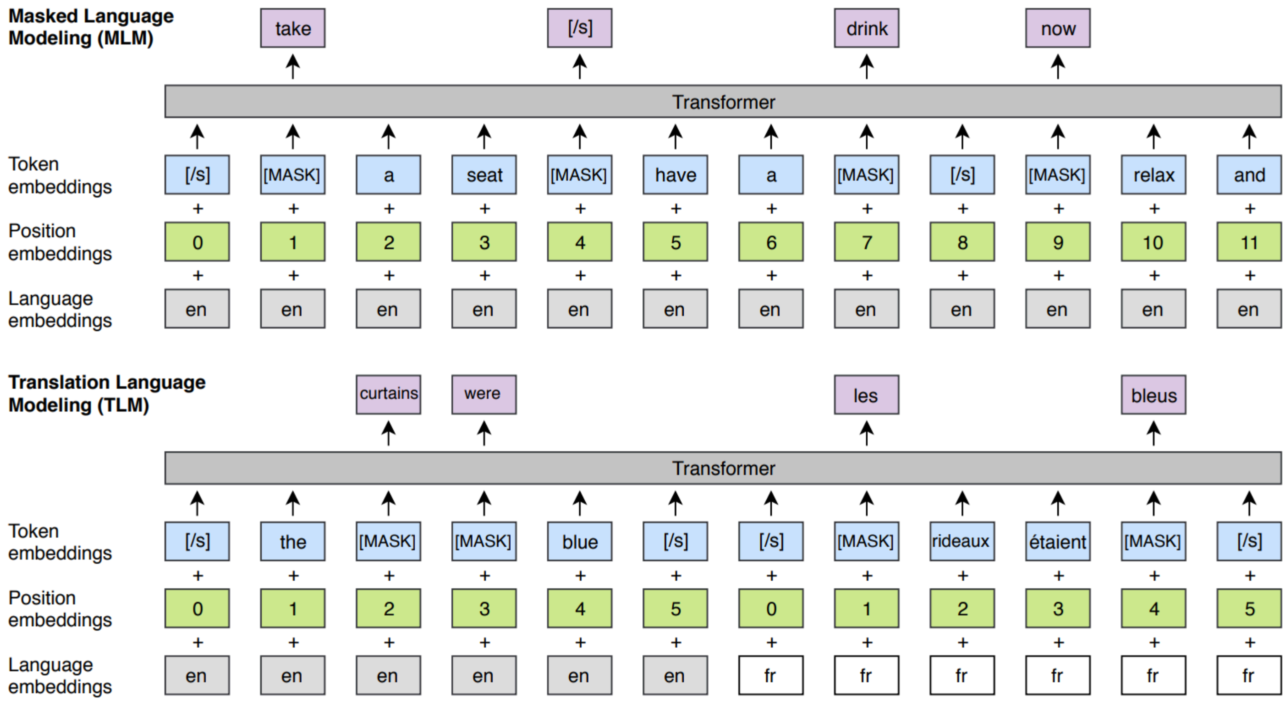
[Artetxe et al. ICLR-2018]

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.
  - Improve cross-lingual representation during UNMT training.

Better Pre-training

- Large-scale masked cross-lingual language model.

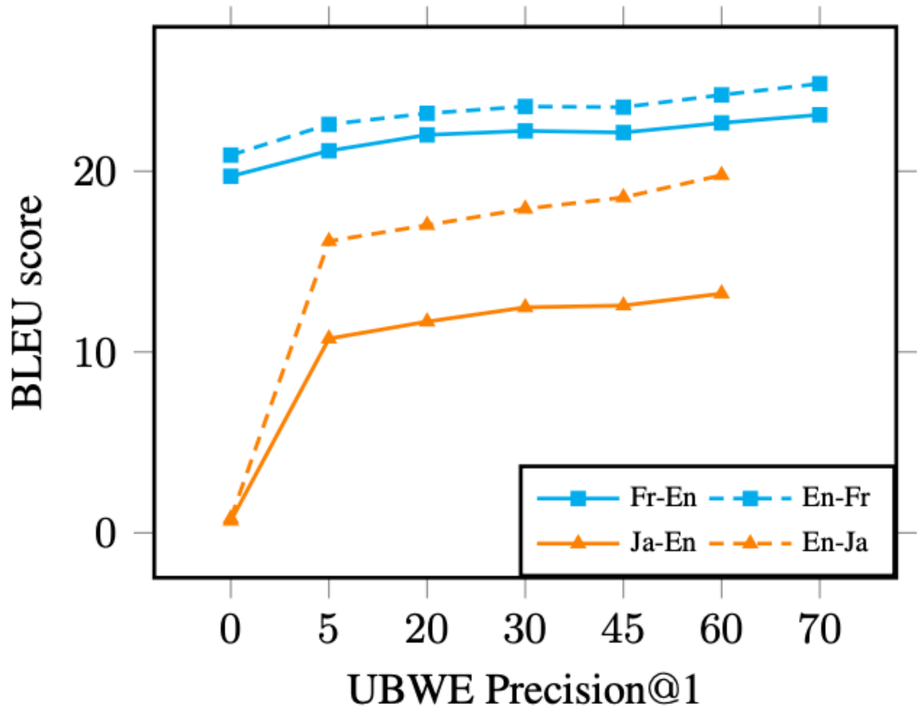


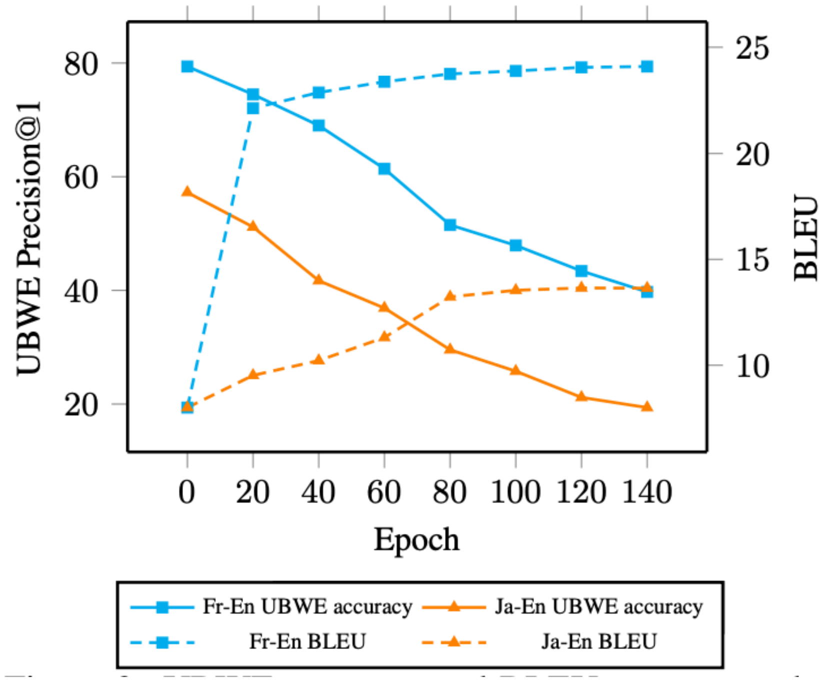
		en-fr	fr-en	en-de	de-en	en-ro	ro-en
Previous state-of-the-art - Lampe et al. (2018b)							
NMT		25.1	24.2	17.2	21.0	21.2	19.4
PBSMT		28.1	27.2	17.8	22.7	21.3	23.0
PBSMT + NMT		27.6	27.7	20.2	25.2	25.1	23.9
Our results for different encoder and decoder initializations							
EMB	EMB	29.4	29.4	21.3	27.3	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]

Better Training

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

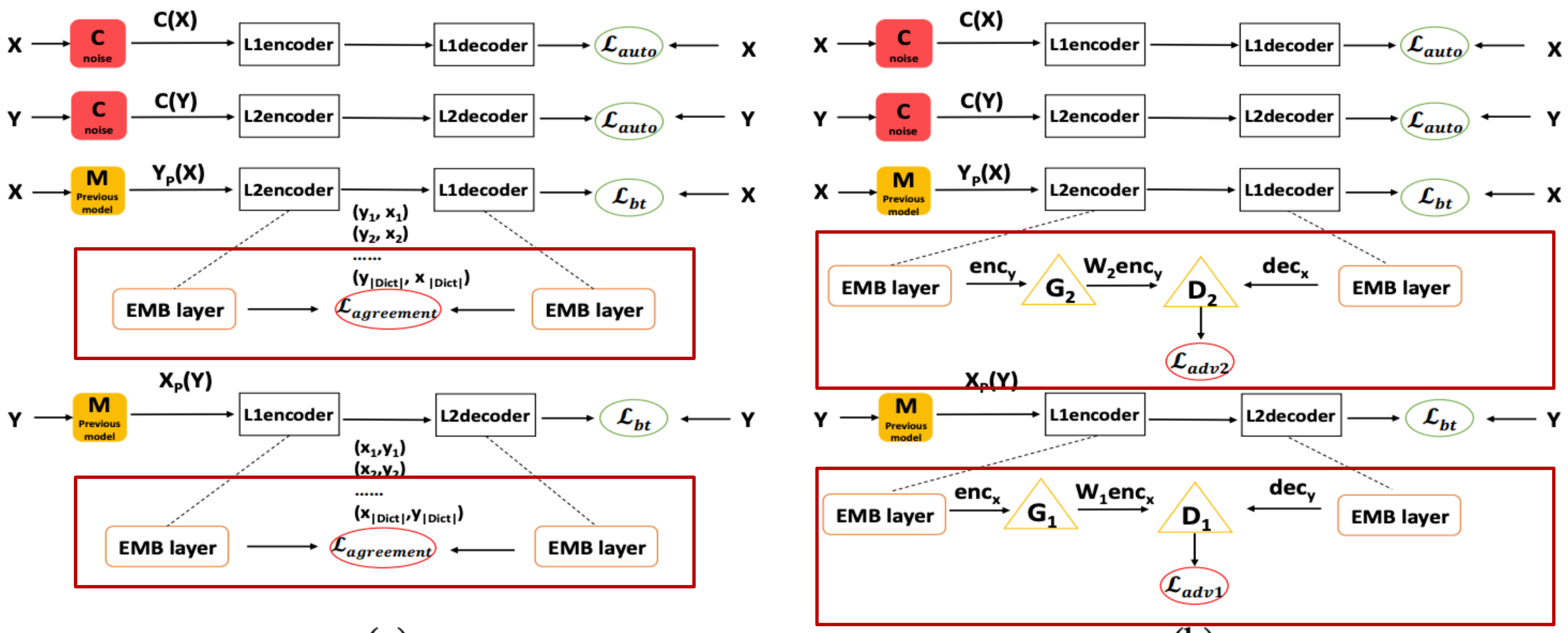




[Sun and **Wang*** et al. ACL-2019]

Joint UBWE and UNMT Training

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

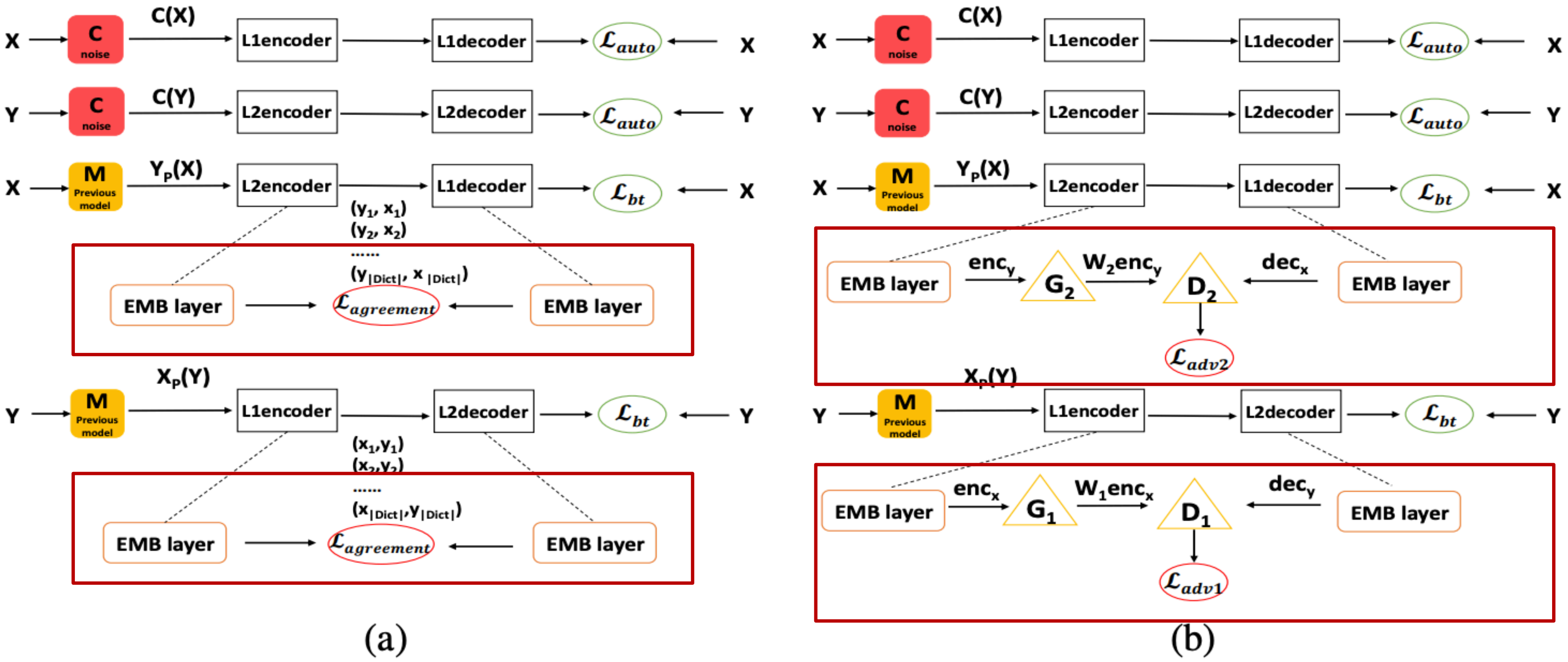


(a) (b)

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

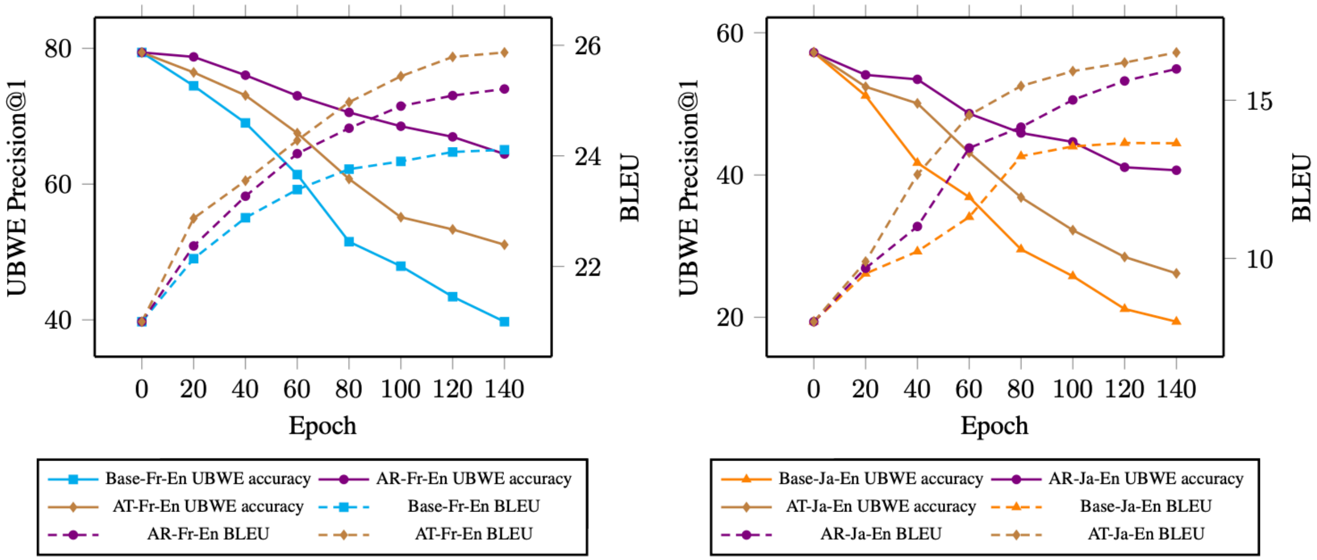
Joint UBWE and UNMT Training

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



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

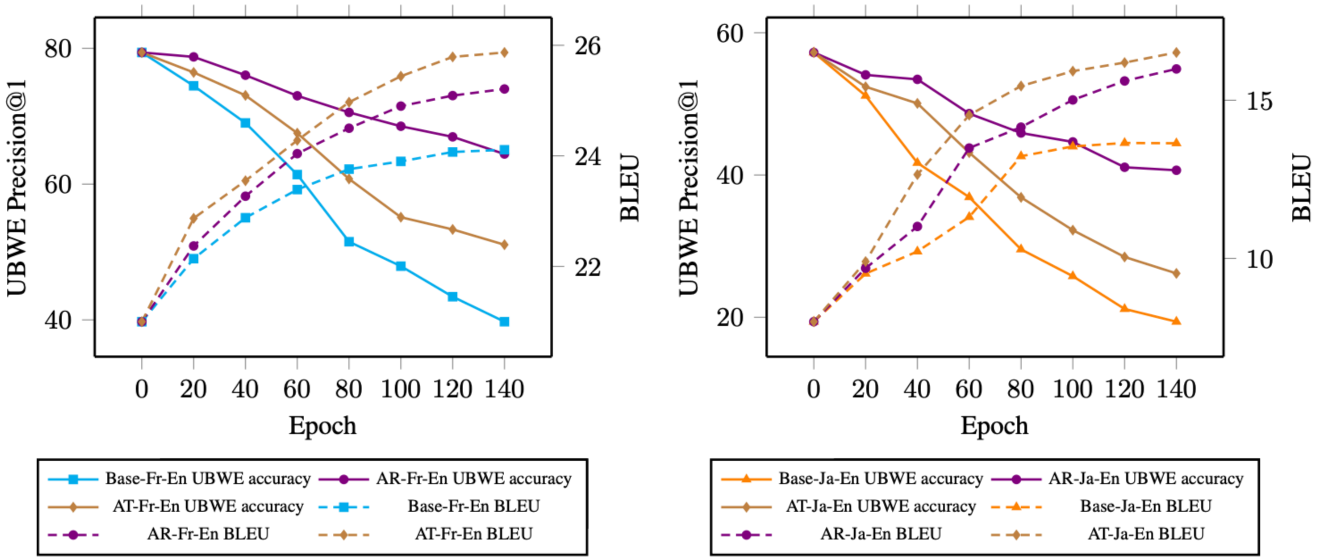
Performance: Unsupervised Translation



(a) Fr-En

(b) Ja-En

Performance: Unsupervised Translation

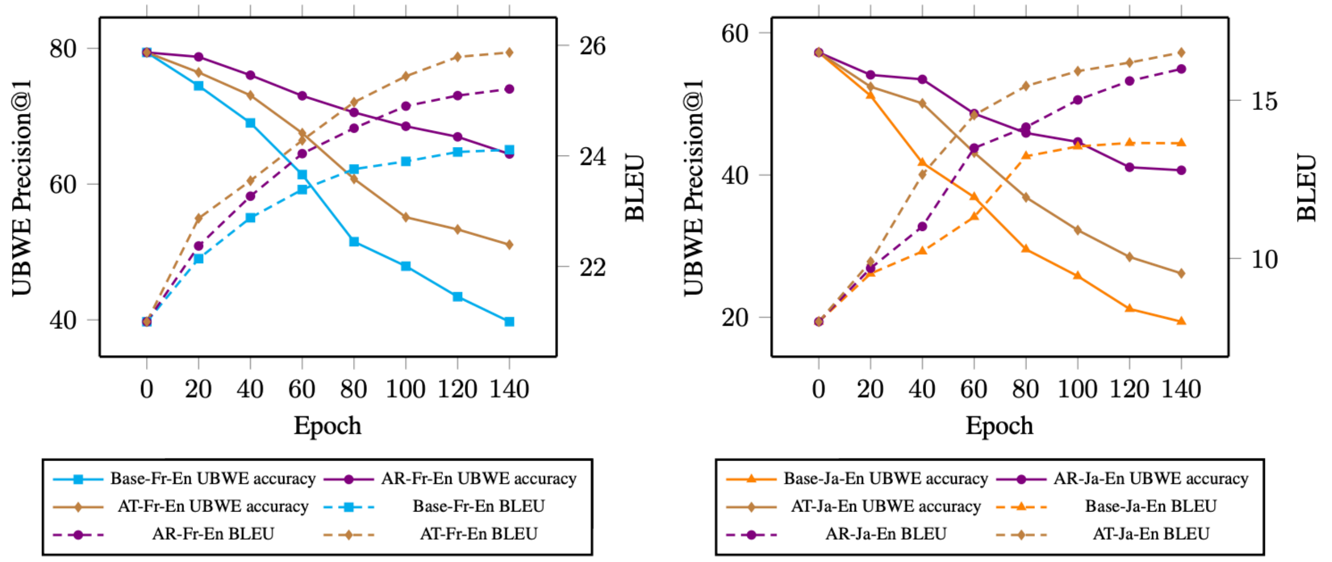


(a) Fr-En (b) Ja-En

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



(a) Fr-En

(b) Ja-En

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?

- 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 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]

What is WMT?

ACL 2019
FOURTH CONFERENCE ON
MACHINE TRANSLATION (WMT19)

August 1-2, 2019
Florence, Italy

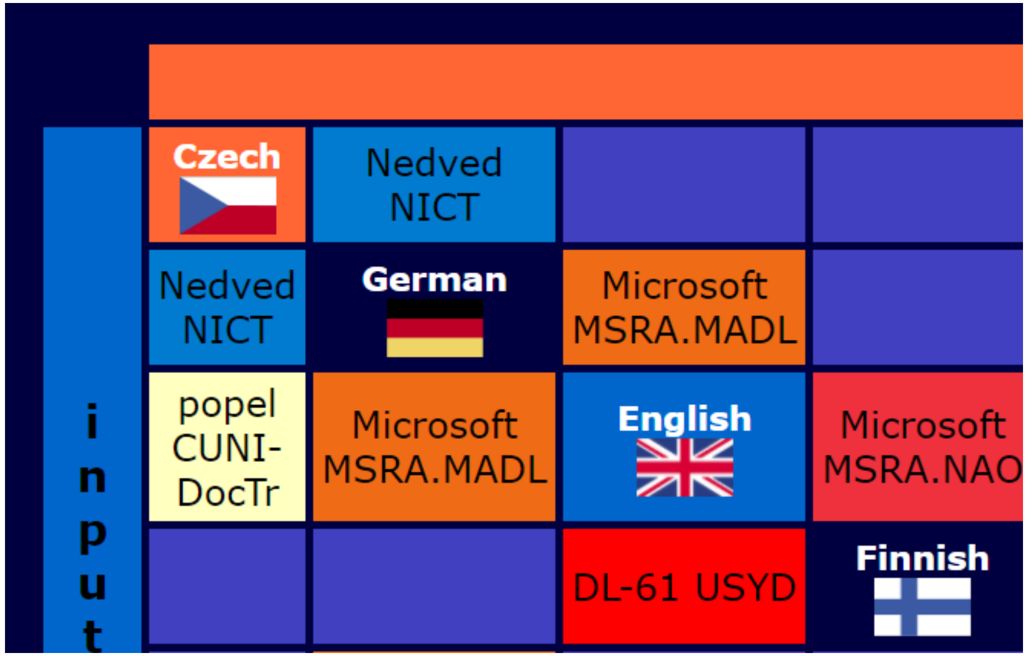
Shared Task: Machine Translation of News

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

Account

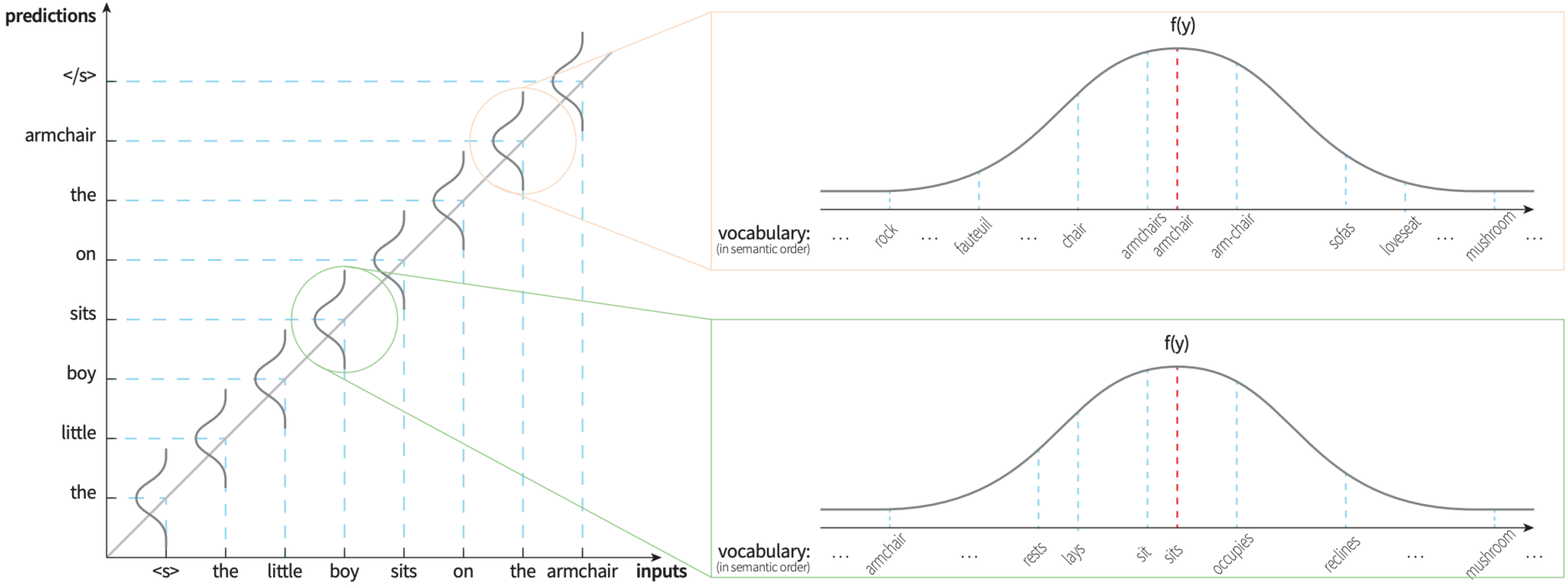
System	Submitter	System Notes	Constraint	Run Notes	BLEU	BLEU-cased	TER	BEER 2.0	CharactTER
NICT (Details)	Nedved *NICT*	Pre-trained cross-lingual LM + UNMT + USMT + pseudo SMT + pseudo NMT + fine-tuning + ensemble + USMT reranking + fixed quotes	yes		20.5	**20.1**	0.726	0.519	0.624
NICT (Details)	Nedvěd *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

Who is Nedved?



What’s More: Better Optimization

- Use the word embedding to calculate the similarity of words.
- Use this similarity as the training objective distribution.



[Li and Wang* et al., ICLR-2020, full-score paper]

State-of-the-art Performance (Till Recently)

System	EN–DE	EN–FR	EN–RO	EN–RO + STD
Vaswani et al. (2017) (base)	27.30	38.10	-	-
Vaswani et al. (2017) (big)	28.40	41.00	-	-
Transformer (base)	27.35	38.44	33.22	36.68
+ **D2GPo**	**27.93++**	**39.23++**	**34.00+**	**37.11+**
Transformer (big)	28.51	41.05	33.45	37.55
+ **D2GPo**	**29.10+**	**41.77++**	**34.13+**	**37.92+**

Supervised NMT

Method	EN–FR	FR–EN	EN–DE	DE–EN	EN–RO	RO–EN
Artetxe et al. (2017)	15.13	15.56	6.89	10.16	-	-
Lample et al. (2017)	15.05	14.31	9.75	13.33	-	-
Yang et al. (2018)	16.97	15.58	10.86	14.62	-	-
Lample et al. (2018)	25.14	24.18	17.16	21.00	21.18	19.44
XLM (Lample & Conneau, 2019)	33.40	33.30	27.00	34.30	33.30	31.80
MASS (Song et al., 2019)	37.50	34.90	28.30	35.20	35.20	33.10
MASS + **D2GPo**	**37.92**	**34.94**	**28.42**	**35.62**	**36.31**	**33.41**

Unsupervised NMT

Future Trends

- Distant Language Pairs
- Multi-Lingual UNMT
- Multi-signal (speech, vision, etc.) in UNMT

Distant Language Pairs (Sorry Chinese Again)

语言	相似语言对		远距离语言对	
	法语-英语	德语-英语	日语-英语	中文-英语
共享单词数	37, 257	43, 642	454	20, 662
共享单词所占比例	23. 30%	25. 40%	0. 18%	4. 91%
非监督机器翻译性能（BLEU）	27. 6	25. 1	14. 1	8. 02

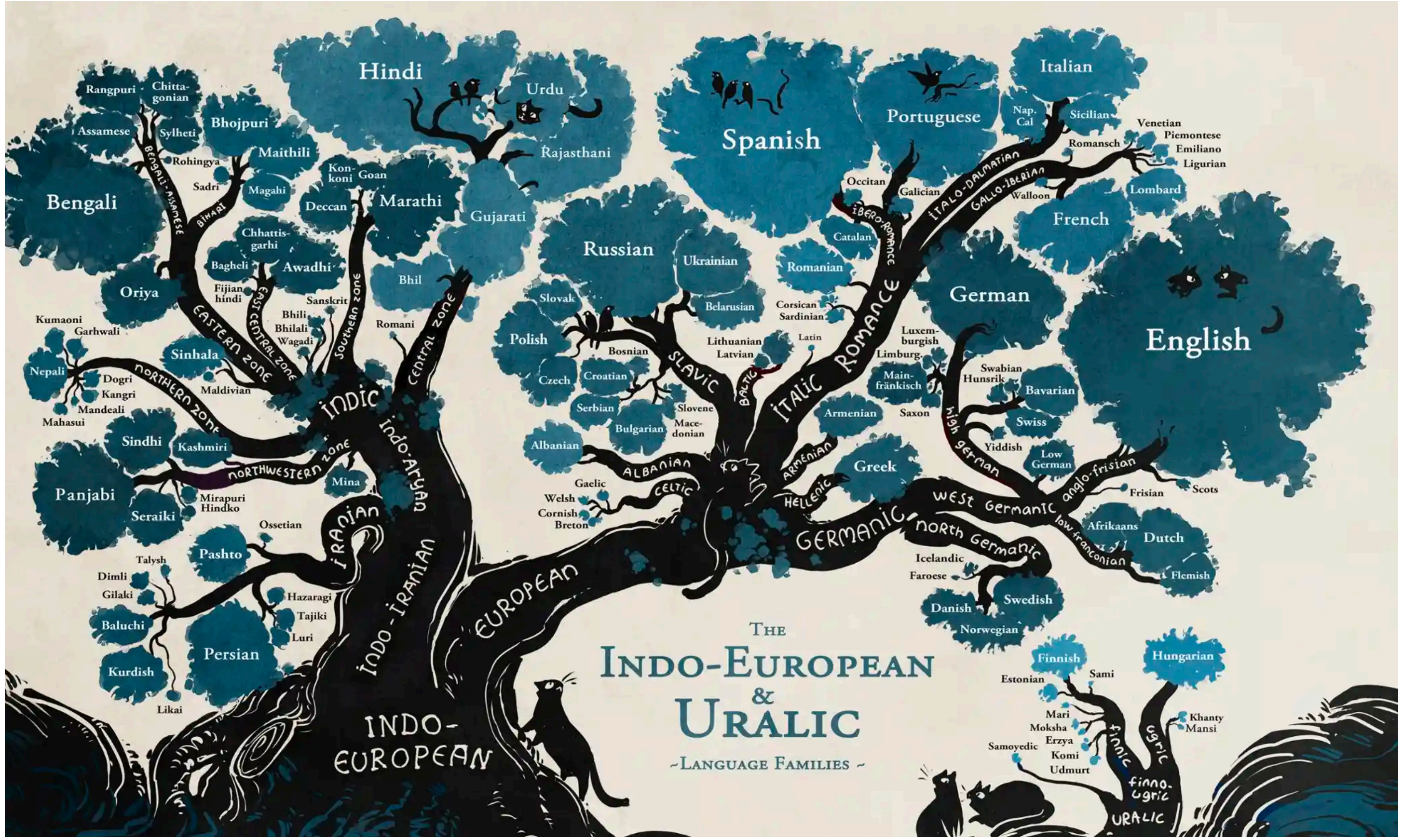
表 2：不同语言对的共享单词数据统计

语言	相似语言对		远距离语言对	
	法语-英语	德语-英语	日语-英语	中文-英语
语序相似度	76. 3%	78. 1%	53. 4%	62. 2%
监督机器翻译性能（BLEU）	40. 2	35. 0	30. 9	26. 4
非监督机器翻译性能（BLEU）	27. 6	25. 1	14. 1	8. 02

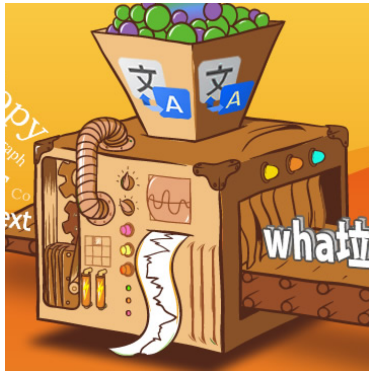
表 3：不同语言对的语序相似度统计

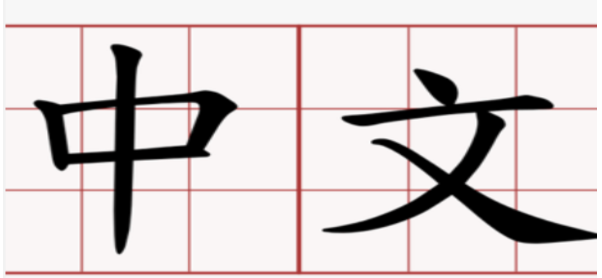
Multi-Lingual Unsupervised Translation

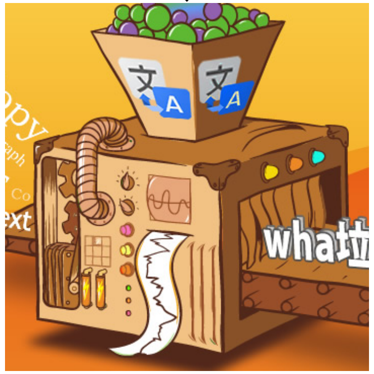
- Challenge
  - There are many language families and groups in the world.
  - The language within certain language families can help each other.



Bilingual & Multi-Lingual Translation

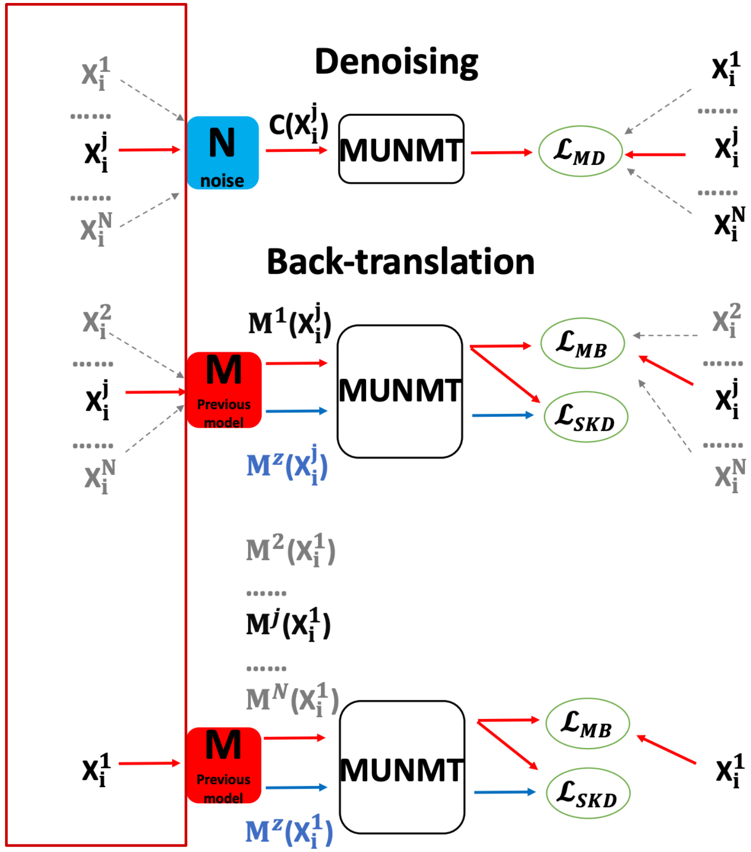




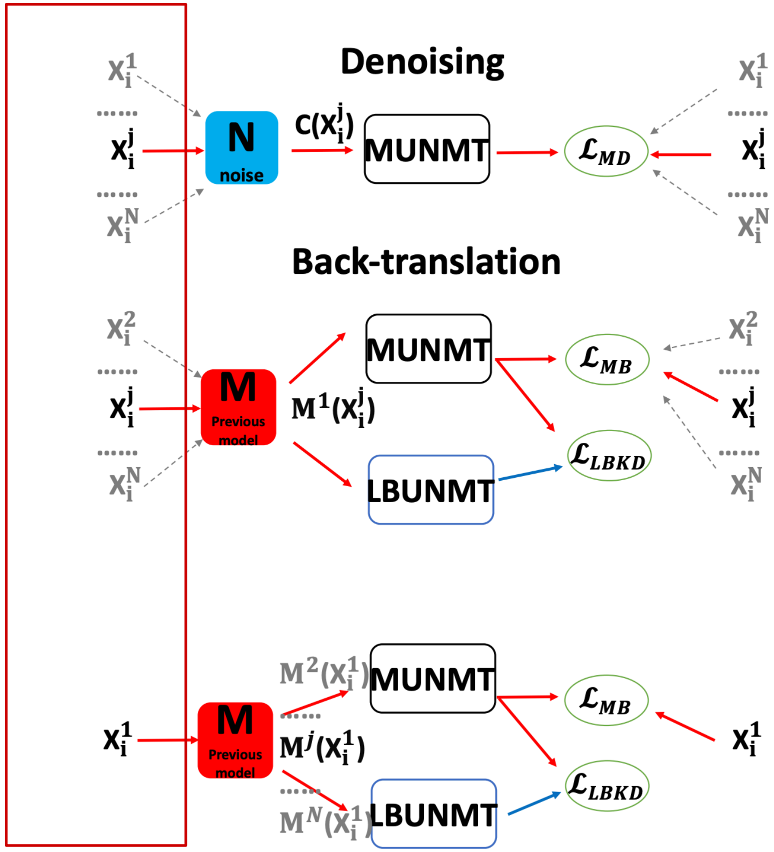


Multi-Lingual UNMT

- Contribution
  - We proposed multi-lingual UNMT.



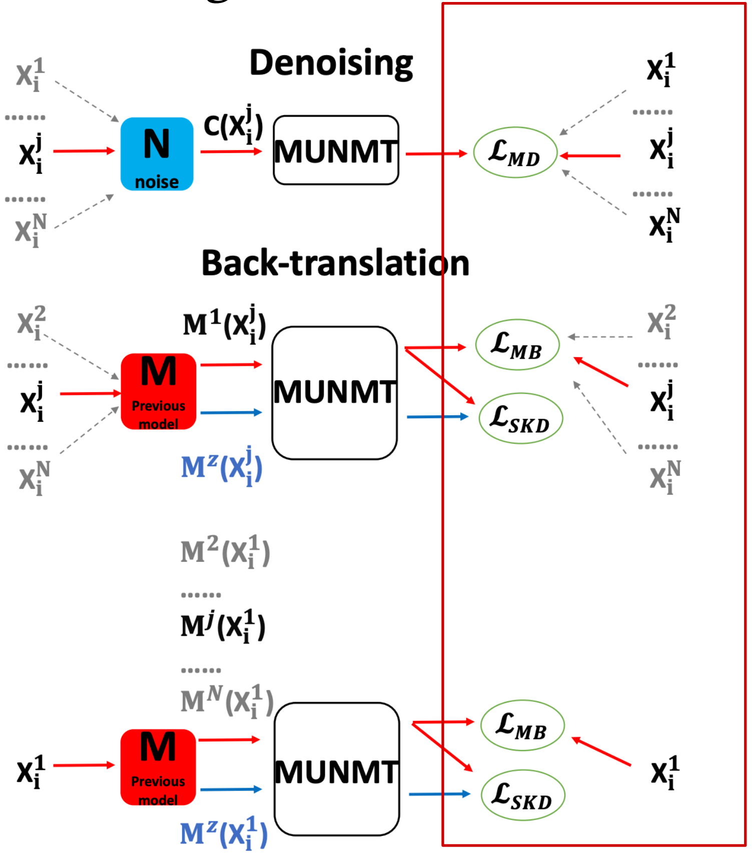
All languages



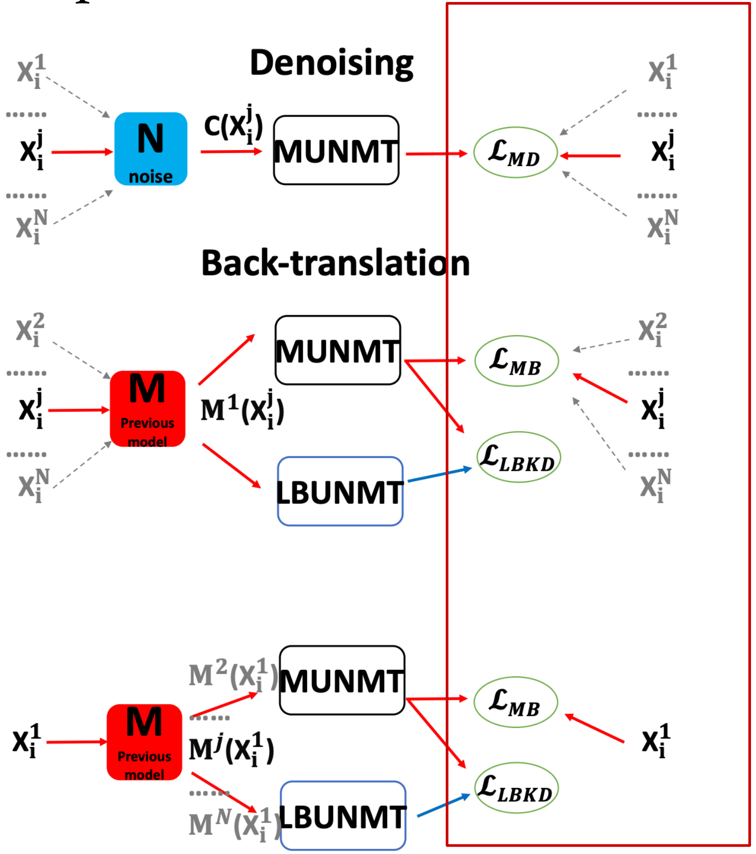
Languages in the same brunch

Multi-Lingual UNMT

- Contribution
  - We proposed multi-lingual UNMT.
  - We use knowledge distillation to enhance UNMT performance.



All languages



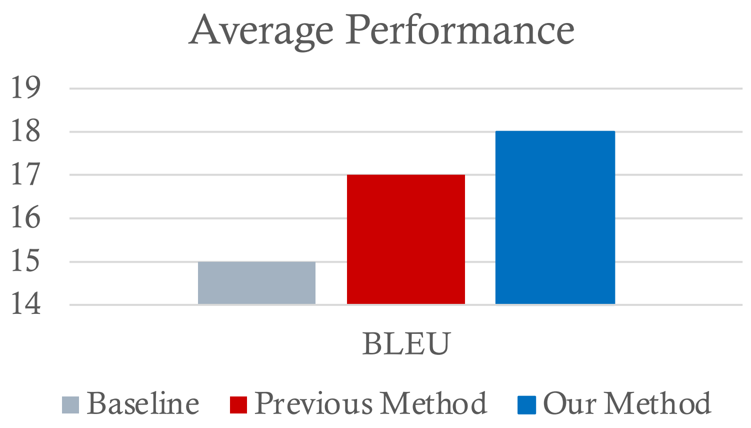
Languages in the same brunch

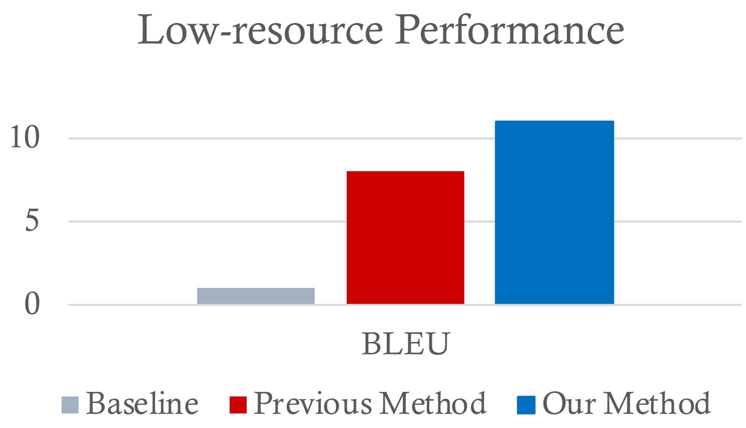
Performance: Multi-Lingual Translation

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**

Our Method

Low Resource

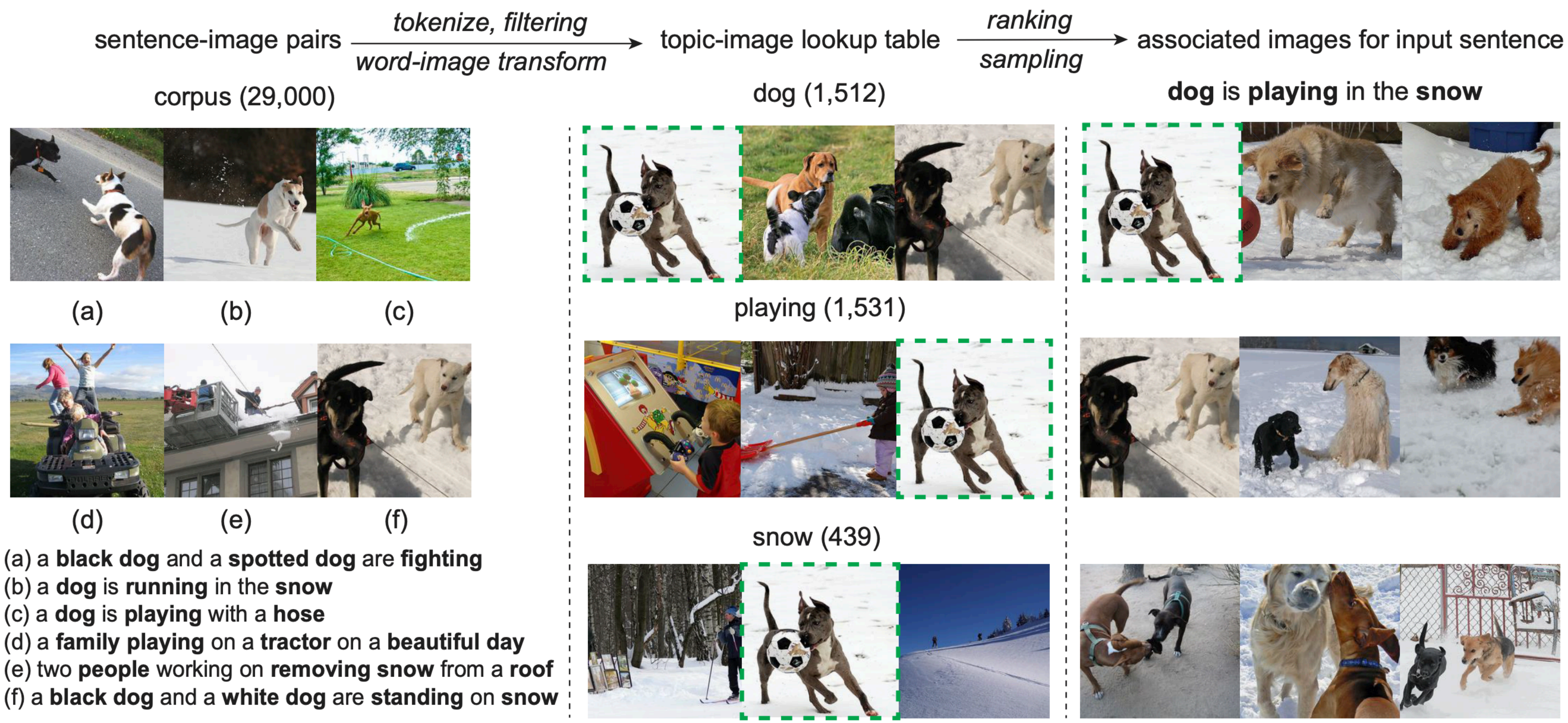




(Sun and **Wang*** et al. ACL-2020)

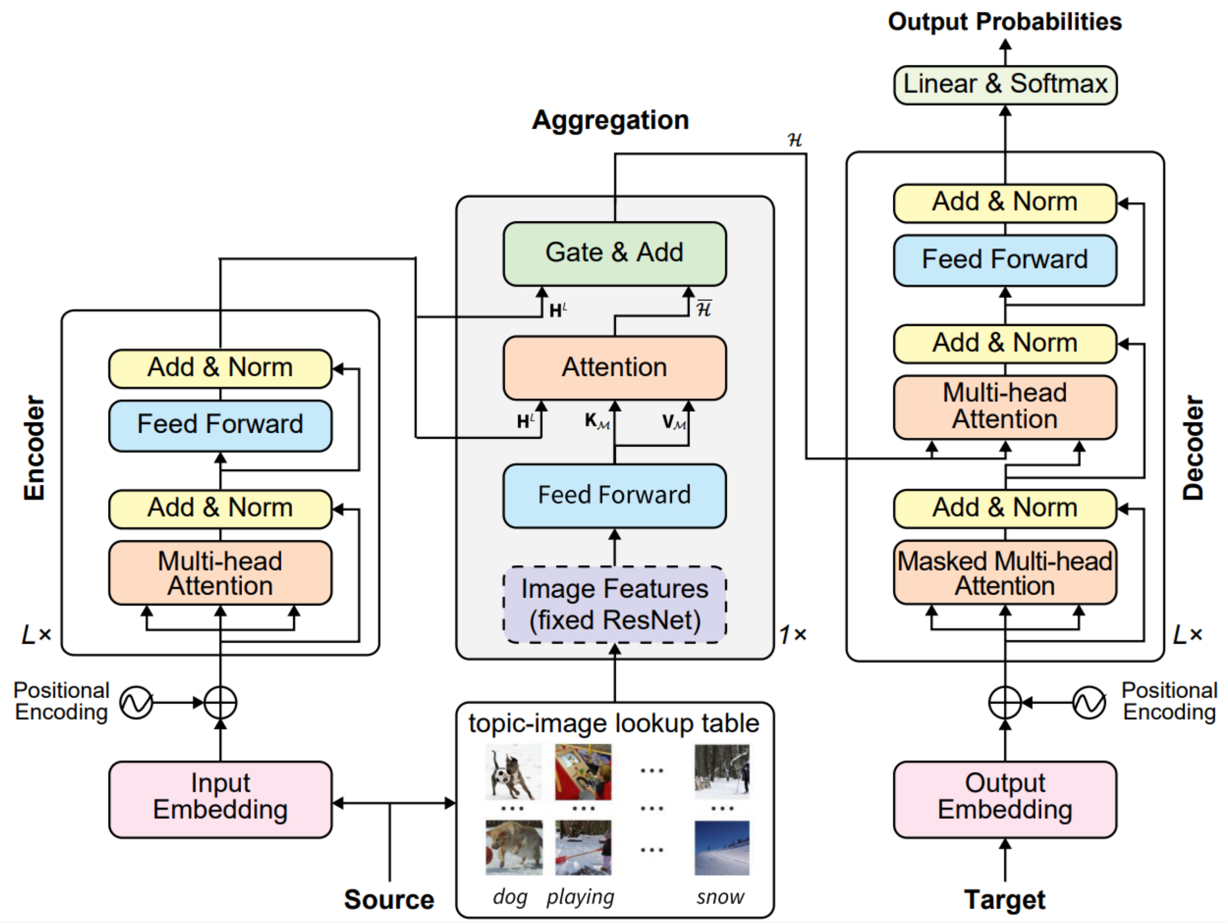
Modeling Visual Information

- No only language information, but also visual, speech information etc., can be modeled in UNMT.



[Zhang and Wang* et al., ICLR-2020]

Modeling Visual Information



System	Architecture	EN-RO		EN-DE		EN-FR	
		BLEU	#Param	BLEU	#Param	BLEU	#Param
Existing NMT systems							
Vaswani et al. (2017)	Trans. (base)	N/A	N/A	27.3	N/A	38.1	N/A
	Trans. (big)	N/A	N/A	28.4	N/A	41.0	N/A
Lee et al. (2018)	Trans. (base)	32.40	N/A	24.57	N/A	N/A	N/A
Our NMT systems							
This work	Trans. (base)	32.66	61.54M	27.31	63.44M	38.52	63.83M
	+VR	**33.78++**	63.04M	**28.14++**	64.94M	**39.64++**	65.33M
	Trans. (big)	33.85	207.02M	28.45	210.88M	41.10	211.66M
	+VR	**34.46+**	211.02M	**29.14++**	214.89M	**41.83+**	215.66M

Table 1: Results on EN-RO, EN-DE, and EN-FR for the NMT tasks. Trans. is short for transformer. N/A denotes that those numbers are not reported in the corresponding literature. "++/+" after the BLEU score indicate that the proposed method was significantly better than the corresponding baseline Transformer (base or big) at significance level p<0.01/0.05.

Conclusion

- My understanding
  - Supervision in linguistic is always necessary.
  - Supervision in machine learning is not always necessary.

- Welcome to join us to work on MT!
  - https://wangruinlp.github.io/
  - wangrui.nlp@gmail.com

Thank You!