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

Nikolay Bogoychev **Pinzhen Chen**
{n.bogoych, pinzhen.chen}@ed.ac.uk

Introduction

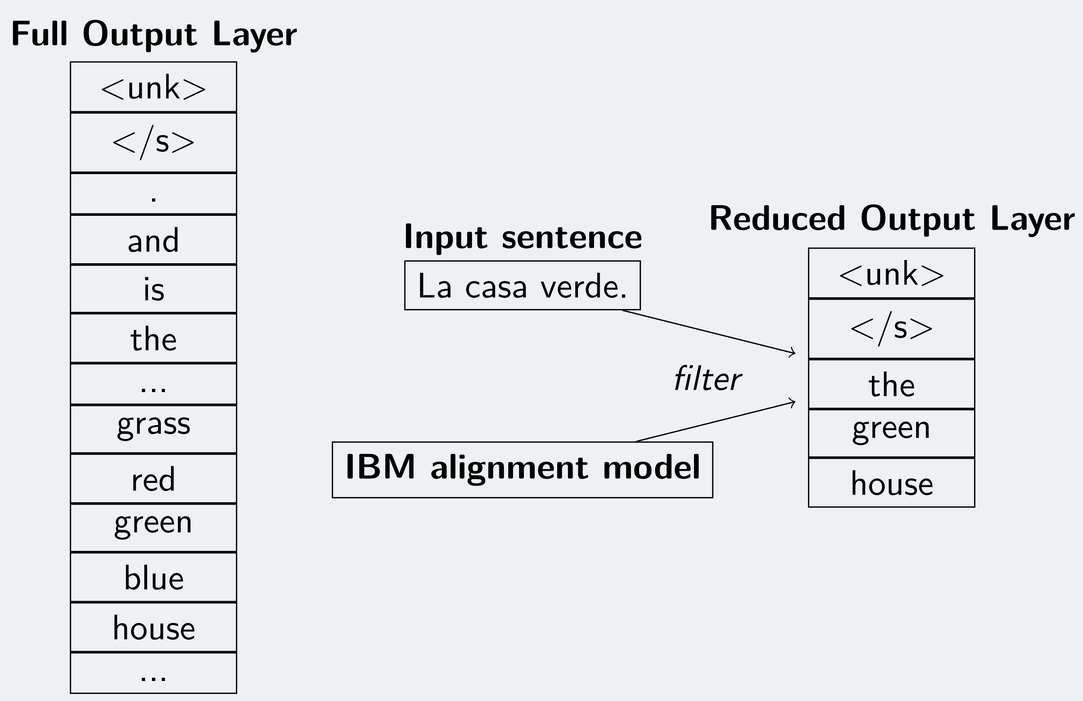
Neural Machine translation deals poorly with domain mismatch.

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

Can we perform *cheap* domain adaptation during decoding?

Lexical shortlisting

- Use IBM model to compute alignment on a small amount of parallel in-domain data.
- Limit the output of the neural network to plausible words according to the IBM model.



n-best list reranking

- 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.

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	...

Table: An illustration of re-ranking hypothesis based on inter-hypothesis similarity

Experimental setup

Domain specific German-English dataset from Opus.

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.
‡ The in-domain model was trained on the *medical* corpus.

The Highs

Some improvements in BLEU.

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**

Results are better if the vocabulary is not overfitted on the in-domain data.

Analysis

Shortlisting improves unigram BLEU score: We get more words correct!

Domain	System	1- to 4-gram precisions				Brevity penalty	BLEU (△)	METEOR (△)
law	baseline	53.0	27.5	16.9	11.0	0.778	17.8	0.36
	shortlist	**56.1**	**29.4**	**17.9**	**11.4**	0.804	19.3 (+1.5)	**0.39 (+0.03)**
	re-rank	51.4	26.4	16.1	10.5	0.906	19.8 (+2.0)	0.31 (−0.05)
	both	53.1	27.6	16.7	10.7	**0.919**	**20.8 (+3.0)**	0.35 (−0.01)
IT	baseline	34.6	18.8	13.1	9.5	0.930	15.7	0.16
	shortlist	**43.9**	**24.7**	**17.1**	**12.1**	0.828	**18.0 (+2.3)**	**0.18 (+0.02)**
	re-rank	33.5	17.2	11.7	8.1	**1.000**	15.3 (−0.4)	0.12 (−0.04)
	both	38.0	20.1	13.6	9.7	**1.000**	17.8 (+2.1)	0.09 (−0.07)

Reranking preys on BLEU length penalty : (

The Lows

Our methods don't work in high resource setting.

	Microsoft WMT19		low-resource
	baseline	shortlist	baseline
medical	14.4	14.4	61.4
Koran	0.0	0.0	0.8
law	8.7	8.7	17.8
IT	15.4	15.4	15.7
subtitles	1.0	1.0	2.6

Table: High-resource German-English results.

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

Table: Very low-resource Burmese-English results. High domain mismatch.

Nor when the domain mismatch is very large.

Y it no work

We have too great domain mismatch

German	English
sein Pilot hat nicht die volle Kontrolle .	its p@@ il@@ ot is@@ n't in control .
und Z@@ eth@@ rid ? nur einen Strei@@ f@@ sch@@ uss .	and , Z@@ eth@@ rid , just gr@@ aze it .

Table: Two random German-English sentence pairs from the *subtitles* dataset after BPE.

Our IBM model can't learn meaningful information from those.

Edinburgh University of Edinburgh Natural Language Processing NLP
MATERIAL
User Focused MARIAN EU Project
THE UNIVERSITY of EDINBURGH informatics