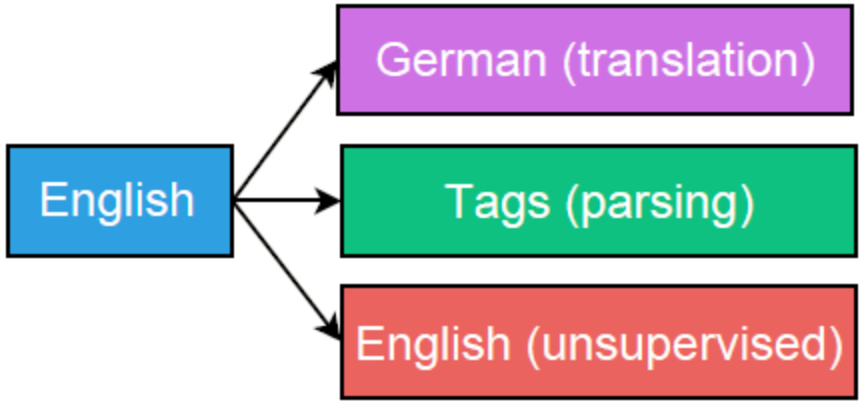
Multi-Task Sequence to Sequence Learning

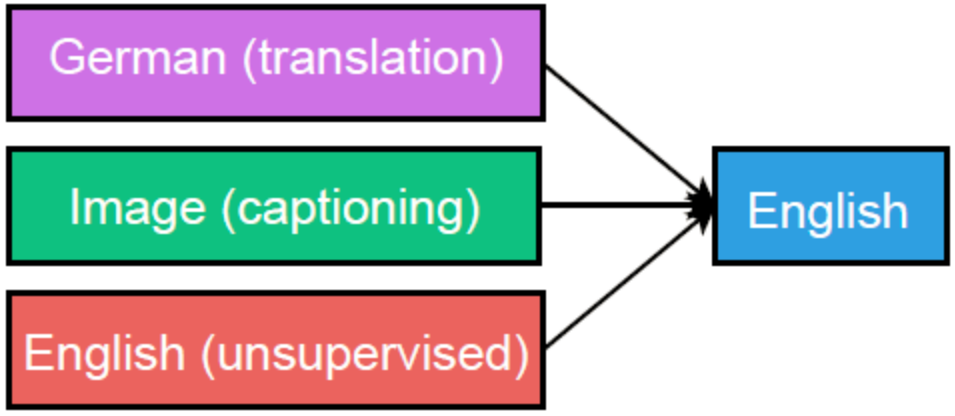
Minh-Thang Luong, Quoc V. Le, Ilya Sutskever,
Oriol Vinyals, Lukasz Kaiser
Google Brain

One-To-Many Setting



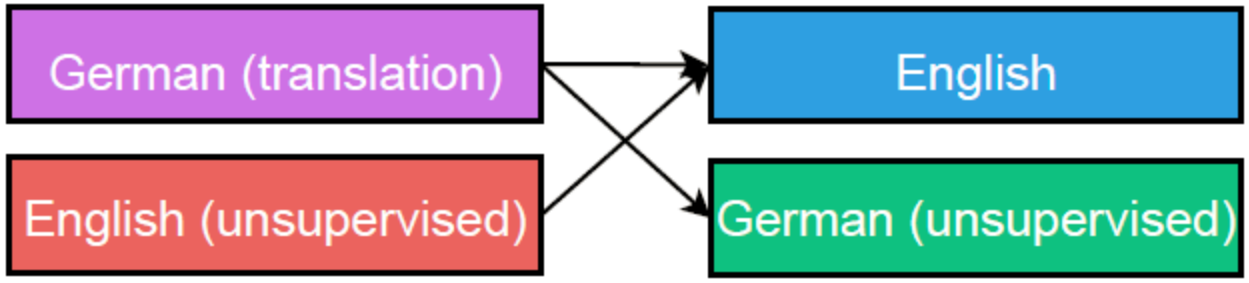
Only encoder shared.

Many-To-One Setting



Only decoder shared.

Many-To-Many Setting



Both encoder and decoder shared.

Auto-Encoder

X -> z -> X

Skip-Thoughts

Xt -> z -> Xt-1

Experiments

Large Tasks with Small Tasks

Task	Translation			Parsing
	Valid ppl	Test ppl	Test BLEU	Test F_1
(Luong et al., 2015a)	-	8.1	14.0	-
Our single-task systems				
Translation	8.8 (0.3)	8.3 (0.2)	14.3 (0.3)	-
PTB Parsing	-	-	-	43.3 (1.7)
Our multi-task systems				
Translation + PTB Parsing (1x)	8.5 (0.0)	8.2 (0.0)	14.7 (0.1)	54.5 (0.4)
Translation + PTB Parsing (0.1x)	8.3 (0.1)	7.9 (0.0)	15.1 (0.0)	**55.2 (0.0)**
Translation + PTB Parsing (0.01x)	**8.2** (0.2)	**7.7** (0.2)	**15.8** (0.4)	39.8 (2.7)

Table 2: **English→German WMT'14 translation & Penn Tree Bank parsing results** – shown are perplexities (ppl), BLEU scores, and parsing F_1 for various systems. For muli-task models, *reference* tasks are in italic with the mixing ratio in parentheses. Our results are averaged over two runs in the format *mean (stddev)*. Best results are highlighted in boldface.

many-to-one

Main task (Translation)

Auxiliary task (Paring-PTB)

Help each other, but the mixing ratio is very important

Large Tasks with Medium Tasks

Task	Translation			Captioning
	Valid ppl	Test ppl	Test BLEU	Valid ppl
(Luong et al., 2015a)	-	14.3	16.9	-
Our single-task systems				
Translation	11.0 (0.0)	12.5 (0.2)	17.8 (0.1)	-
Captioning	-	-	-	30.8 (1.3)
Our multi-task systems				
Translation + Captioning (1x)	11.9	14.0	16.7	43.3
Translation + Captioning (0.1x)	10.5 (0.4)	12.1 (0.4)	18.0 (0.6)	**28.4** (0.3)
Translation + Captioning (0.05x)	**10.3** (0.1)	**11.8** (0.0)	**18.5** (0.0)	30.1 (0.3)
Translation + Captioning (0.01x)	10.6 (0.0)	12.3 (0.1)	18.1 (0.4)	35.2 (1.4)

Table 3: **German→English WMT'15 translation & captioning results** – shown are perplexities (ppl) and BLEU scores for various tasks with similar format as in Table 2. *Reference* tasks are in italic with mixing ratios in parentheses. The average results of 2 runs are in *mean (stddev)* format.

one-to-many

Main task (Translation)

Auxiliary task (image caption)

Help each other, but the mixing ratio is very important

Large Tasks with Large Tasks

Task	Translation		
	Valid ppl	Test ppl	Test BLEU
(Luong et al., 2015a)	-	8.1	14.0
Our systems			
Translation	8.8 (0.3)	8.3 (0.2)	14.3 (0.3)
Translation + HC Parsing (1x)	8.5 (0.0)	8.1 (0.1)	15.0 (0.6)
Translation + HC Parsing (0.1x)	**8.2** (0.3)	**7.7** (0.2)	**15.2** (0.6)
Translation + HC Parsing (0.05x)	8.4 (0.0)	8.0 (0.1)	14.8 (0.2)

Table 4: **English→German WMT'14 translation** – shown are perplexities (ppl) and BLEU scores of various translation models. Our multi-task systems combine translation and parsing on the high-confidence corpus together. Mixing ratios are in parentheses and the average results over 2 runs are in *mean (stddev)* format. Best results are bolded.

many-to-one

Main task (Translation)

Auxiliary task (Parsing-HC)

Help each other, but the mixing ratio is very important

Large-Corpus Parsing Experiment

Task	Parsing	
	Valid ppl	Test F_1
LSTM+A (Vinyals et al., 2015a)	-	92.5
LSTM+A+E (Vinyals et al., 2015a)	-	**92.8**
Our systems		
HC Parsing	1.12/1.12	92.2 (0.1)
HC Parsing + Autoencoder (1x)	1.12/1.12	92.1 (0.1)
HC Parsing + Autoencoder (0.1x)	1.12/1.12	92.1 (0.1)
HC Parsing + Autoencoder (0.01x)	1.12/1.13	92.0 (0.1)
HC Parsing + Translation (1x)	1.12/1.13	91.5 (0.2)
HC Parsing + Translation (0.1x)	1.13/1.13	92.0 (0.2)
HC Parsing + Translation (0.05x)	**1.11/1.12**	**92.4 (0.1)**
HC Parsing + Translation (0.01x)	1.12/1.12	92.2 (0.0)
Ensemble of 6 multi-task systems	-	**93.0**

Table 5: **Large-Corpus parsing results** – shown are perplexities (ppl) and F_1 scores for various parsing models. Mixing ratios are in parentheses and the average results over 2 runs are in *mean (stddev)* format. We show the individual perplexities for all runs due to small differences among them. For Vinyals et al. (2015a)'s parsing results, LSTM+A represents a single LSTM with attention, whereas LSTM+A+E indicates an ensemble of 5 systems. Important results are bolded.

many-to-one

Auto-encoder and Translation help HC Parsing less.

Multi-Tasks and Unsupervised Learning

Task	Translation			German	English
	Valid ppl	Test ppl	Test BLEU	Test ppl	Test ppl
(Luong et al., 2015a)	-	14.3	16.9	-	-
Our single-task systems					
Translation	11.0 (0.0)	12.5 (0.2)	17.8 (0.1)	-	-
Our multi-task systems with Autoencoders					
Translation + autoencoders (1.0x)	12.3	13.9	16.0	**1.01**	2.10
Translation + autoencoders (0.1x)	11.4	12.7	17.7	1.13	**1.44**
Translation + autoencoders (0.05x)	**10.9** (0.1)	**12.0** (0.0)	**18.3** (0.4)	1.40 (0.01)	2.38 (0.39)
Our multi-task systems with Skip-thought Vectors					
Translation + skip-thought (1x)	**10.4** (0.1)	**10.8** (0.1)	17.3 (0.2)	**36.9** (0.1)	**31.5** (0.4)
Translation + skip-thought (0.1x)	10.7 (0.0)	11.4 (0.2)	17.8 (0.4)	52.8 (0.3)	53.7 (0.4)
Translation + skip-thought (0.01x)	11.0 (0.1)	12.2 (0.0)	**17.8** (0.3)	76.3 (0.8)	142.4 (2.7)

Table 6: **German→English WMT'15 translation & unsupervised learning results** – shown are perplexities for translation and unsupervised learning tasks. We experiment with both *autoencoders* and *skip-thought vectors* for the unsupervised objectives. Numbers in *mean (stddev)* format are the average results of 2 runs; others are for 1 run only.

many-to-many

auto-encoder and skip-thought is benefit for translation task, and the mixing ratio is very important.

Conclusion

Multi-related-tasks learnt jointly can help each other, but the mixing learning ratio is very important.