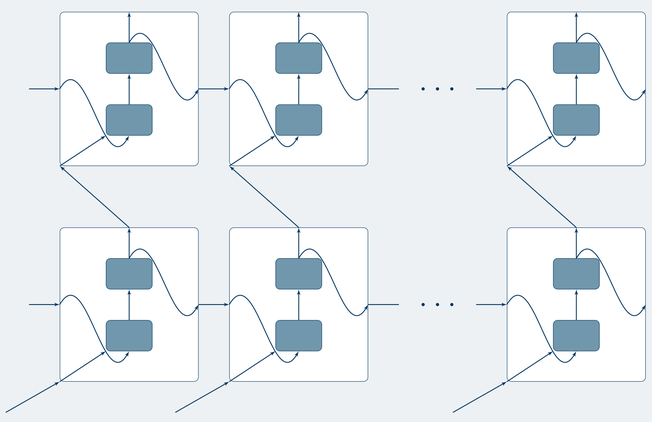
Samsung & Edinburgh's Submission to IWSLT 17

Pawel Przybysz[1], Marcin Chochowski[1], Rico Sennrich[2], Barry Haddow[2] and Alexandra Birch[2]
[1]Samsung R&D Institute Poland
[2]School of Informatics, University of Edinburgh
{m.chochowski,p.przybysz}@samsung.com,
bhaddow@inf.ed.ac.uk, {rico.sennrich,a.birch}@ed.ac.uk

Overview

Deep Neural MT systems using attentional encoder-decoder

- MT Tasks for TED talks
- SLT Tasks Lectures and TED talks
- English $\longleftrightarrow$ German
- Filtering via language identification, sentence alignment scores, and cross-entropy
- Back-translation to use in-domain monolingual data
- Bi-deep model



- Fine-tuning with MAP-L2 regularization

All systems trained using open-source Nematus toolkit

Baseline Training Data

- Amount of available data a lot larger than previous years

Corpus	raw	aligned	filtered
Commoncrawl	2.40M	2.22M	1.62M
Europarl v7	1.92M	1.90M	1.85M
GoldAligment	509	508	486
MultiUN	0.16M	0.16M	0.15M
News Com. v12	0.27M	0.26M	0.26M
Opensubtitles2016	13.88M	12.08M	9.04M
QED Corpus	0.07M	0.07M	0.06M
Rapid 2016	1.33M	1.28M	1.12M
Wikipedia Corpus	2.46M	2.16M	1.18M
WIT3 (in-domain)	0.22M	0.21M	0.20M
Total	22.72M	20.35M	15.47M

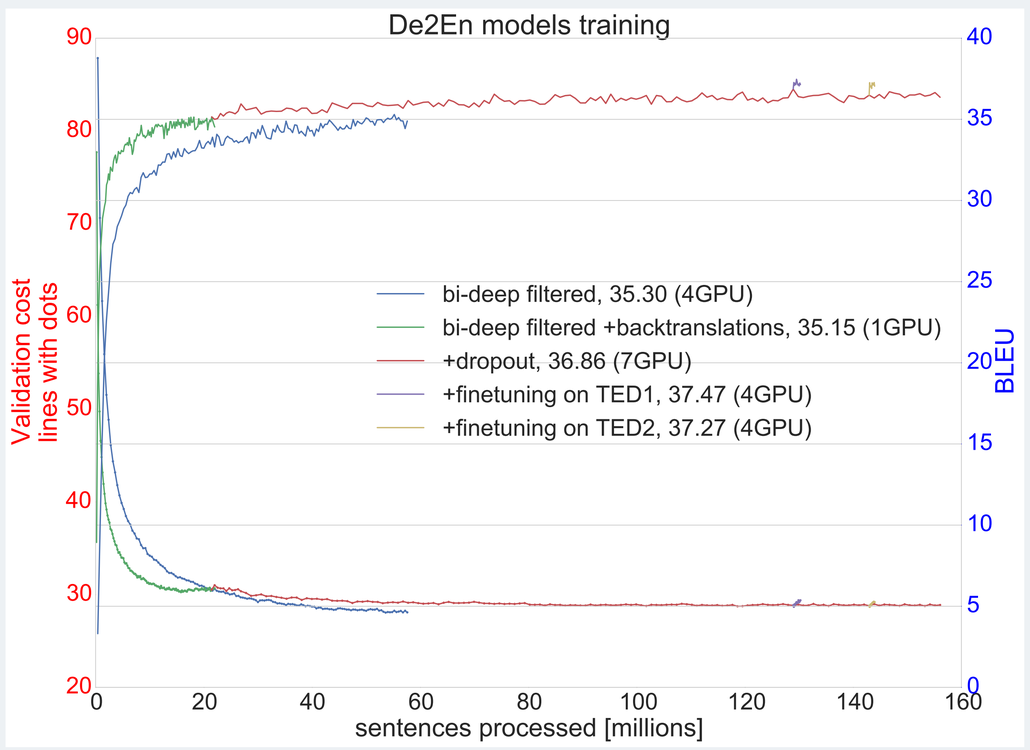
Admissible parallel corpora used for training, with number of segments before and after filtering

Data Selection

Goal: Improve training data quality, domain adaptation

- Sentence Alignment score filtering: 5-50% corpus reduction
- Remove lines with wiki markup
- Use Moore-Lewis filtering of monolingual data to select data for back-translation

Training Progress



Evaluation Results

Translation	Progress set (2016)		Test set (2017)	
	de-en	en-de	de-en	en-de
IWSLT16	32.56	-	27.34	-
baseline	32.52	26.05	27.84	24.33
BiDeep raw	33.92	27.27	29.28	25.14
BiDeep filtered	34.07	27.66	29.94	25.61
+backtranslations	36.27	28.81	30.93	25.24
+dropout	36.50	29.83	31.41	26.66
+finetune on TED	37.08	30.21	32.26	27.38
+checkpoint ens.	37.61	30.34	32.37	27.56
independent ens.	37.56	29.91	32.71	27.23
+right to left	**37.85**	**30.93**	**33.08**	**28.00**

Results for the IWSLT TED translation task

Spoken Language Translation

Our pipeline:

- Punctuation model (NMT system) predicts commas, full stops, question marks, exclamation marks and three dots.
- Input is segmented into sentences based on punctuation.
- Sentences are translated with same NMT system as used for TED.

Links

Code `https://github.com/EdinburghNLP/nematus`