The Highs and Lows of Simple Lexical Domain Adaptation Approaches for Neural Machine Translation

Nikolay Bogoychev Pinzhen Chen

University of Edinburgh
{n.bogoych,pinzhen.chen}@ed.ac.uk

The premise

Machine translation suffers badly from domain mismatch.

Source	Jetzt bin ich nicht mal würdig, ein Paladin zu sein.
out of domain model	In very rare cases, cladribine may not be a palonosetron.
in domain model	Now I’m not even worthy of being a paladin.

The premise

Machine translation suffers badly from domain mismatch.

Source	Jetzt bin ich nicht mal würdig, ein Paladin zu sein.
out of domain model	In very rare cases, cladribine may not be a palonosetron.
in domain model	Now I’m not even worthy of being a paladin.

Exposure bias kills quality. How can we make it better?

Existing solutions

- MRT training

Existing solutions

- MRT training
- Training towards BLEU

Existing solutions

Existing solutions

- MRT training
- Training towards BLEU
- Minimum Bayesian risk decoding

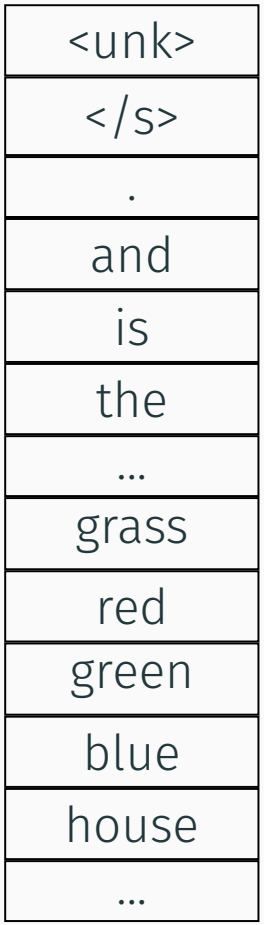
All of those are computationally expensive. Can we do something cheaper

Simple domain adaptation

Lexical shortlisting

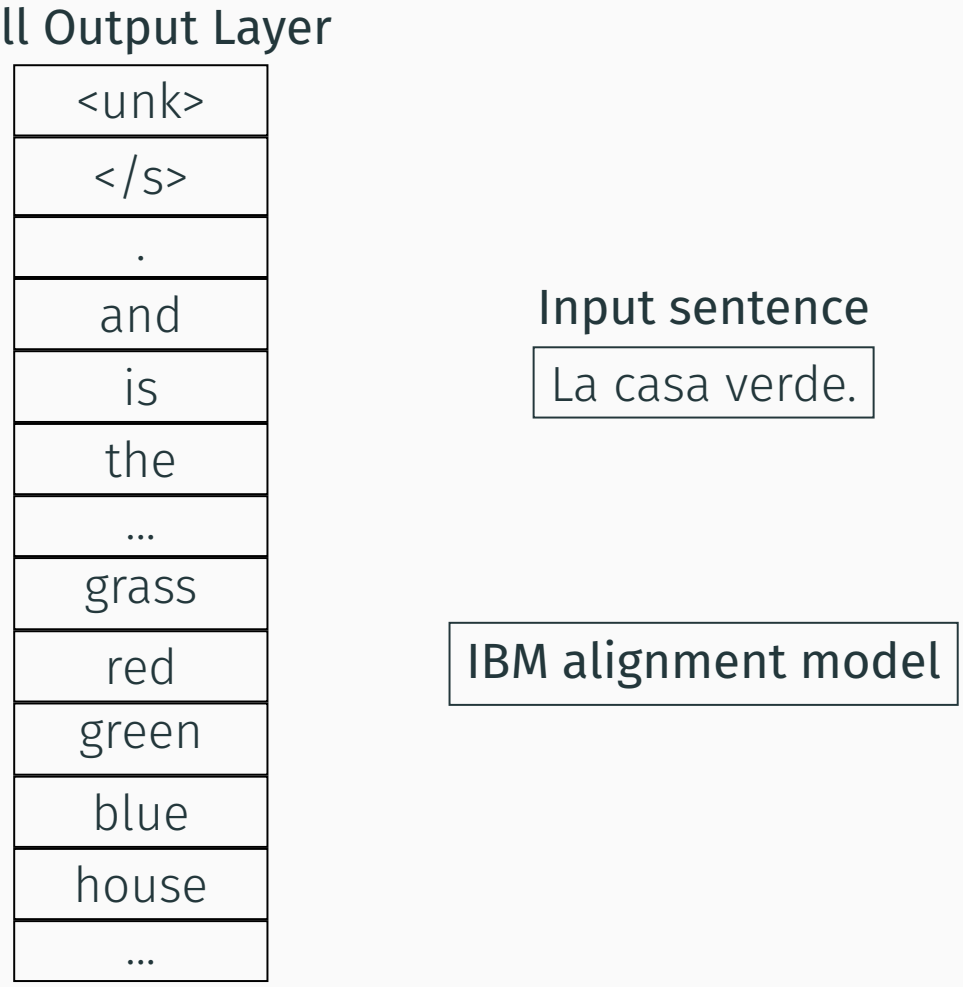
Lexical shortlisting is used to speed up inference.

Full Output Layer



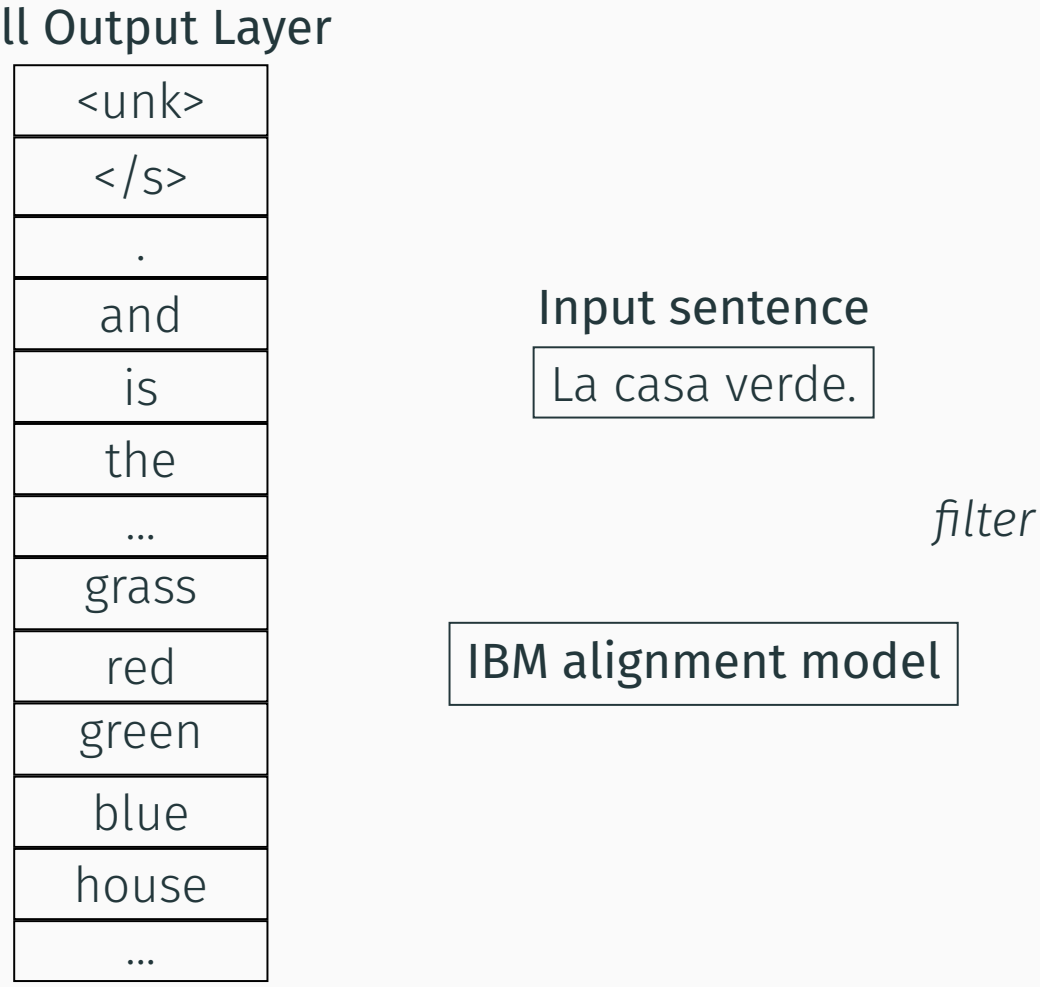
Lexical shortlisting

Lexical shortlisting is used to speed up inference.



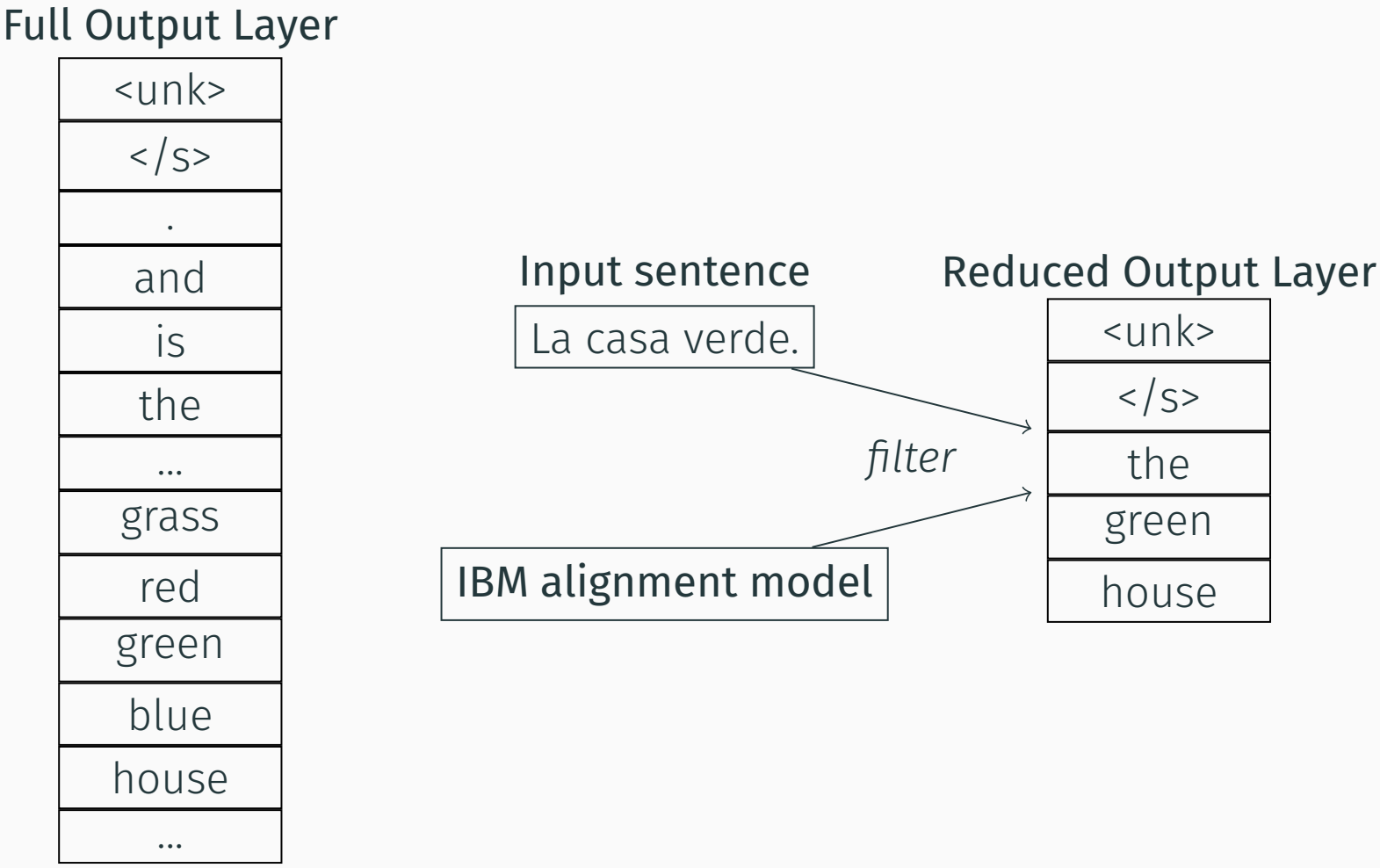
Lexical shortlisting

Lexical shortlisting is used to speed up inference.



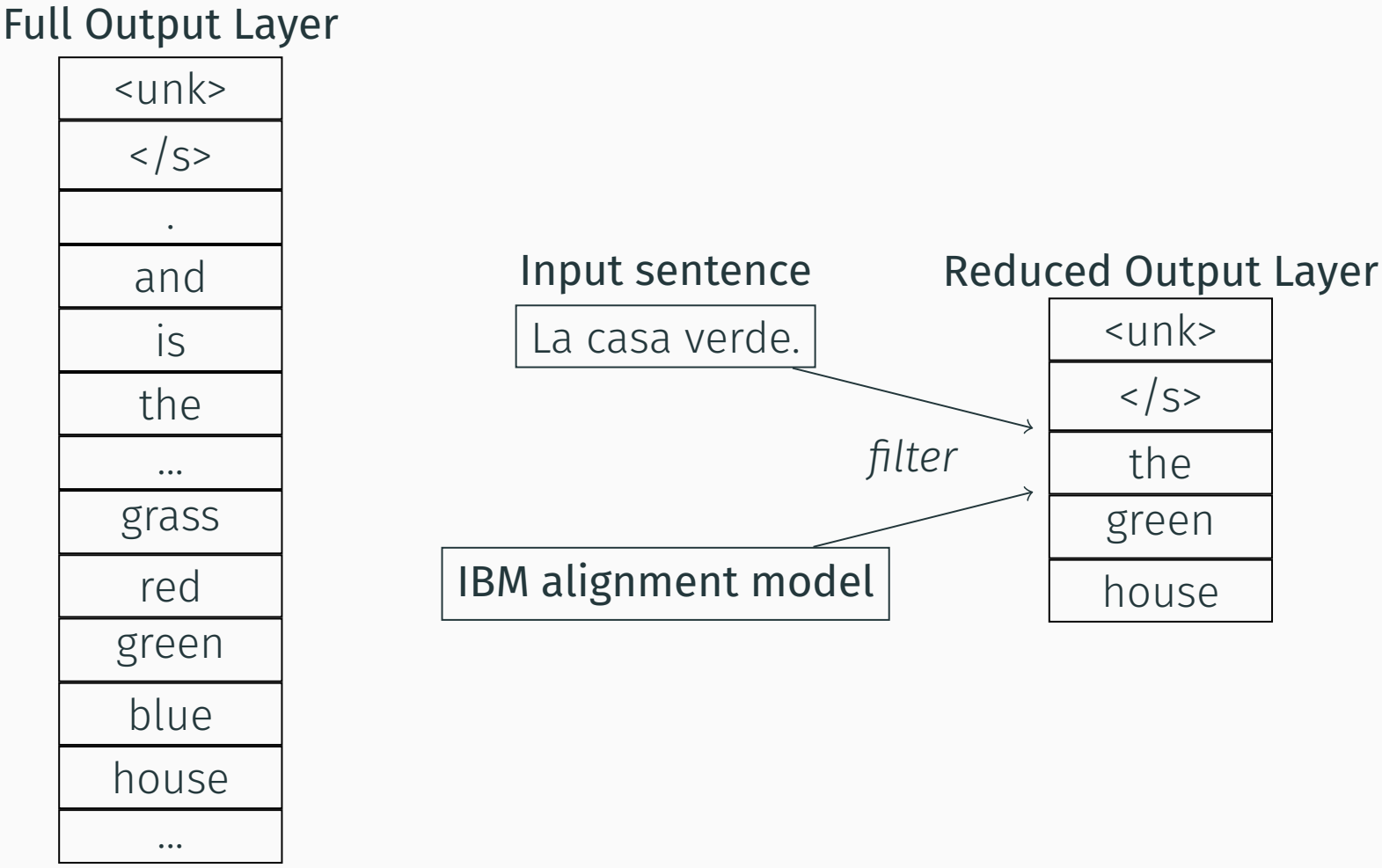
Lexical shortlisting

Lexical shortlisting is used to speed up inference.



Lexical shortlisting

Lexical shortlisting is used to speed up inference.



Can the IBM model help domain adaptation

Reranking

n-best list reranking is a known post-processing steps.

- Inter-hypotheses similarity can reflect a model's confidence.
- Pick the hypothesis that is the most similar to others by re-ranking.
- Selected sentBLEU as the similarity metric after trials.

Reranking

An illustration of re-ranking, ignoring the original hypothesis score.

System output			After re-ranking
Rank	Hypothesis	x-entropy per word	
1	mental :	-1.45	from the age of :
2	from the age of :	-1.63	from the age of 1
3	from the age of 1	-1.74	...
4	from the age of years :	-1.78	...
5	from the tests :	-1.85	...
6	lot	-2.27	...

Experimental Setup

- OPUS German-English dataset.
- In-domain 1M *medical* domain training sentences.
- Out-of-domain test sets *law*, *subtitles*, *it*, *koran*
- Balance Vocabulary. Include out of domain data when training BPE

Results

Domain	BPE trained on *medical* only				BPE trained on all except *subtitles*			
	baseline	shortlist	re-rank	both	baseline	shortlist	re-rank	both
medical	*60.0*	*59.5*	***60.3***	*59.1*	***61.4***	*58.2*	*57.6*	*60.4*

Results

Domain	BPE trained on *medical* only				BPE trained on all except *subtitles*			
	baseline	shortlist	re-rank	both	baseline	shortlist	re-rank	both
medical	*60.0*	*59.5*	***60.3***	*59.1*	***61.4***	*58.2*	*57.6*	*60.4*
Koran	0.9	1.0	0.7	**1.1**	0.8	0.9	0.9	**1.0**

Results

Domain	BPE trained on *medical* only				BPE trained on all except *subtitles*			
	baseline	shortlist	re-rank	both	baseline	shortlist	re-rank	both
medical	*60.0*	*59.5*	***60.3***	*59.1*	***61.4***	*58.2*	*57.6*	*60.4*
Koran	0.9	1.0	0.7	**1.1**	0.8	0.9	0.9	**1.0**
law	19.6	**20.6**	16.6	17.8	17.8	19.3	19.8	**20.8**

Results

Domain	BPE trained on *medical* only				BPE trained on all except *subtitles*			
	baseline	shortlist	re-rank	both	baseline	shortlist	re-rank	both
medical	*60.0*	*59.5*	***60.3***	*59.1*	***61.4***	*58.2*	*57.6*	*60.4*
Koran	0.9	1.0	0.7	**1.1**	0.8	0.9	0.9	**1.0**
law	19.6	**20.6**	16.6	17.8	17.8	19.3	19.8	**20.8**
IT	15.0	**16.3**	10.1	11.5	15.7	**18.0**	15.3	17.8
subtitles	2.8	**3.1**	1.4	1.9	2.6	**2.8**	2.4	**2.8**

Having a balanced vocabulary is key

Analysis

Domain	System	1- to 4-gram precisions				Brevity penalty	BLEU (△)
law	baseline	53.0	27.5	16.9	11.0	0.778	17.8
	shortlist	**56.1**	**29.4**	**17.9**	**11.4**	0.804	19.3 (+1.5)

Analysis

Domain	System	1- to 4-gram precisions				Brevity penalty	BLEU (△)
law	baseline	53.0	27.5	16.9	11.0	0.778	17.8
	shortlist	**56.1**	**29.4**	**17.9**	**11.4**	0.804	19.3 (+1.5)
	re-rank	51.4	26.4	16.1	10.5	0.906	19.8 (+2.0)
	both	53.1	27.6	16.7	10.7	**0.919**	**20.8 (+3.0)**

- Shortlisting improves unigram accuracy :)
- Reranking preys on BLEU length penalty : (

The negative results

High resource setting

	Microsoft WMT19	
	baseline	shortlist
medical	14.4	14.4
Koran	0.0	0.0
law	8.7	8.7
IT	15.4	15.4
subtitles	1.0	1.0

IBM model doesn't offer meaningful performance boost here.

The high domain mismatch setting

Very low-resource Burmese-English (18k sentence pairs on sports news).

	baseline	shortlist
news (in-domain)	18.00	15.7
Bible	0.2	0.2

Y it no work : (

Y it no work

In a nutshell: Vocabulary mismatch

Domain	law	**medical**	subtitles[†]	IT	Koran
Number of sentences	695k	1M	1M	372k	529k

Y it no work

In a nutshell: Vocabulary mismatch

Domain	law	**medical**	subtitles†	IT	Koran
Number of sentences	695k	1M	1M	372k	529k
Avg. original sentence length	22.1	12.5	8.0	7.5	20.4
Avg. BPE sentence length	30.4	14.3	11.1	12.7	24.1

Y it no work

In a nutshell: Vocabulary mismatch

Domain	law	**medical**	subtitles†	IT	Koran
Number of sentences	695k	1M	1M	372k	529k
Avg. original sentence length	22.1	12.5	8.0	7.5	20.4
Avg. BPE sentence length	30.4	14.3	11.1	12.7	24.1
Vocab size, appearing >20 times	34k	36k	30k	15k	20k
Vocab overlap with *medical*	11.5k	36k	9.0k	5.8k	5.1k

† The *subtitles* corpus was sampled down from 20M to 1M sentence pairs.

Runaway subwords

Sample sentence pairs from *subtitles* with BPE segmentation.

German und Z@@ eth@@ rid ? nur einen Strei@@ f@@ sch@@ uss .

English and , Z@@ eth@@ rid , just gr@@ aze it .

Runaway subwords

Sample sentence pairs from *subtitles* with BPE segmentation.

German und Z@@ eth@@ rid ? nur einen Strei@@ f@@ sch@@ uss .

English and , Z@@ eth@@ rid , just gr@@ aze it .

How can the IBM model learn any meaningful alignment?

Conclusion

- Fast and cheap domain adaptation strategies.

Conclusion

- Fast and cheap domain adaptation strategies.
- Shorlisting does offer genuine improvement :)

Conclusion

- Fast and cheap domain adaptation strategies.
- Shorlisting does offer genuine improvement :)
- Reranking preys on BLEU length penalty : (

Conclusion

- Fast and cheap domain adaptation strategies.
- Shorlisting does offer genuine improvement :)
- Reranking preys on BLEU length penalty : (
- Doesn’t work on high resource setting : (
- Doesn’t work with large domain mismatch : (

Conclusion

- Fast and cheap domain adaptation strategies.
- Shorlisting does offer genuine improvement :)
- Reranking preys on BLEU length penalty : (
- Doesn’t work on high resource setting : (
- Doesn’t work with large domain mismatch : (
- The main issue is vocabulary mismatch and subword segmentation : (

Thank you for your time!