Survey on Transfer Learning for Neural Machine Translation

Wang Yong, Longyue

5, June 2019

Content

- Multi-lingual NMT
- Domain Adaptation NMT
- Conclusion

Multi-lingual NMT

Develop one model for translation between all possible languages by effective use of available linguistic resources.

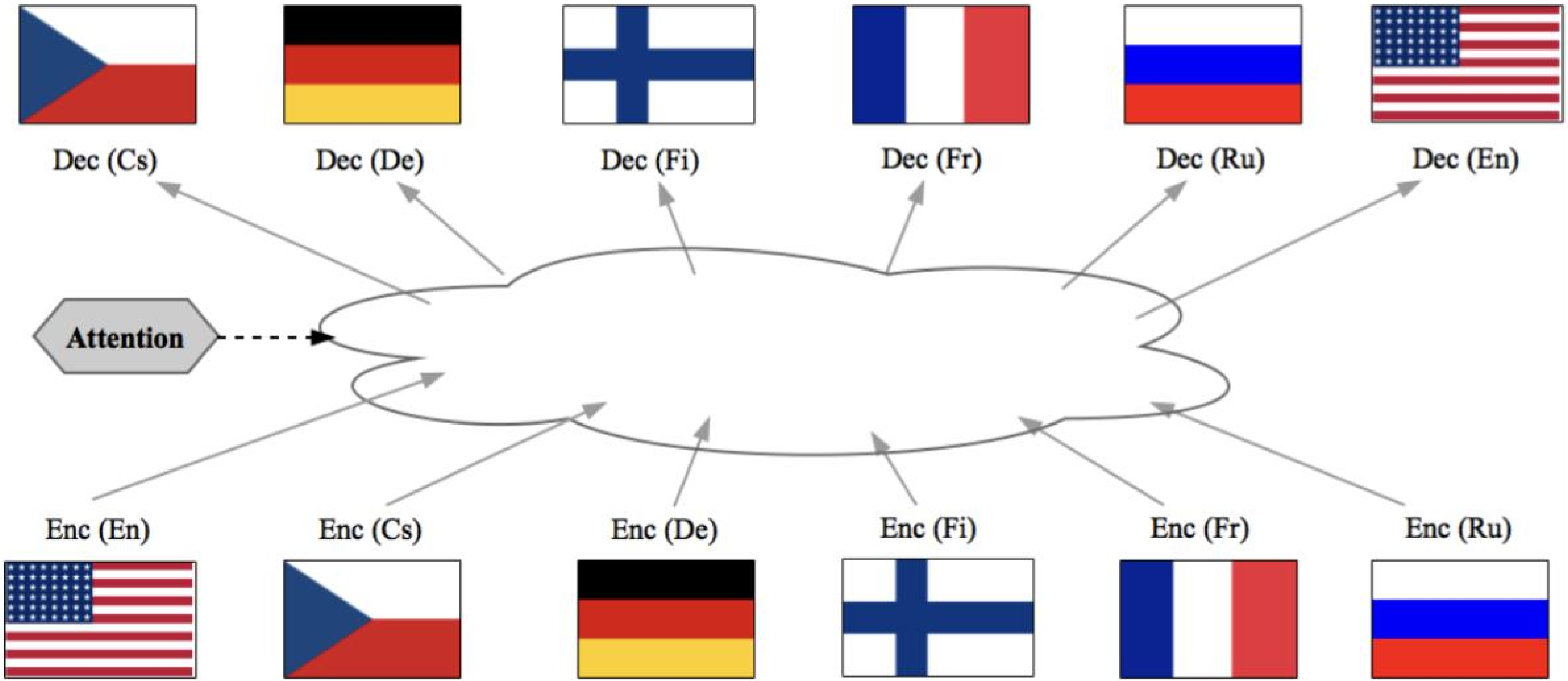
Language pairs:

$$(s_i, t_i) \in L \quad (i \in \{1, ..., l\}) \qquad L \subset S \times T$$

S, T are sets of source and target languages

respectively.

Multi-lingual NMT



Multi-lingual NMT

Benefits:

1. Knowledge transfer (Cross-lingual knowledge)
2. Compact (One model)
3. Interlingua (Universal representation)

Scenarios:

1. Multiway Translation
2. Low or Zero-Resource Translation: Pivot and Zero-shot
3. Multi-Source Translation

Multiway NMT

1. Separate encoder/decoder (Firat et al. (2016a); Firat et al. (2016b)).
2. Universal encoder/decoder
   Shared/Non-shared vocab: Johnson et al. (2017); Ha et al. (2016),
   Capacity bottleneck: Aharoni et al. (2019),
   Lexical similarity: Lee et al. (2017); Nguyen et al. (2017); Wang et al. (2019),
   Architecture comparison: Lakew et al. (2018a).
3. Partial Parameter Sharing
   Sharing strategies: Sachan et al. (2018),
   Routing network: Zaremoodi et al. (2018),
   Parameter generation: Platanios et al. (2018).
4. Training Strategy
   Joint training,
   Knowledge distillation: Tan et al. (2019).

Low or Zero-Resource NMT

1. Low-resource NMT

    Training strategy: Finetune (Zoph et al. (2016)), Meta-learning (Gu et al. (2018b)),

    Language relatedness: Zoph et al. (2016), Neubig et al. (2018),

    Lexical-level transfer: Nguyen et al. (2017), Gu et al. (2018a), Murthy et al. (2018).

2. Pivoting NMT

    Run-time pivoting: Firat et al. (2016a),

    Pivoting during training: Firat et al. (2016b) , Cheng et al. (2017), Chen et al. (2017).

3. Zero-shot NMT

    Training strategy: Johnson et al. (2017), Lakew et al. (2017), Arivazhagan et al. (2018),

$$D^{a\to b} = \{(X_1^a, Y_1^b), ..., (X_N^a, Y_N^b)\} \quad |D^{a\to b}| = 0 \quad |D^{a\to c}| > 0 \quad |D^{c\to b}| > 0$$

    Corpus size: Aharoni et al. (2019),

$$\log p(Y^b|\hat{X}^c) \quad \hat{X}^c = \arg\max_X \log p(X^c|X^a)$$

Ha et al. (2017).

MNMT paper1

Massively Multilingual Neural Machine Translation (NAACL-2019)

Supervised performance:

	Ar-En	En-Ar	Fr-En	En-Fr	Ru-En	En-Ru	Uk-En	En-Uk	Avg.
5-to-5	**23.87**	**12.42**	**38.99**	**37.3**	29.07	**24.86**	**26.17**	16.48	**26.14**
25-to-25	23.43	11.77	38.87	36.79	**29.36**	23.24	25.81	**17.17**	25.8
50-to-50	23.7	11.65	37.81	35.83	29.22	21.95	26.02	15.32	25.18
75-to-75	22.23	10.69	37.97	34.35	28.55	20.7	25.89	14.59	24.37
103-to-103	21.16	10.25	35.91	34.42	27.25	19.9	24.53	13.89	23.41

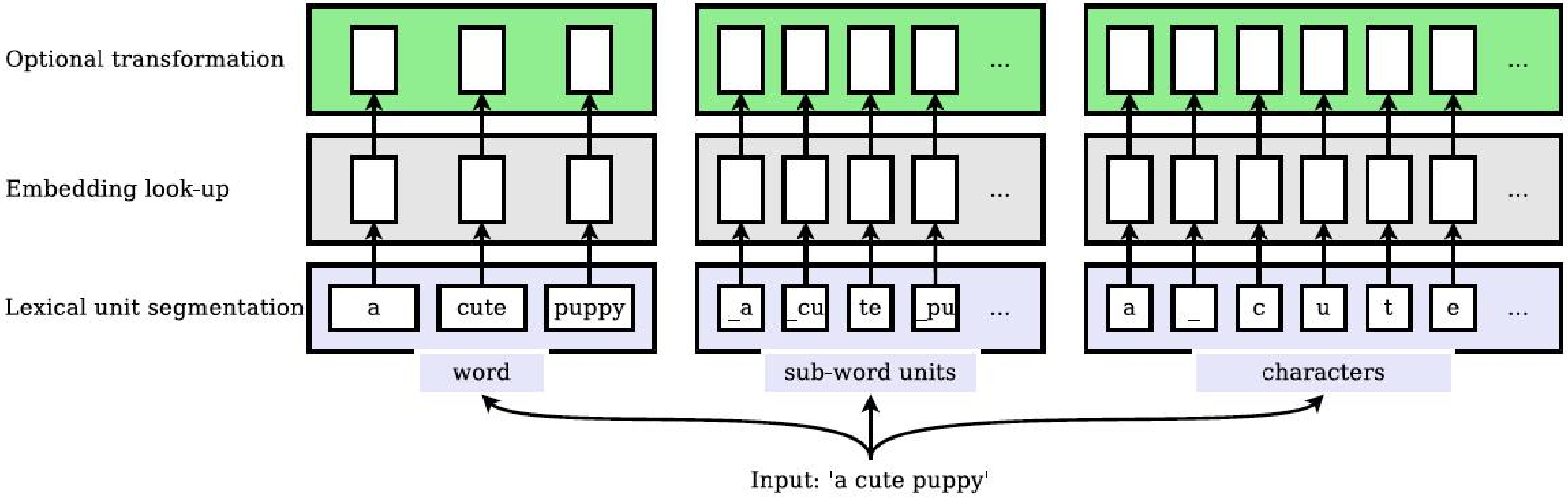
Zero-Shot performance:

	Ar-Fr	Fr-Ar	Ru-Uk	Uk-Ru	Avg.
5-to-5	1.66	4.49	3.7	3.02	3.21
25-to-25	1.83	**5.52**	**16.67**	4.31	7.08
50-to-50	**4.34**	4.72	15.14	**20.23**	**11.1**
75-to-75	1.85	4.26	11.2	15.88	8.3
103-to-103	2.87	3.05	12.3	18.49	9.17

Problem: Capacity bottleneck

MNMT paper2

Multilingual Neural Machine Translation With Soft Decoupled Encoding (ICLR-2019)

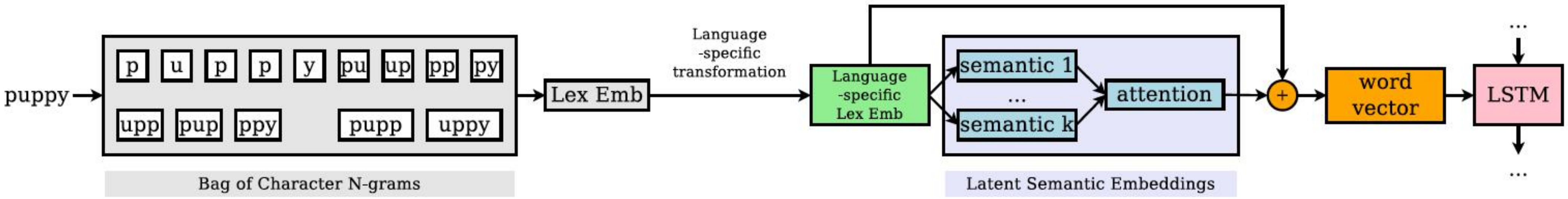


Word-based: unbounded, independent parameters for same concepts
Character-based: slow training, large pressure on model
Subword-based: sub-optimal for MNMT

Desiderata: accurately represent words in all of the languages under considerati
maximize the sharing of parameters across languages

MNMT paper2

Multilingual Neural Machine Translation With Soft Decoupled Encoding (ICLR-2019)



Two stages: (1) Modeling the language-specific spelling of the word,
(2) Modeling the language-agnostic semantics of the word.

Three steps: (1) Lexical Embedding,
(2) Language-specific Transformation,
(3) Latent Semantic Embedding.

MNMT paper2

Multilingual Neural Machine Translation With Soft Decoupled Encoding (ICLR-2019)

LRL	Train	Dev	Test	HRL	Train
aze	5.94k	671	903	tur	182k
bel	4.51k	248	664	rus	208k
glg	10.0k	682	1007	por	185k
slk	61.5k	2271	2445	ces	103k

LRL and HRL mean Low-Resource and High-Resource Language

Lex Unit	Model	aze	bel	glg	slk
Word	Lookup	7.66	13.03	28.65	25.24
Sub-joint	Lookup	9.40	11.72	22.67	24.97
Sub-sep	Lookup (Neubig & Hu, 2018)[2]	10.90	16.17	28.10	28.50
Sub-sep	UniEnc (Gu et al., 2018)[3]	4.80	8.13	14.58	12.09
Word	SDE	11.82*	18.71*	30.30*	28.77†

MNMT paper3

Effective Cross-lingual Transfer of Neural Machine Translation Models without Shared Vocabularies (ACL-2019)

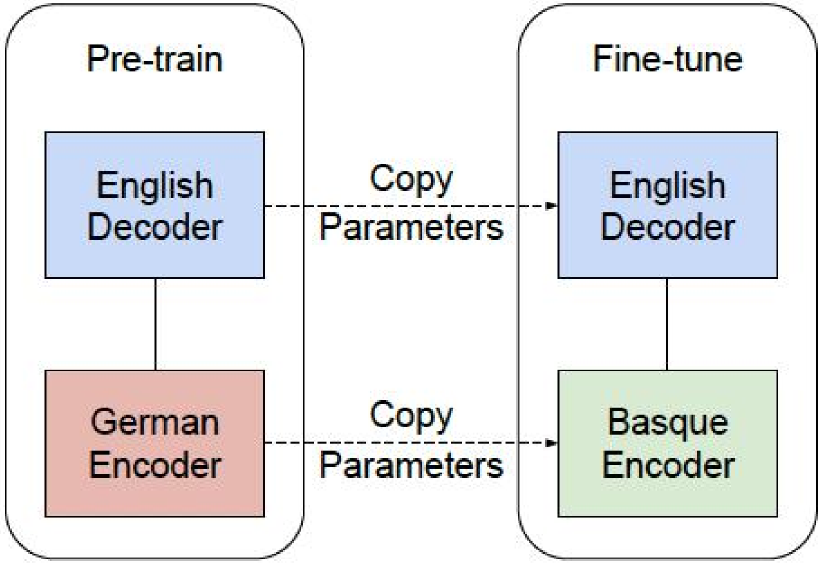
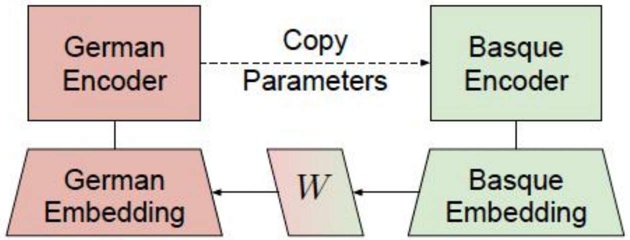
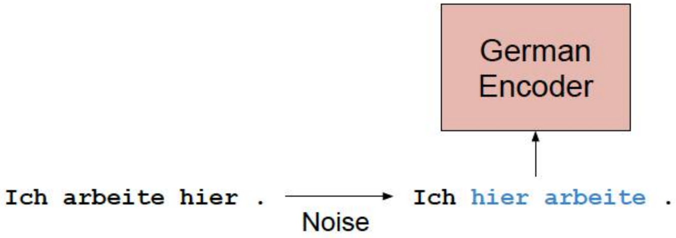


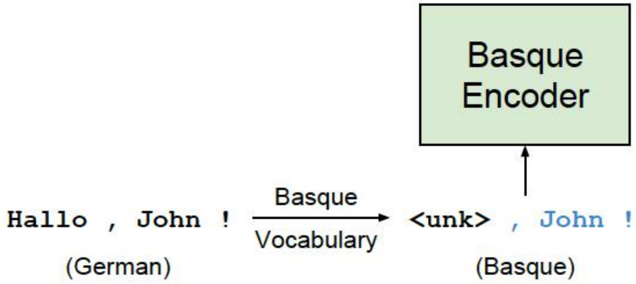
Diagram of transfer learning for NMT from
German => English to Basque => English.

MNMT paper3

Effective Cross-lingual Transfer of Neural Machine Translation Models without Shared Vocabularies (ACL-2019)







Cross-lingual Word Embedding:

$$E^{\text{src}}_{\text{parent}} \leftarrow W E^{\text{mono}}_{\text{child}}$$

Artificial Noises

Synthetic Data from Parent Model Training Data

MNMT paper3

Effective Cross-lingual Transfer of Neural Machine Translation Models without Shared Vocabularies (ACL-2019)

Family	Source Language	Data (→English) [#sents]
Germanic	German	10,111,758
Isolate	Basque	5,605
Slavic	Slovenian	17,103
	Belarusian	4,509
Turkic	Azerbaijani	5,946
	Turkish	9,998

	BLEU [%]				
System	eu-en	sl-en	be-en	az-en	tr-en
Baseline	1.7	10.1	3.2	3.1	0.8
Multilingual (Johnson et al., 2017)	5.1	16.7	4.2	4.5	8.7
Transfer (Zoph et al., 2016)	4.9	19.2	8.9	5.3	7.4
+ Cross-lingual word embedding	7.4	20.6	12.2	7.4	9.4
+ Artificial noises	8.2	21.3	12.8	8.1	10.1
+ Synthetic data	**9.7**	**22.1**	**14.0**	**9.0**	**11.3**

Summary on MNMT

- Problems

  (1) Cross-lingual transfer without shared vocabularies, including unsupervised machine translation.

  (2) Capacity bottleneck: high-resource deficiency and low-resource efficiency,

  (3) Zero-shot cannot beat pivot-based methods.

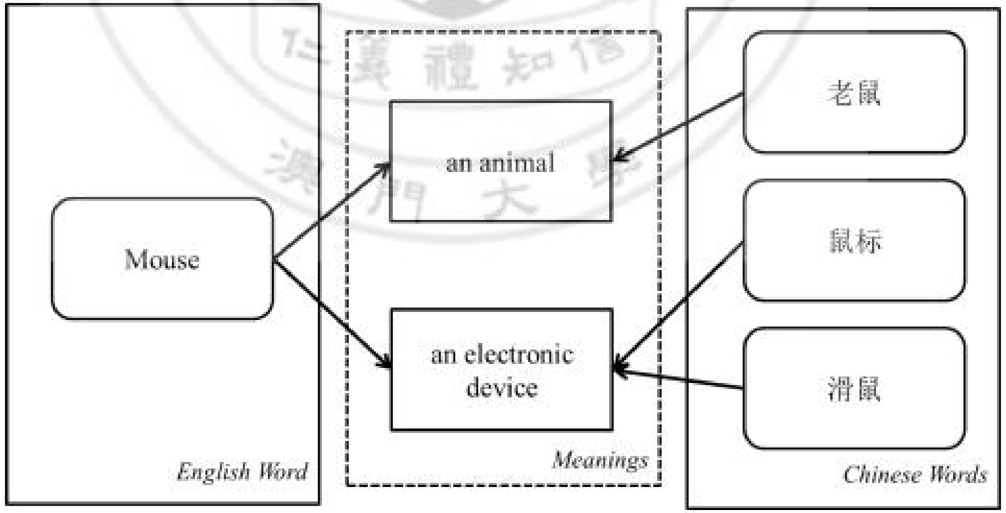
- Methods

  (1) Can denoising autoencoder help problem without shared vocabularies,

  (2) Have more efficient strategy to handle capacity bottleneck?

Domain Adaptation NMT

- [Domain] Different domains may vary by topic or text style.
- [Domain Adaptation] A mismatch between the domain for which training data are available and the target domain of a machine translation system.



When an international treaty that relates to a contract and **which** the People's Republic of China has concluded on participated into has provisions of the said treaty shall be applied, but with the exception of clauses to which the People's Republic of China has declared reservation.

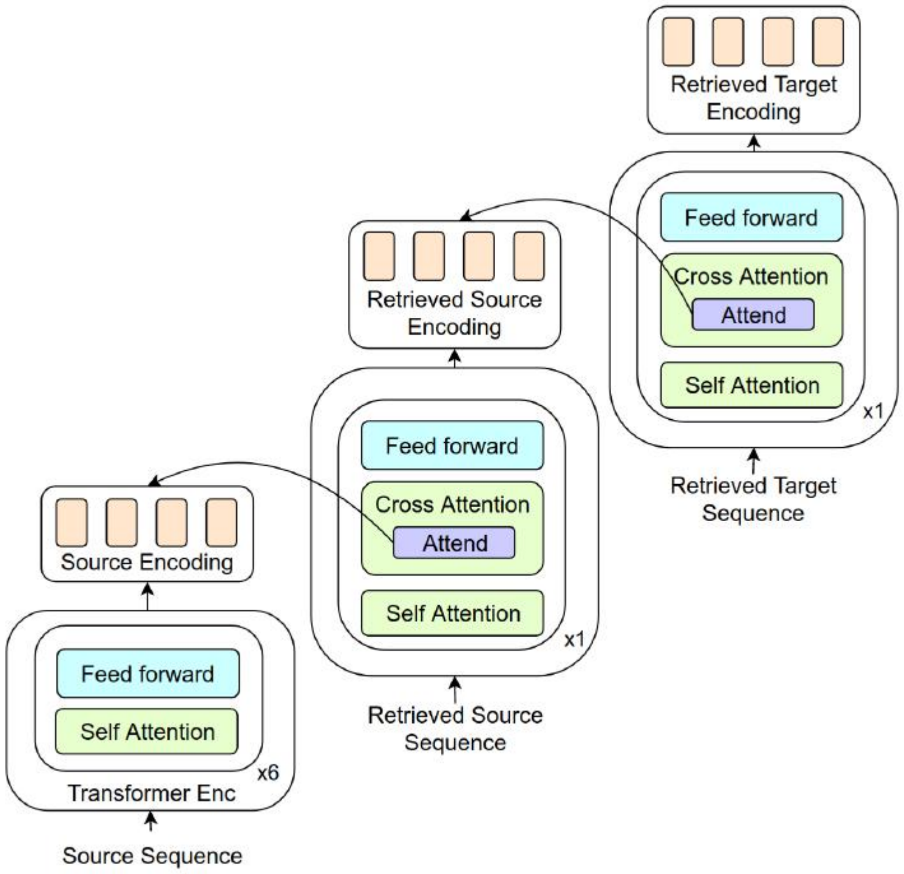
中华人民共和国缔结或者参加的与合同有关的国际条约同中华人民共和国法律有不同规定的,适用该国际条约的规定。但是,中华人民共和国声明保留的条款除外。

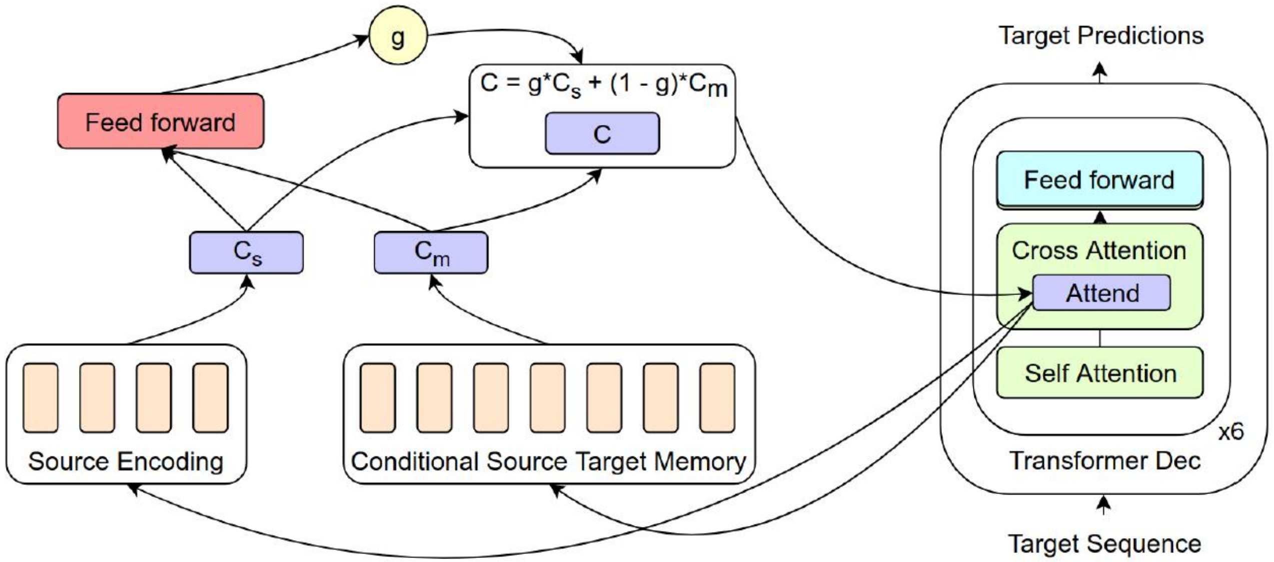
China's Li Duihong won the women's 25-meter sport pistol Olympic gold with a total of 687.9 points early this morning Beijing time. (Guangming Daily, 1996/07/02)

我国女子运动员李对红今天在女子运动手枪决赛中，以687.9环战胜所有对手，并创造新的奥运记录。（《光明日报》1996年7月2日）

DA-NMT paper1

Non-Parametric Adaptation for Neural Machine Translation (NAACL-2019)





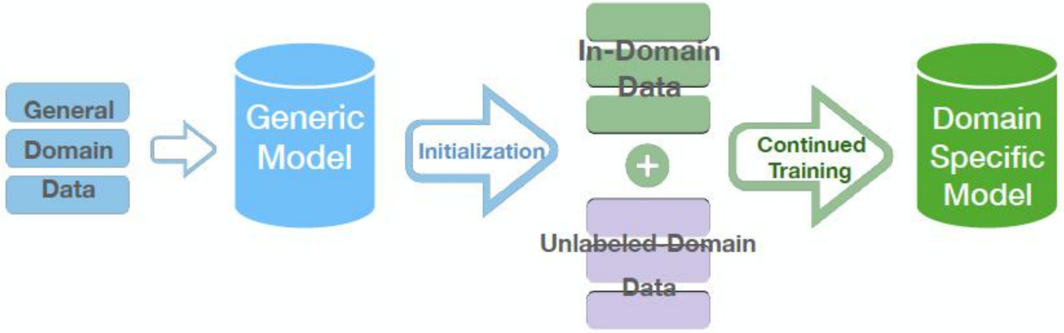
DA-NMT paper1

Non-Parametric Adaptation for Neural Machine Translation (NAACL-2019)

Model	Data	newstest 14	IWSLT 2015	OpenSub	JRC-Acquis
TransformerBase	Multi Domain (MD)	41.92	43.17	26.67	56.19
+ CSTM	MD + IDF Sentence	40.89	42.35	28.25	65.38
+ CSTM	MD + IDF N-Gram	41.92	**45.09**	28.74	66.39
+ CSTM	MD + Dense N-Gram	**42.41**	**45.02**	**29.06**	**66.92**

DA-NMT paper2

Curriculum Learning for Domain Adaptation in Neural Machine Translation (NAACL-2019)



Data Selection => Shard data => Training

	TED(de)	TED(ru)	patent(de)	patent(ru)
GEN	34.59	23.40	35.95	23.41
IN	2.53	1.76	12.09	16.81
IN_CT	36.16	25.04	54.70	35.61
std_rand	35.32	24.33	50.00	34.70
std_ML	36.02	24.73	50.40	30.96
CL_ML	38.78	26.45	52.91	34.18
Δ_ML	2.76	1.72	2.51	3.22
std_CDS	35.83	24.60	52.58	34.54
CL_CDS	**38.88**	**26.49**	**55.51**	**36.59**
Δ_CDS	3.05	1.89	2.93	2.05

- **std_ML**: standard continued training with Moore-Lewis scores
- **CL_ML**: curriculum learning approach to continued training with Moore-Lewis scores
- **std_CDS**: standard continued training with scores from Cynical Data Selection
- **CL_CDS**: curriculum learning approach to continued training with scores from Cynical Data Selection

DA-NMT paper3

Domain Adaptive Inference for Neural Machine Translation (ACL-2019-short)

Problem: Catastrophic forget

$$L(\theta) = L_B(\theta) + \Lambda \sum_j F_j(\theta_j - \theta^*_{A,j})^2$$

- **No-reg**, where $\Lambda = 0$
- **L2**, where $F_j = 1$ for each parameter index j
- **EWC**, where $F_j = \mathbb{E}\left[\nabla^2 L_A(\theta_j)\right]$, a sample estimate of task A Fisher information. This effectively measures the importance of θ_j to task A.

	Training scheme	**Health**	**Bio**
1	Health	**35.9**	33.1
2	Bio	29.6	36.1
3	Health and Bio	35.8	37.2
4	1 then Bio, No-reg	30.3	36.6
5	1 then Bio, L2	35.1	37.3
6	1 then Bio, EWC	35.2	**37.8**

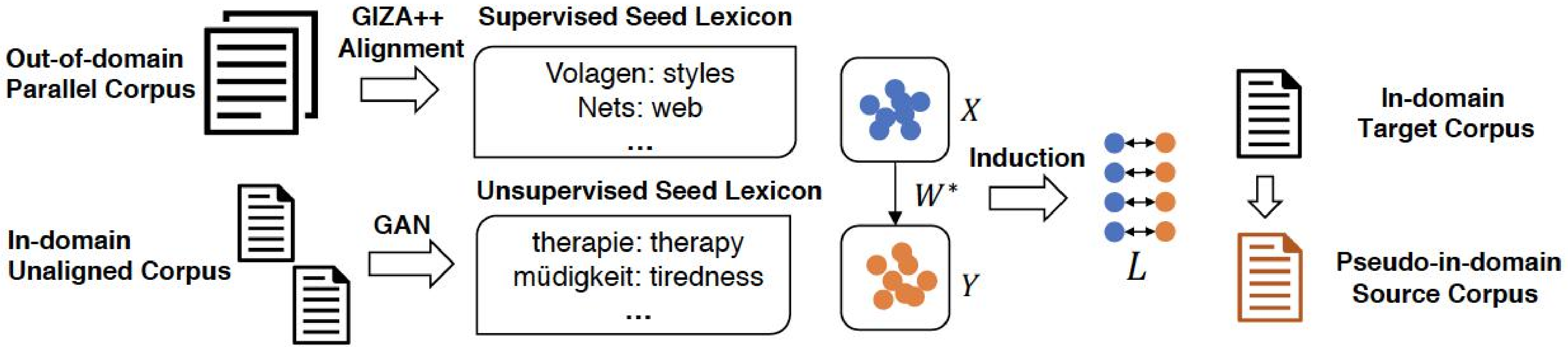
DA-NMT paper4

Domain Adaptation of Neural Machine Translation by Lexicon Induction

(ACL-2019)

Unsupervised domain adaptation: Training: De-En (News), De (Spoken), En (Spoken) (Domain Mismatch

Testing: De-En (Spoken).



Approach: 1. Lexicon Induction (Unsupervised or supervised)

2. NMT Data Generation and Training

DA-NMT paper4

Domain Adaptation of Neural Machine Translation by Lexicon Induction

(ACL-2019)

Corpus	Words	Sentences	W/S
Medical	12,867,326	1,094,667	11.76
IT	2,777,136	333,745	8.32
Subtitles	106,919,386	13,869,396	7.71
Law	15,417,835	707,630	21.80
Koran	9,598,717	478,721	20.05

	Medical	Subtitles	Law	Koran
Unadapted	7.43	5.49	4.10	2.52
Copy	13.28	6.68	5.32	3.22
BT	18.51	11.25	11.55	**8.18**
DALI-U	20.44	9.53	8.63	4.90
DALI-S	19.03	9.80	8.64	4.91
DALI-U+BT	**24.34**	**13.35**	**13.74**	8.11
DALI-GIZA++	28.39	9.37	11.45	8.09
In-domain	46.19	27.29	40.52	19.40

Comparison among different methods on adapting NMT from IT to {Medical, Subtitles, Law, Koran} domains, along with two oracle results.

Summary on DA-NMT

- Problems
  (1) Catastrophic forgetting: embedding?

- Potential Methods
  (1) Can memory-based methods relieve the problem?
  (2) Is non-parametric method necessary?