The JHU Machine Translation Systems for WMT 2016

Shuoyang Ding, Kevin Duh, Huda Khayrallah, Philipp Koehn, and Matt Post

Phrase-Based System

- Baseline system from WMT 2015
- Och clusters for language model, OSM, reordering model, sparse features
- Huge language models on CommonCrawl data

Language	Tokens	LM Size
Czech	6.7 billion	13GB
German	65.2 billion	107GB
English	65.1 billion	89GB
Finnish	2.9 billion	8GB
Romanian	8.1 billion	13GB
Russian	23.3 billion	41GB
Turkish	11.9 billion	23GB

- Neural Network Joint Model

Domain Adapted Neural Language Model

- LM training data from different domains

1. Train model on all data, additional training with tuning set
2. Method 1, but only back-propagate through last layer of the network
3. Take the interpolation weights $w_1, w_2, ..., w_n$ of the traditional language model interpolation. Monolingual data with word count $c_1, c_2, ..., c_n$. Compute the normalized interpolation weights as follows:

$$\tilde{w}_i = \frac{w_i}{c_i}$$

Weight (by repeating) domain data with these weights

- Results on English–Romanian

System	newsdev2016b
baseline	23.1
w/o untuned nplm on all data	23.5 (+.4)
w/o untuned nplm on setimes2	23.2 (+.1)
w/o all data nplm + method 1	23.4 (+.3)
w/o all data nplm + method 2	23.8 (+.7)
w/o all data nplm + method 3	24.0 (+.9)

Neural Sequence Model Reranking

- Neural attention model (TensorFlow implementation)
- Reranked 50-best list
- Gains: Russian–English +.04, German–English +.10

Syntax-Based System

- Syntax-based system with Moses (German parser ParZu, English parser Berkeley)
- Hierarchical system: Joshua with Brown clusters

Language Pair	Phrase	Syntax	Joshua
English-Turkish	9.2	-	**9.8**
Turkish-English	12.9	**13.9**	-
English-Finnish	**13.8**	-	11.9
Finnish-English	19.1	-	-
English-Romanian	23.5	-	-
Romanian-English	32.2	-	-
English-Russian	24.0	-	-
Russian-English	27.9	-	-
English-Czech	23.6	-	-
Czech-English	30.4	-	-
English-German	**28.3**	27.3	-
German-English	**34.5**	32.3	-

Morphological Segmentation

- Segment rare words
- Unsupervised segmentation: Morfessor & Byte Pair encoding
- Supervised segmentation: Chipmunk
- Turkish–English morphology results on newsdev2016b

Method	Processing	Thresh.	BLEU
baseline	-	-	13.9
Byte-Pair	preprocessing	-	13.7
Chipmunk	replace-rare	2	14.3
Chipmunk	replace-rare	10	14.9
Chipmunk	replace-rare	20	**15.4**
Chipmunk	replace-rare	20	14.7
Morfessor	replace-rare	0	13.5
Morfessor	replace-rare	2	13.7
Morfessor	replace-rare	5	14.0
Morfessor	replace-rare	10	**14.1**
Morfessor	replace-rare	20	14.2

Phrase-Based with Neural Joint Model

Language Pair	Best 2015	Baseline	w/clusters	w/CC LM	w/both	w/NNJM	w/all&ttl100
English-Turkish	-	7.8	8.2 +0.3	**9.4 +1.6**	8.9 +1.1		
Turkish-English	-	14.0	**14.3 +0.3**	13.9 –0.1	14.1 +0.1		
English-Finnish	15.5	11.9	12.6 +0.7	12.2 +0.3	**12.9 +1.0**		
Finnish-English	19.7	16.5	**16.9 +0.4**	16.4 –0.1	**16.9 +0.4**		
English-Romanian	-	23.4	**24.6 +1.2**	23.4 +0.0	23.5 +0.1	23.7 +0.4	23.5 +0.1
Romanian-English	-	32.0	32.5 +0.5	32.5 +0.5	**32.8 +0.8**	32.0 +0.0	**32.8 +0.8**
English-Russian	24.3	23.9	25.0 +1.1	23.9 +0.0	24.9 +1.0	24.4 +0.5	**25.2 +1.3**
Russian-English	27.9	27.5	28.3 +0.7	28.1 +0.6	28.2 +0.7	27.8 +0.3	**28.7 +1.2**
English-Czech	18.8	18.2	19.2 +1.0	18.8 +0.6	**19.6 +1.4**		
Czech-English	26.2	27.0	27.7 +0.6	27.7 +0.7	**28.1 +1.1**		
English-German	24.9	22.7	**23.0 +0.3**	22.5 –0.2	22.7 +0.0	22.6 –0.1	22.9 +0.2
German-English	29.3	29.0	29.6 +0.6	29.6 +0.6	29.9 +0.9	29.6 +0.6	**30.0 +1.0**