Exploiting Out-of-Domain Parallel Data through Multilingual Transfer Learning for Low-Resource NMT

Aizhan Imankulova†, Raj Dabre‡,
Atsushi Fujita‡, and Kenji Imamura‡

†Tokyo Metropolitan University
‡NICT, Japan

Theme of this work

- What is translation quality in an extremely low-resource scenario?
- Difficult to obtain a high-resource parallel data
  - Upper limit quantity in budget-poor situations: 10k-20k sentence pairs
- Aim of this research: Attest a feasibility of
  - Multilingualism
    - How to use another helping languages (e.g., English)?
  - Back-translation
    - How to use additional monolingual data?
  - Out-of-domain data
    - What are the effective ways to exploit out-of-domain data?

Our Japanese–Russian setting

- Domain: News
  - Important and challenging
  - Large presence of out-of-vocabulary (OOV)

- Data conditions
  - Extremely low-resource in-domain parallel data
    - Japanese ↔ Russian — 12k only
    - Distant: scripts, grammar, syntax, etc.
  - Larger but still small-scale helping in-domain parallel data
    - Japanese ↔ English — 47k only
    - Russian ↔ English — 82k only
  - Large-scale in-domain monolingual data and helping out-of-domain parallel data

Test bed

Bilingual corpora

Lang	Partition	Origins	#sent.	#tokens
Ja↔Ru	Train	Global Voices	12,356	341k / 229k
	Dev	News Commentary	486	16k / 11k
	Test	News Commentary	600	22k / 15k
Ja↔En	Train	Global Voices	47,082	1.27M / 1.01M
	Dev	News Commentary	589	21k / 16k
	Test	News Commentary	600	22k / 17k
Ru↔En	Train	Global Voices	82,072	1.61M / 1.83M
	Dev	News Commentary	313	7.8k / 8.4k
	Test	News Commentary	600	15k / 17k

Extremely small in-domain data

Larger but still small-scale in-domain helping data

https://github.com/aizhanti/JaRuNC

Monolingual corpora

	Corpus	Ru	Ja	En
News domain	Global Voices	24k	26k	842k
	Other	71M	19M	193M
Other Domains (IWSLT, Tatoeba)		121k	414k	273k

RQs in our study

- **[RQ1]** What kind of translation quality can we obtain in an extremely low-resource scenario?
  - Extensive comparisons of multiple architectures and MT paradigms
  - Multilingualism
  - Back-translation
- **[Contribution 1]**
  - Using only low-resource in-domain data limits the performance

- **[RQ2]** What are the effective ways to exploit out-of-domain data for extremely low-resource in-domain translation?
  - Multistage fine-tuning
  - Iterative fine-tuning on pseudo-parallel data
- **[Contribution 2]**
  - Achieved consistent improvements on all six translation directions

[RQ1] What kind of translation quality can we obtain in an extremely low-resource scenario?

Settings (see details in our paper)

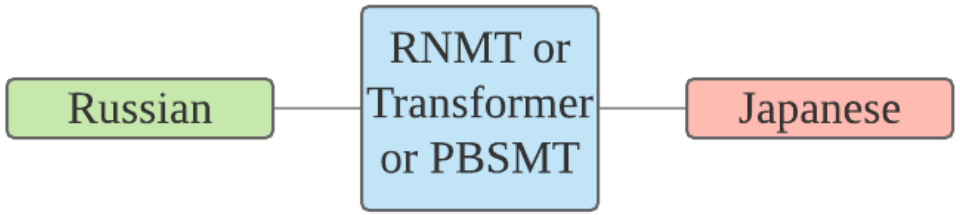
- NMTs (tensor2tensor)
  - Models: Transformer [Vaswani et al., 2017]
    - (Bidirectional RNN + attention [Bahdanau et al., 2015])
  - Sub-wording: Internal sub-word segmentation
    - Uni- and bi-directional: 16k
    - M2M: 32k

- PBSMT (Moses)
  - Unidirectional
  - English as the pivot language

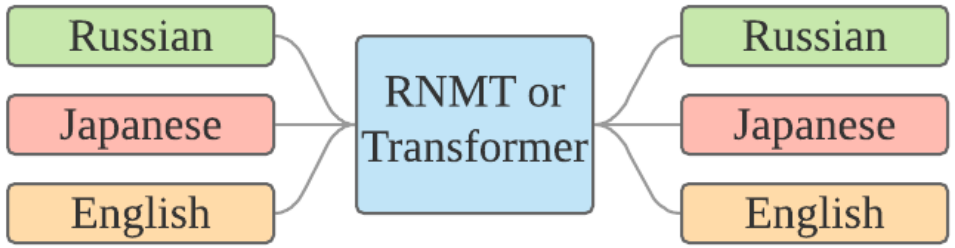
- Evaluation: Moses's script
  - BLEU: case-sensitive, tokenized

Models examined (subset)

- M2M setting with Ja ↔ Ru : Ja ↔ En : Ru ↔ En = 1:1:1



Uni-directional



Multilingual (M2M)

Bi-directional

Pivot-based (Cascade, Synthesize, Triangulate, Induced)

Baseline results (excerpts)

Best

Model	Ja→Ru	Ru→Ja	Ja→En	En→Ja	Ru→En	En→Ru
Uni-directional Transformer	0.70	1.96	4.36	7.97	20.70	16.24
Bi-directional Transformer	0.19	0.87	6.48	10.63	22.25	16.03
M2M Transformer	3.72	**8.35**	**10.24**	**12.43**	22.10	**16.92**
Uni-directional PBSMT	2.02	4.45	8.19	10.27	**22.37**	16.52
Best Pivot-based Transformer	2.41	6.84				
Best Pivot-based PBSMT	**4.02**	7.62	-	-	-	-
Unsupervised PBSMT	0.37	0.65				

- Multilingualism is helpful
  - M2M models >> uni-directional and bi-directional models
- PBSMT models were the best for Ja→Ru and Ru→En
  - Unsupervised approach did not push the limit
- Further exploration: M2M Transformer as the baseline

Augmentation with back-translation

- Monolingual target data selection [Moore and Lewis, 2010]
  1. For each language, select low-perplexity sentences using 4-gram LM
     - In-domain (Global Voices)
     - General-domain (Global Voices, IWSLT, and Tatoeba data)
  2. Discard sentences containing OOVs
  3. For each translation direction, select the *T*-best monolingual sentences
     - *T* is set to match a ratio 1:1 between parallel and monolingual data
- Process
  1. Create pseudo-parallel corpora (PPC) [Sennrich et al., 2016]
     - Six directions: Source*→Target
     - Using back-translator: M2M Transformer
  2. Train a new M2M system from scratch on a mixture of original and back-translated data
     - Combine different directions of PPC
     - The ratio of original and back-translated data: 1:1

Results for systems augmented by back-translation

>> M2M Transformer

ID	Used Pseudo-parallel Corpora						BLEU scores for each test set					
	Ja*→Ru	Ru*→Ja	Ja*→En	En*→Ja	Ru*→En	En*→Ru	Ja→Ru	Ru→Ja	Ja→En	En→Ja	Ru→En	En→Ru
M2M Transformer							3.72	8.35	10.24	12.43	22.10	16.92
1	✓						**4.59**	8.63	10.64	12.94	22.21	17.30
2		✓					3.74	**8.85**	10.13	13.05	22.48	17.20
3	✓	✓					**4.56**	**9.09**	10.57	**13.23**	22.48	**17.89**
4			✓				3.71	8.05	**11.00**	12.66	22.17	16.76
5				✓			3.62	8.10	9.92	**14.06**	21.66	16.68
6			✓	✓			3.61	7.94	**11.51**	**14.38**	22.22	16.80
7					✓		3.80	8.37	10.67	13.00	22.51	**17.73**
8						✓	3.77	8.04	10.52	12.43	**22.85**	17.13
9					✓	✓	3.37	8.03	10.19	12.79	22.77	17.26
10	✓	✓	✓	✓	✓	✓	**4.43**	**9.38**	**12.06**	**14.43**	**23.09**	17.30

- Best: Using all six-way PPC significantly improved M2M transformer
- One-way: One direction benefits when PPC for that specific direction was used
- In most cases, using two-way PPC is better than using one way PPC

Baseline established

- **[RQ1]** What kind of translation quality can we obtain in an extremely low-resource scenario?
- **[Answer 1]** BLEU scores do not exceed 10 BLEU points
  - Best architecture: Transformer NMT
  - Exploiting other helping data benefits extremely low-resource MT
    - M2M, Pivot-based PBSMT results
  - Unsupervised approach did not work
    - Needs further investigation
  - Using additional PPC generated by weak back-translator has limitations

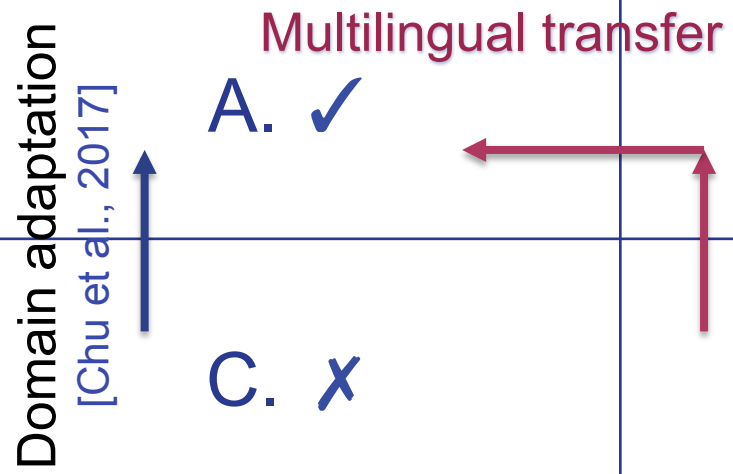
Investigation step	Ja→Ru	Ru→Ja
Uni-directional Transformer	0.70	1.96
M2M Transformer	3.72	8.35
+ six-way pseudo-parallel data (10)	4.43	9.38

Baseline

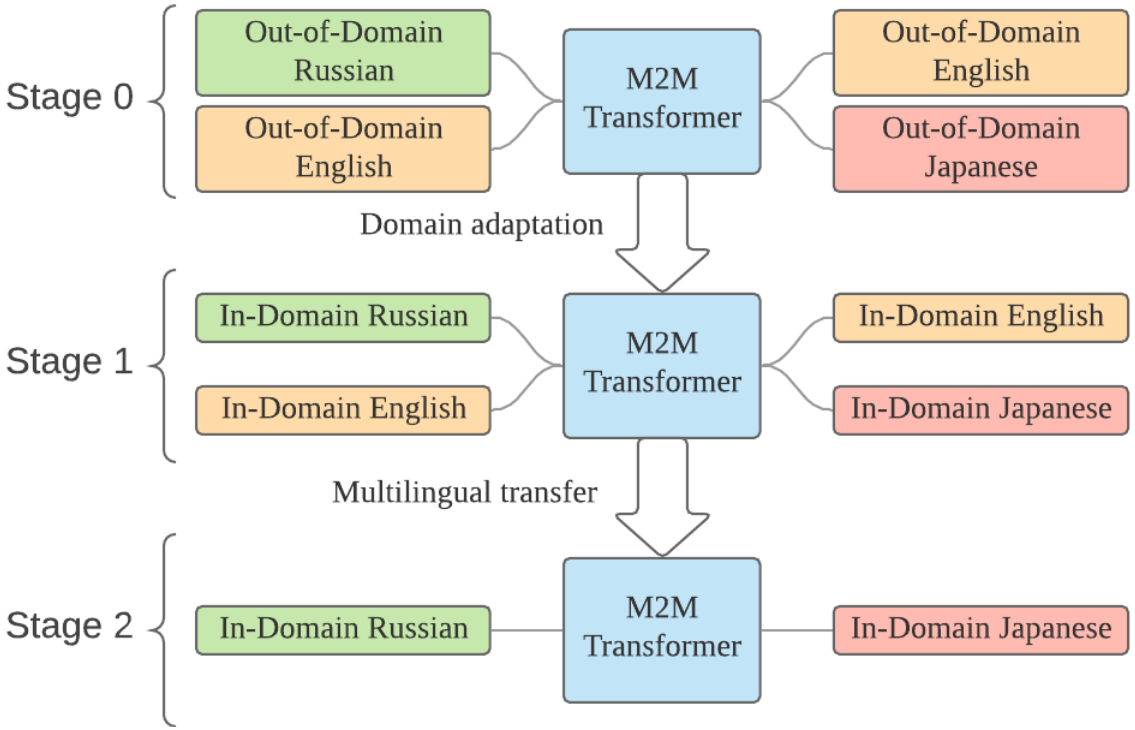
[RQ2] What are the effective ways to exploit out-of-domain data for extremely low-resource in-domain translation?

Available data and our scenario

Domain \ language pair	Direct (Ja↔Ru)	One-side shared (Ja↔En, Ru↔En)
in-domain	A. ✓	B. ✓
out-of-domain	C. ✗	D. ✓



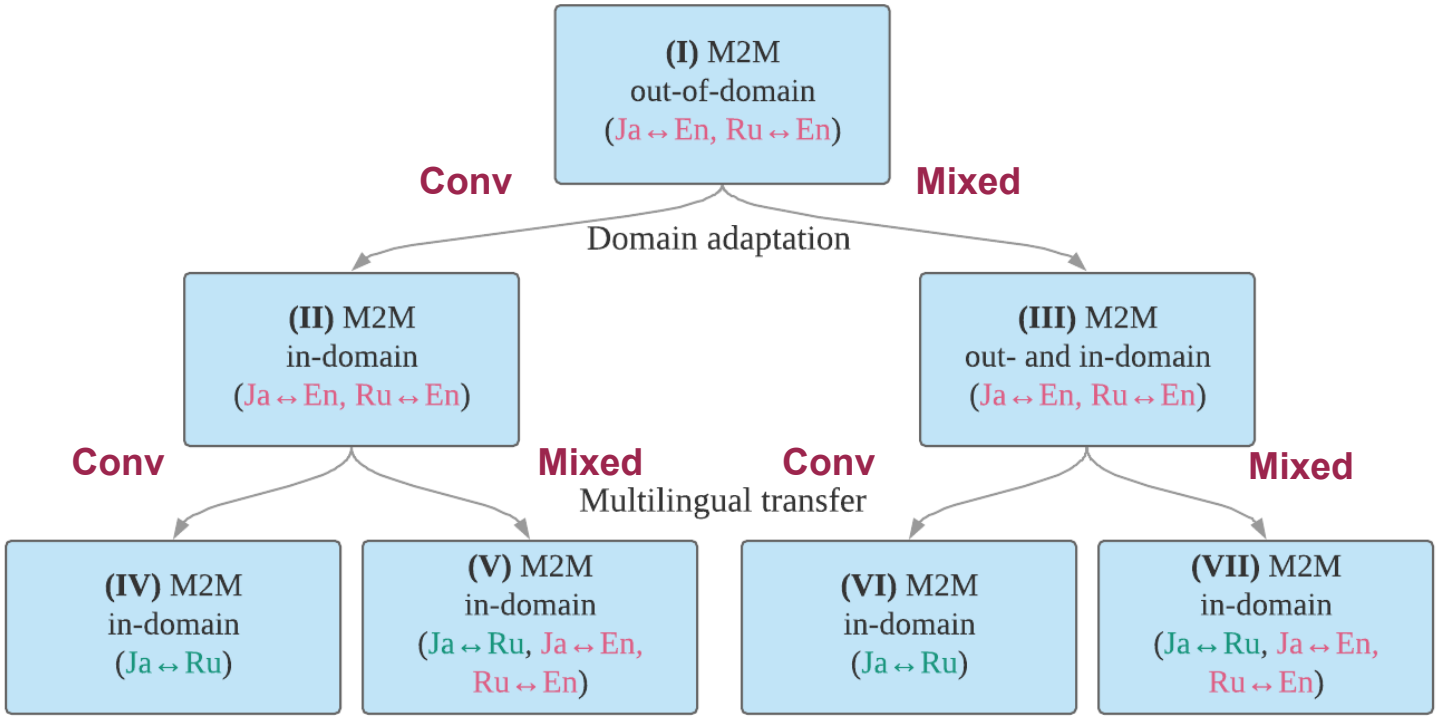
Multistage fine-tuning



- Rationale
  - Need to balance between several different parallel corpora sizes
  - Solve each sub-problem enabling gradual shift of parameters

Variants of fine-tuning methods

- Two options at each stage
  - Conventional fine-tuning: use only in-domain (target pair) data
  - Mixed fine-tuning: use both out-of-domain (other pair) and in-domain (target pair) data
    - Size should be balanced [Johnson et al., 2017]
- Intermediate models and resulting variants



Out-of-domain parallel data

- Ja↔En
  - ASPEC: Asian Scientific Paper Excerpt Corpus
    - Cleanest 1.5M sentence pairs
- Ru↔En
  - UN and Yandex from WMT 2018
    - Sampled sentence pairs that contain at least one OOV token in both sides

Lang	Corpus	#sent.
Ja↔En	ASPEC	1,500,000
Ru↔En	UN	2,647,243
	Yandex	320,325

31x in-domain

36x in-domain

- NB:
  - The vocabularies of all models are determined based only on in-domain data

Results of multistage fine-tuning

Best

	ID	Initialized	Out-of-domain data		In-domain data			BLEU scores for each test set					
			Ja↔En	Ru↔En	Ja↔Ru	Ja↔En	Ru↔En	Ja→Ru	Ru→Ja	Ja→En	En→Ja	Ru→En	En→Ru
	M2M	-	-	-	✓	✓	✓	3.72	8.35	10.24	12.43	22.10	16.92
	I	-	✓	✓	-	-	-	0.00	0.15	4.59	4.15	25.22	20.37
Multi stage	II	I	-	-	-	✓	✓	0.20	0.70	14.10	**17.80**	28.23	24.35
	III	I	✓	✓	-	✓	✓	0.23	1.07	13.31	17.74	**28.73**	**25.22**
	IV	II	-	-	✓	-	-	5.44	10.67	0.12	3.97	0.11	3.66
	V	II	-	-	✓	✓	✓	6.90	11.99	14.34	16.93	27.50	23.17
	VI	III	-	-	✓	-	-	5.91	10.83	0.26	2.18	0.18	1.10
	VII	III	-	-	✓	✓	✓	7.49	12.10	**14.63**	17.71	28.51	24.60
Fewer stages	I'	-	✓	✓	✓	✓	✓	5.31	10.73	14.41	16.34	27.46	23.21
	II'	I	-	-	✓	✓	✓	6.30	11.64	14.29	16.83	27.53	23.00
	III'	I	✓	✓	✓	✓	✓	**7.53**	**12.33**	14.19	16.77	27.94	23.97

- **[RQ2]** What are the effective ways to exploit out-of-domain data for extremely low-resource in-domain translation?
- **[Answer 2]** Multistage fine-tuning led to a robust model (VII)

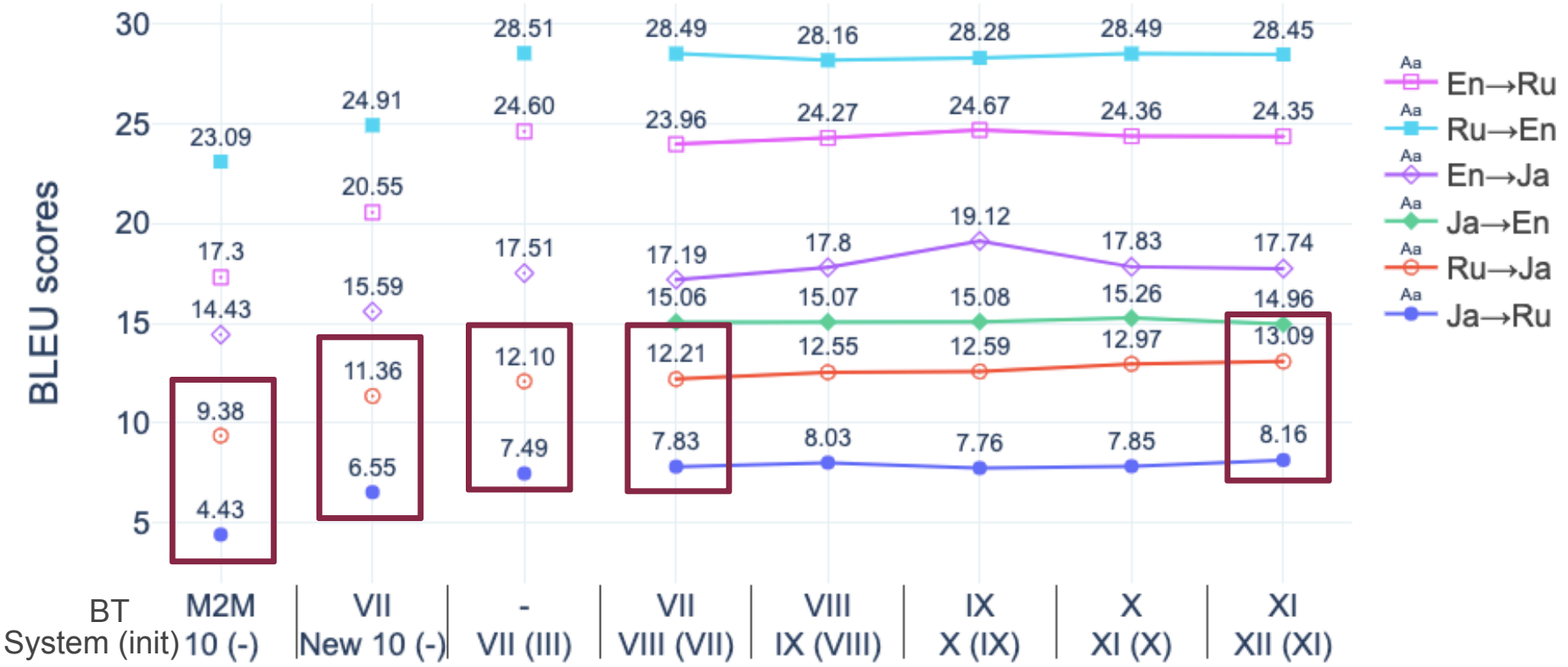
Further augmentation with back-translation

- Given better translation models
  - Does better pseudo-parallel data lead to further improvement?
  - Does one more stage of fine-tuning result in a better model?

- Monolingual data: same as the baseline experiments

- Iteratively create PPC
  1. Create pseudo-parallel corpora (PPC)
     - Six directions: Source*→Target
     - Using back-translator: Current M2M model
  2. Fine-tune the current M2M system
     - Data: a mixture of original and back-translated data with ratio 1:1
  3. Repeat 1-2

Final results



- New 10 >> 10: back-translations of better quality, but underperform VII
- Indirectly explain that 10 and VII can be better than M2M

Conclusion & Future work

- Extremely low-resource settings: Ja↔Ru news domain
  - Test bed based on News Commentary: https://github.com/aizhanti/JaRuNC
  - Contribution 1: Limited success of solutions using only in-domain data
  - Contribution 2: Proposed a multilingual multistage fine-tuning approach
    - Improved Ja↔Ru translation by 3.7 BLEU over a strong baseline

Investigation step	Ja→Ru	Ru→Ja	
Uni-directional Transformer	0.70	1.96	
M2M Transformer	3.72	8.35	
+ six-way pseudo-parallel data (10)	4.43	9.38	Baseline
M2M multistage fine-tuning (VII)	7.49	12.10	
+ six-way pseudo-parallel data (XII)	**8.16**	**13.09**	**Best**

- Future work
  - Use of out-of-domain pseudo-parallel data
  - Better domain-adaptation approaches
  - Evaluation on other challenging language pairs