Яндекс

Yandex Translate

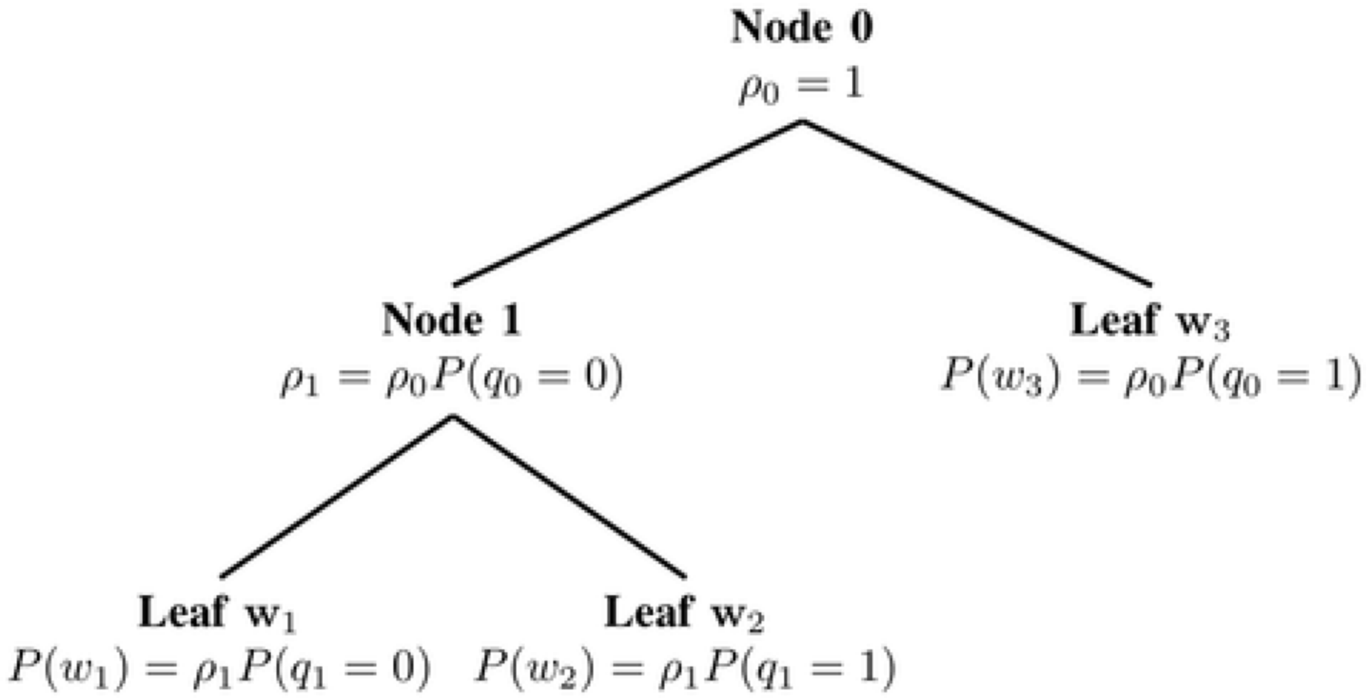
Advanced Neural Machine Translation

David Talbot

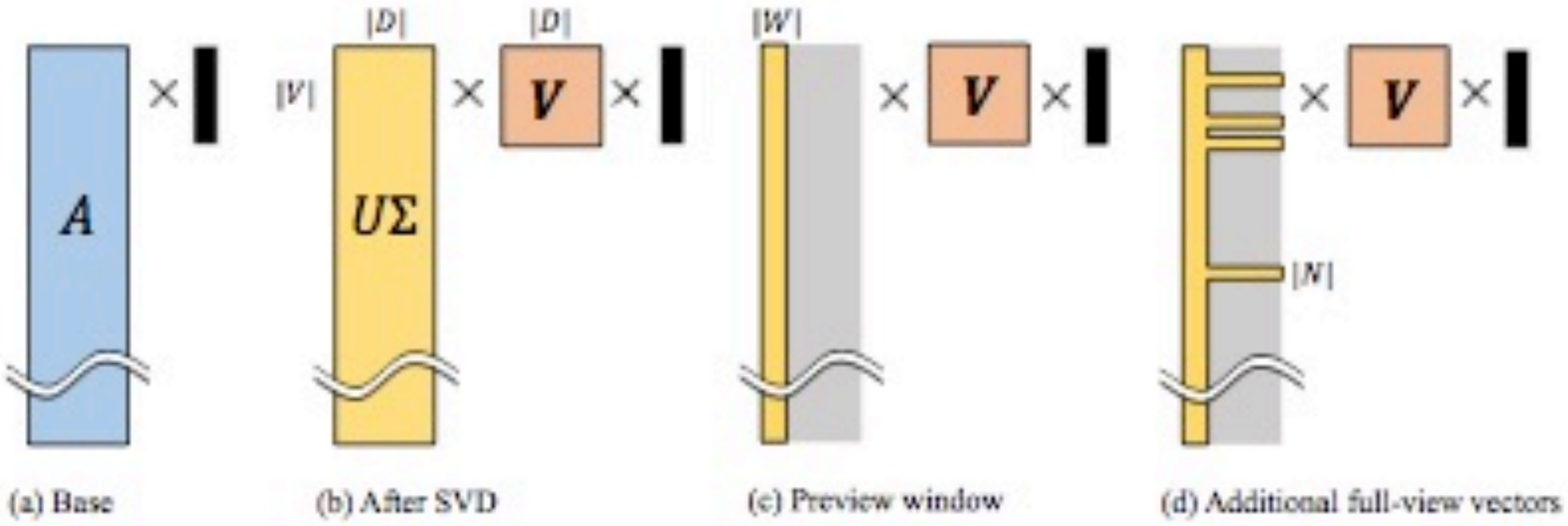
What are the main computational and statistical bottlenecks in NMT?

Computation Bottlenecks in NMT

- SoftMax dominates most architectures
- Naively O(vocab)
- Hierarchical O(log(vocab))



Computation Bottlenecks in NMT



- SVD SoftMax
- Importance sampling
- Vocabulary selection

Vocabulary Manipulation

› Per sentence (or batch) vocabulary for translating sentence f

$$V_f^D = \bigcup_{j=1}^{J} Dict(f_j)$$

$$V_f^P = \bigcup_{\forall f_i,\ldots,f_j \in subseq(f)} Phrases(f_i,\ldots,f_j)$$

$$V_f^{Freq} = FreqTarget(f)$$

$$V_f^R = \bigcup_{\forall\, e_i \in Reference(f)} e_i$$

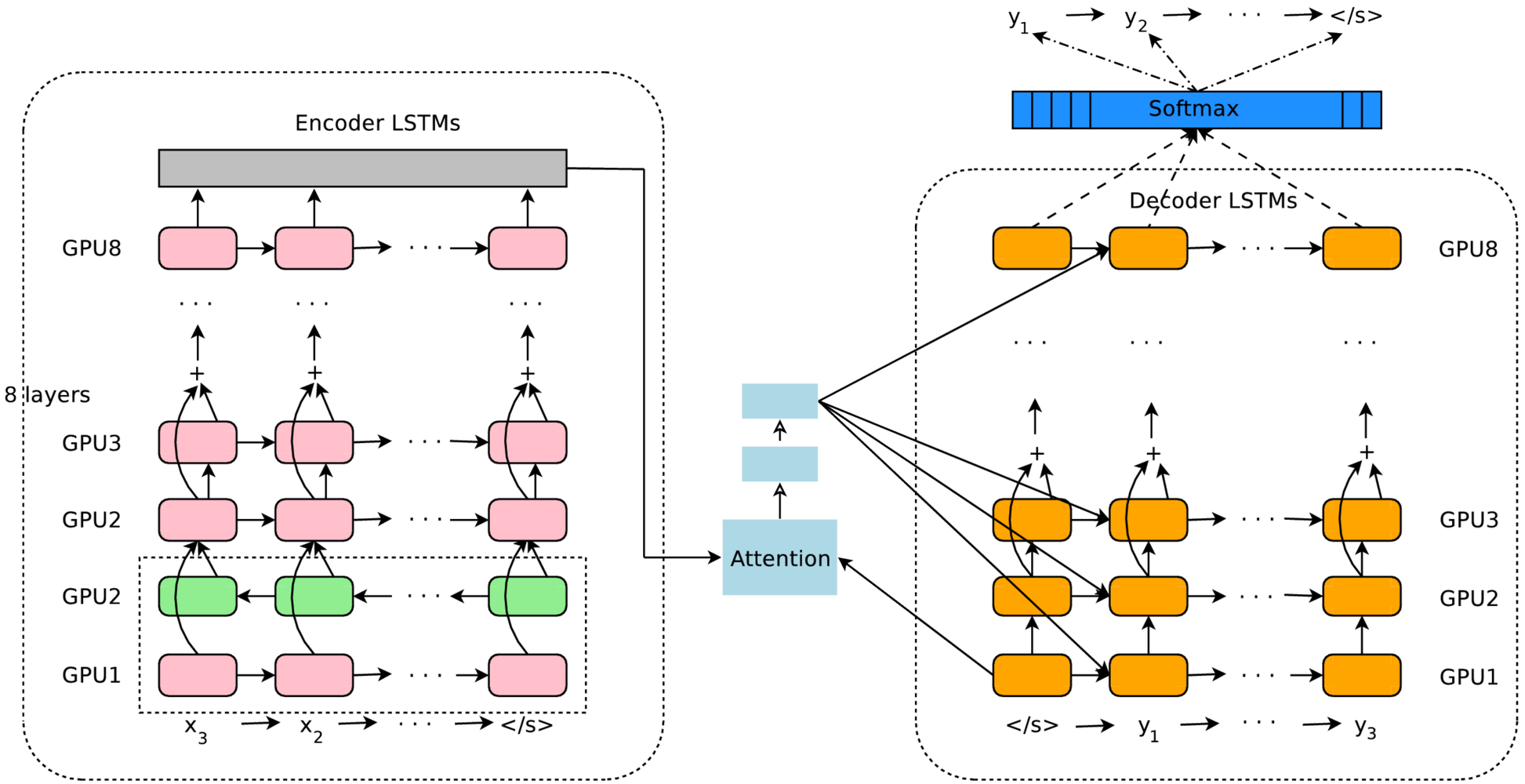
During training only

Statistical Bottlenecks in NMT

- Number of parameters dominated by size of word embeddings
- Unknown word rate depends on vocab size
- UNK replacement (using alignments)
- Byte-pair encoding
- Character-based models smaller but slower

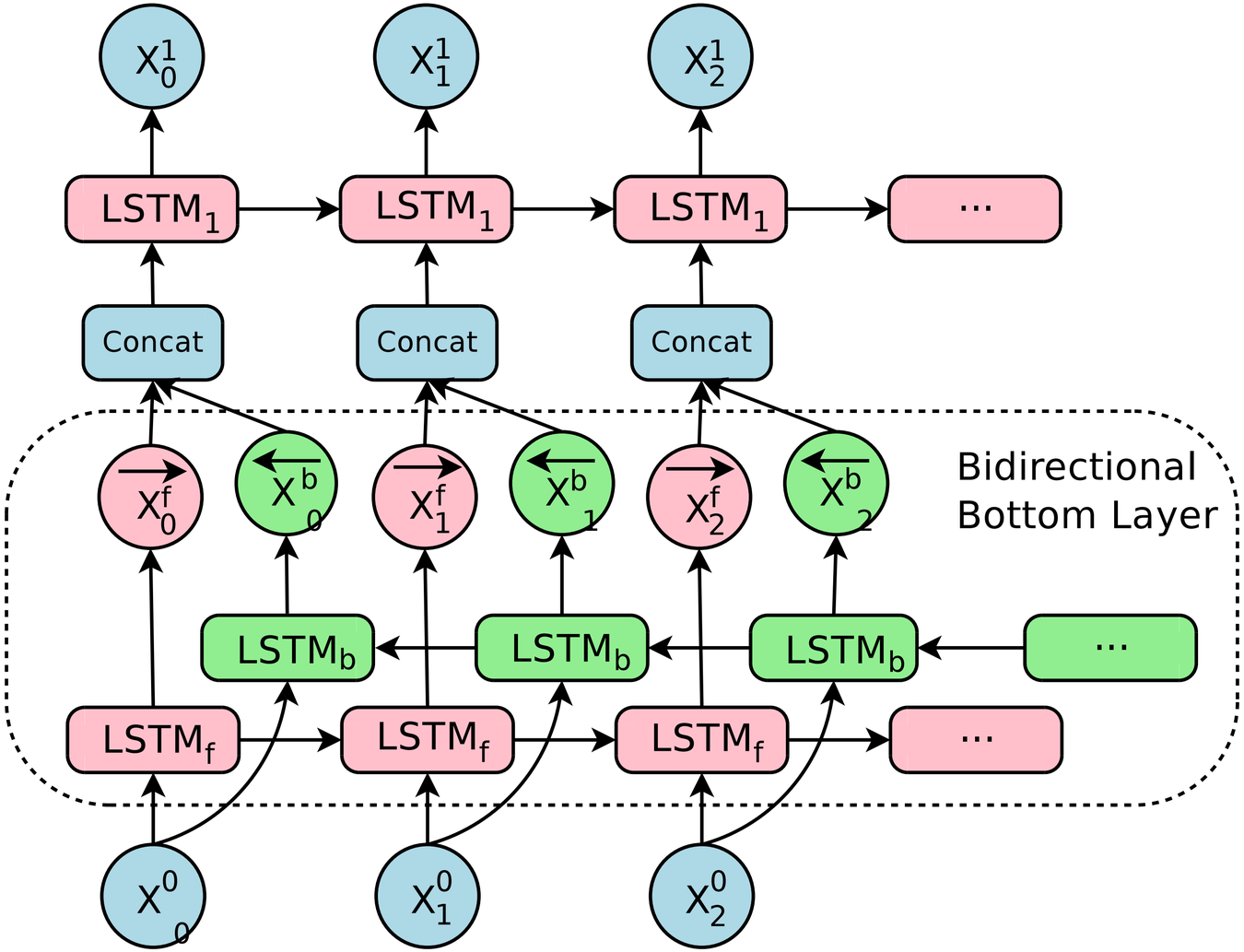
What are the pros/cons of different Encoder-Decoder architectures?

Deep Encoder-Decoder Models (GNMT)

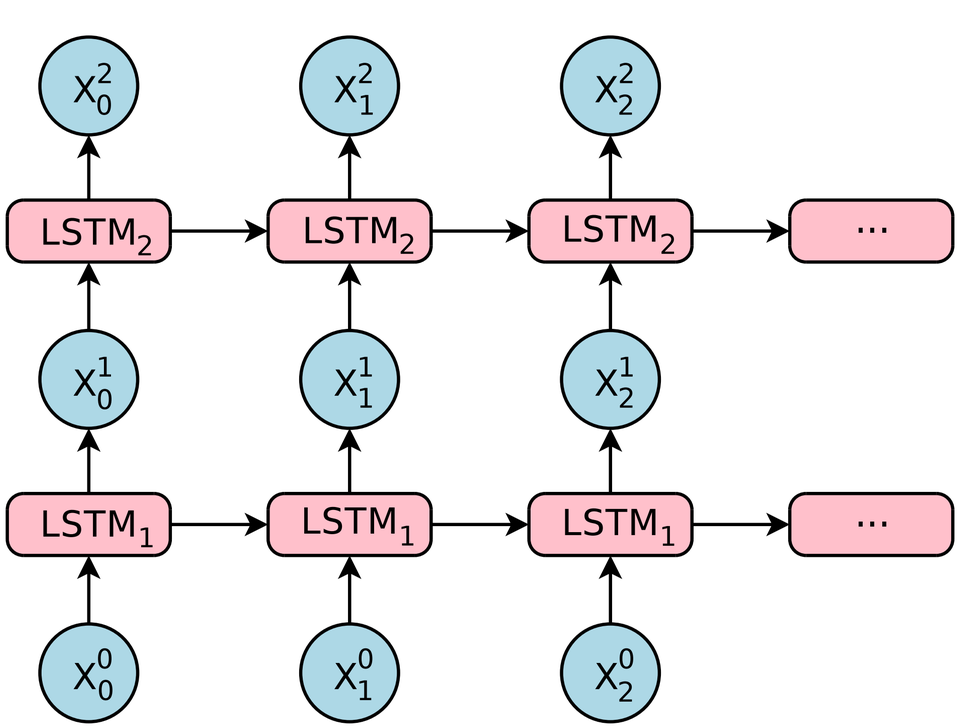


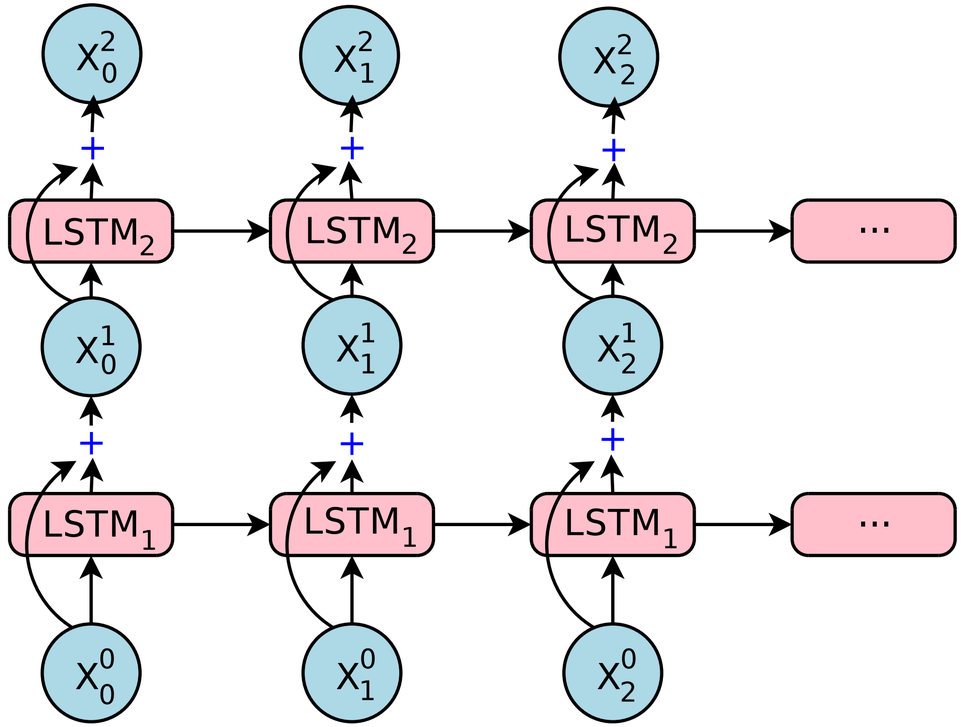
Wu et al. 2016

Deep Encoder-Decoder Models (GNMT)



Deep Encoder-Decoder Models (GNMT)

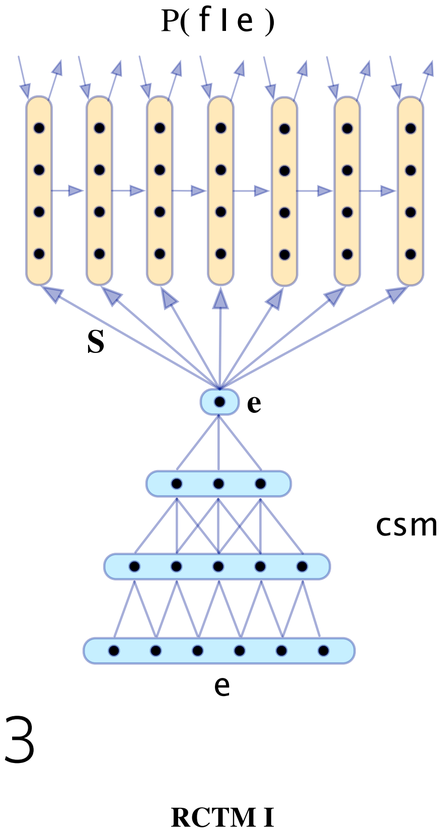




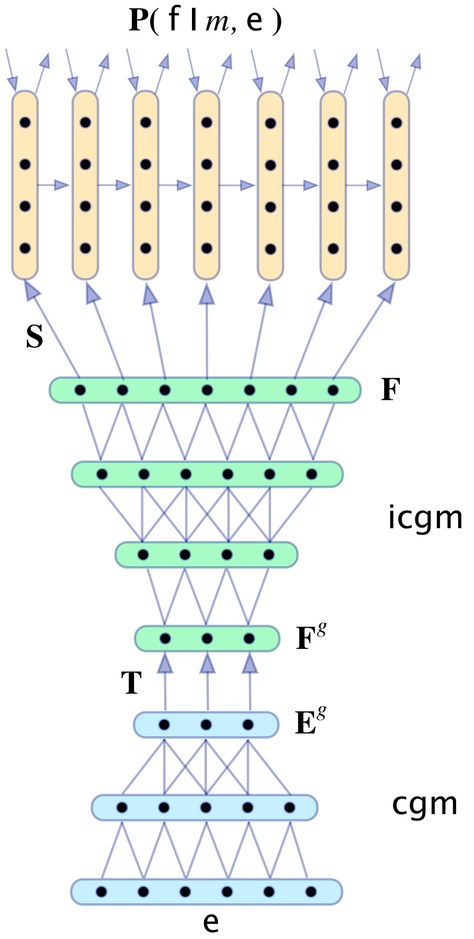
Convolutional Sequence-to-Sequence Models

- Convolutions can be parallelized easily (unlike RNNs)
- Can be stacked to model more dependencies
- Naturally build hierarchical representations

Kalkbrenner & Blunsom 2013



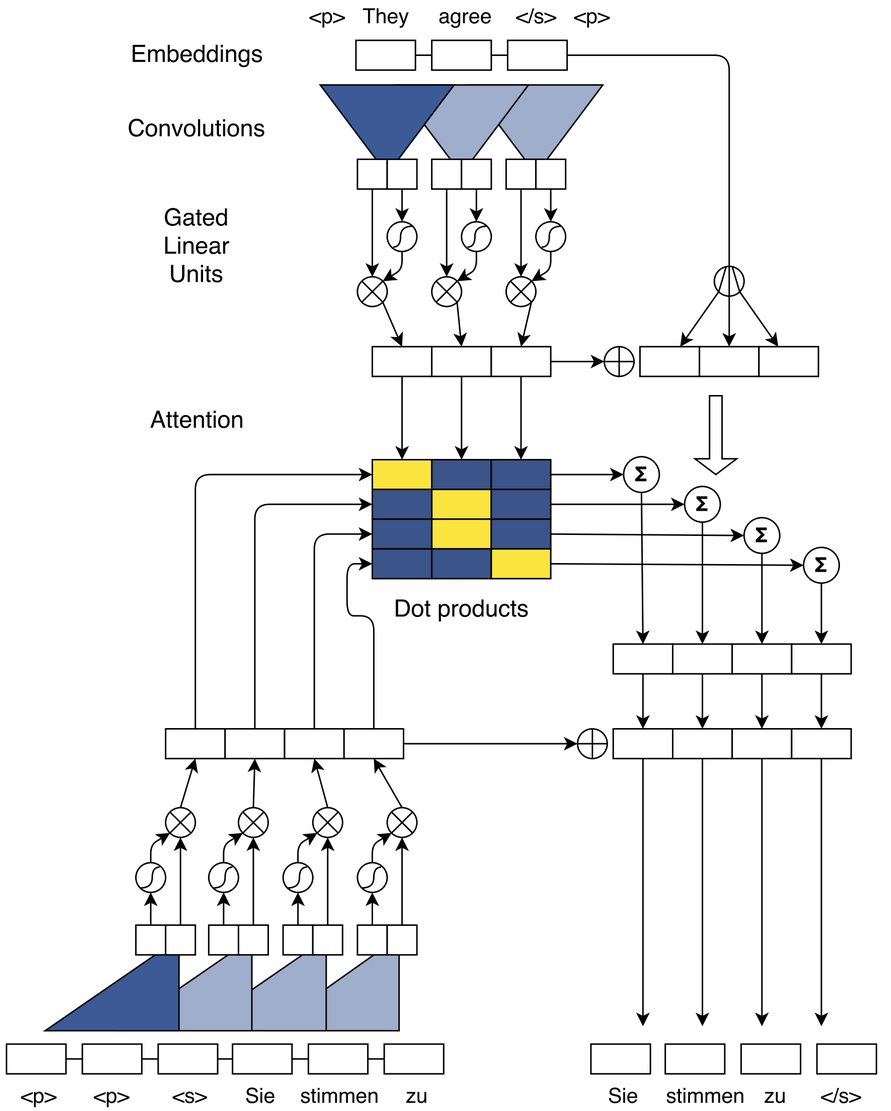
RCTM I



RCTM II

Convolutional Sequence-to-Sequence Models

- Add positional embeddings to inputs
- Input field limited by kernel width k and number of layers n
- Residual connections between layers
- Careful normalization (uniform variance)
- Multi-attention (computed in each layer)
- Source embeddings provided to attention
- Decoder has access to attention history

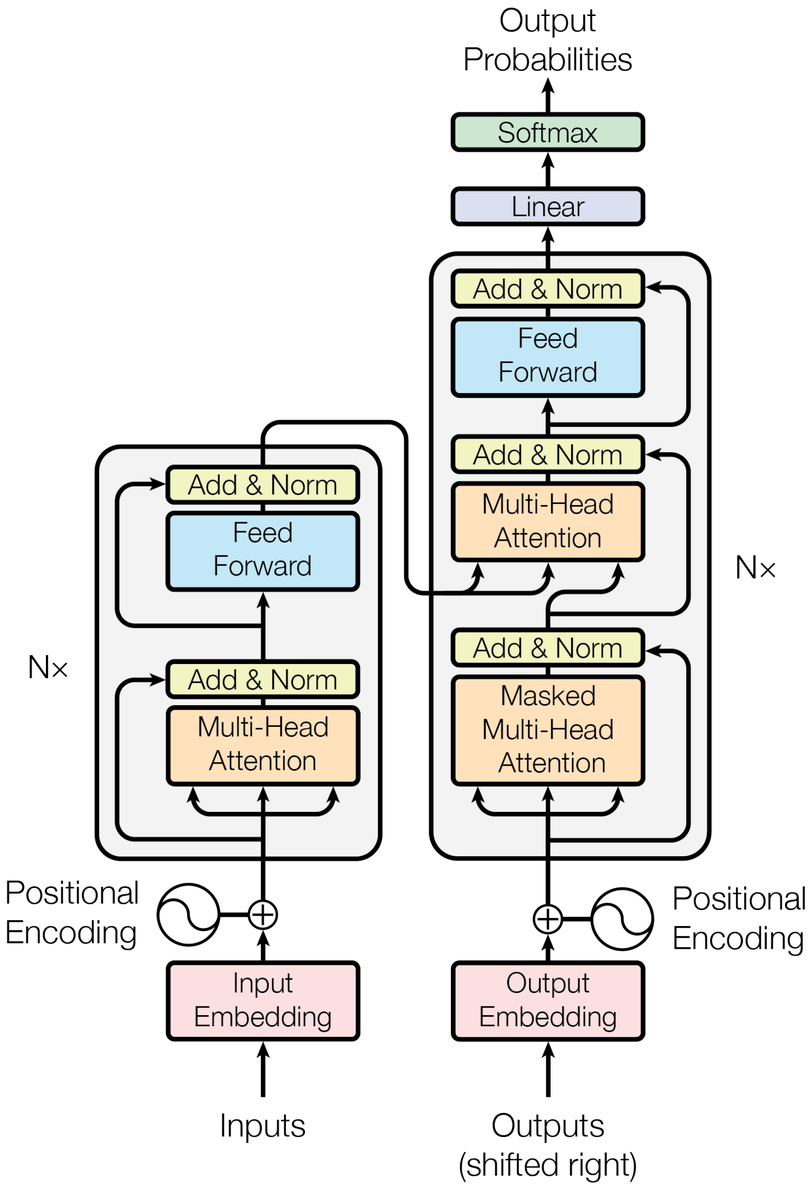


Transformer

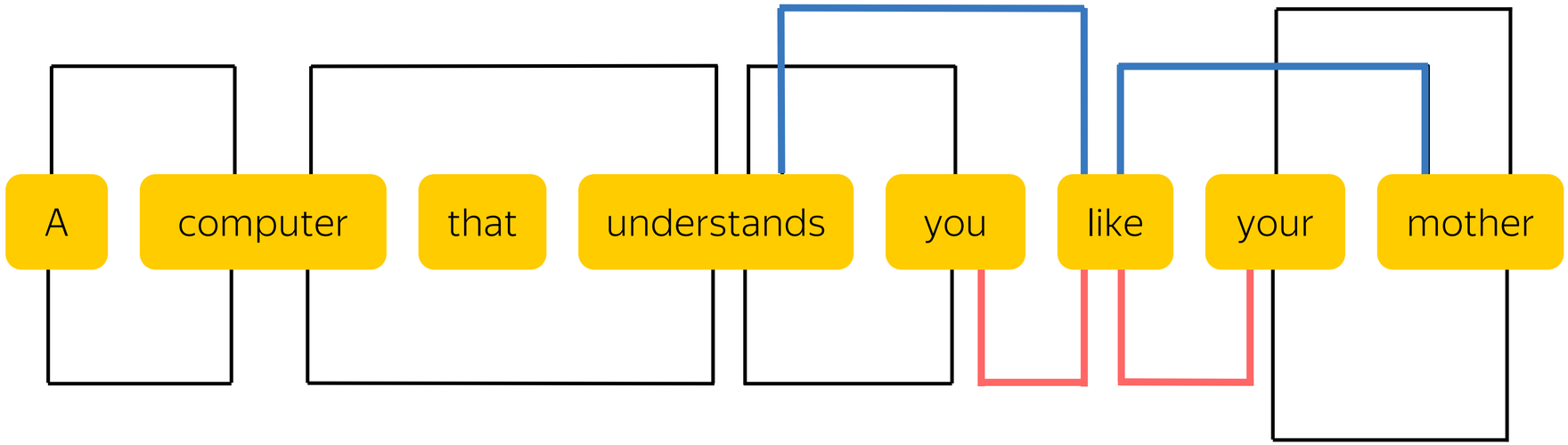
https://ai.googleblog.com/2017/08/transformer-novel-neural-network.html

Transformer

- Lots going on in the paper
- Use feed-forward network for sequence-to-sequence task
- Self-attention
- Multi-head attention
- Label smoothing
- Layer normalization
- …



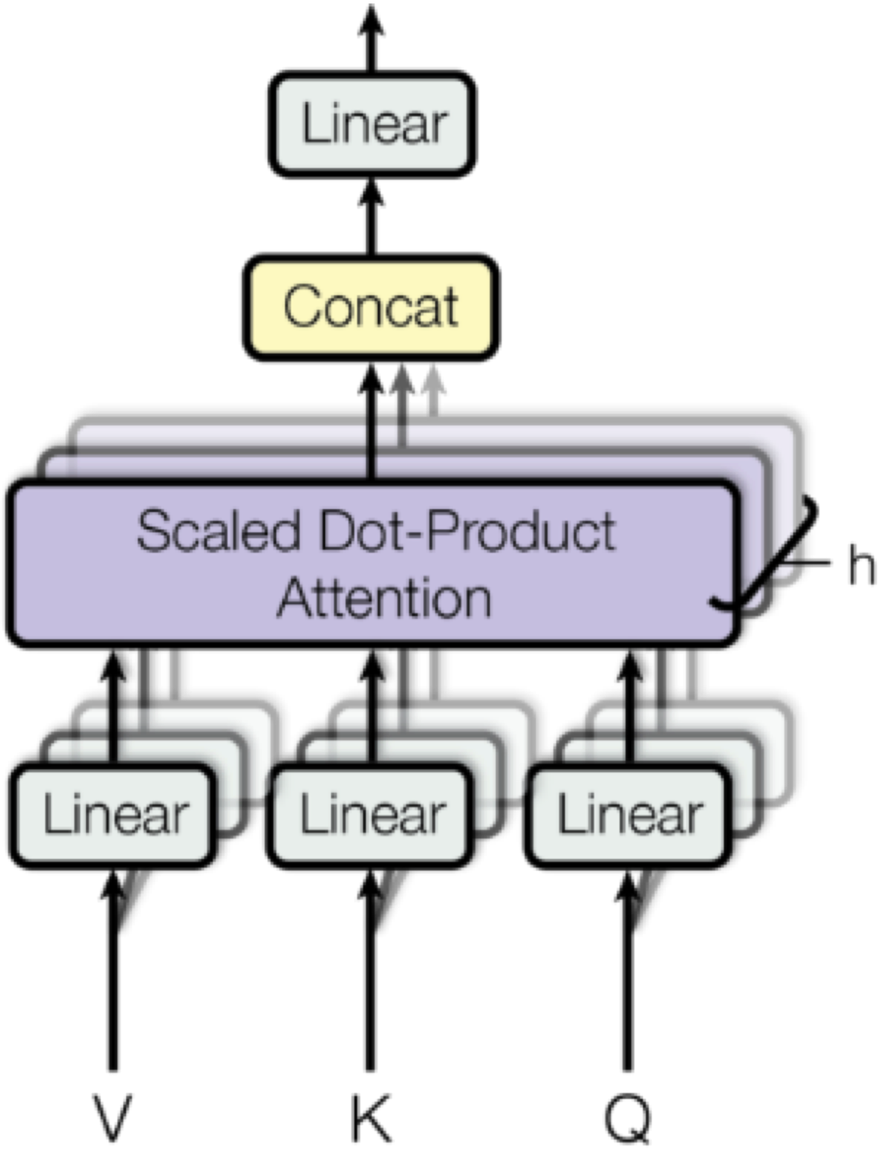
Self Attention



Multi-Head Attention

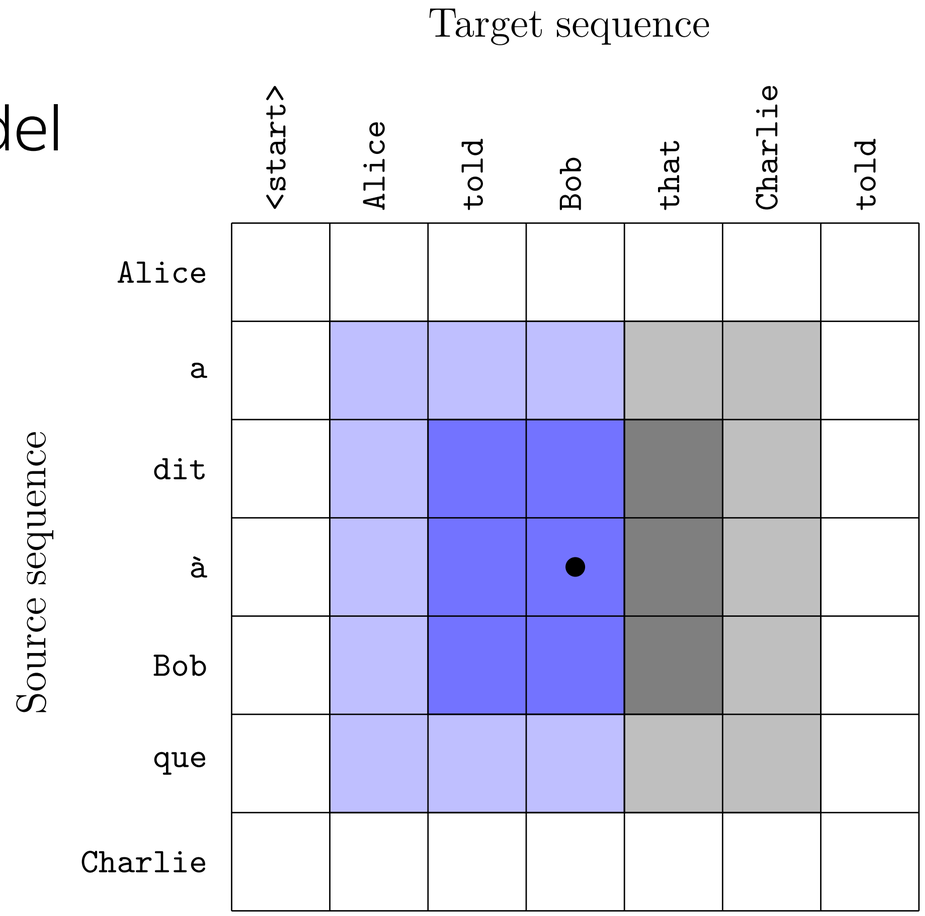
Она руководит **новым** проектом

- Gender agreement
- Case government
- Lexical preferences
- …



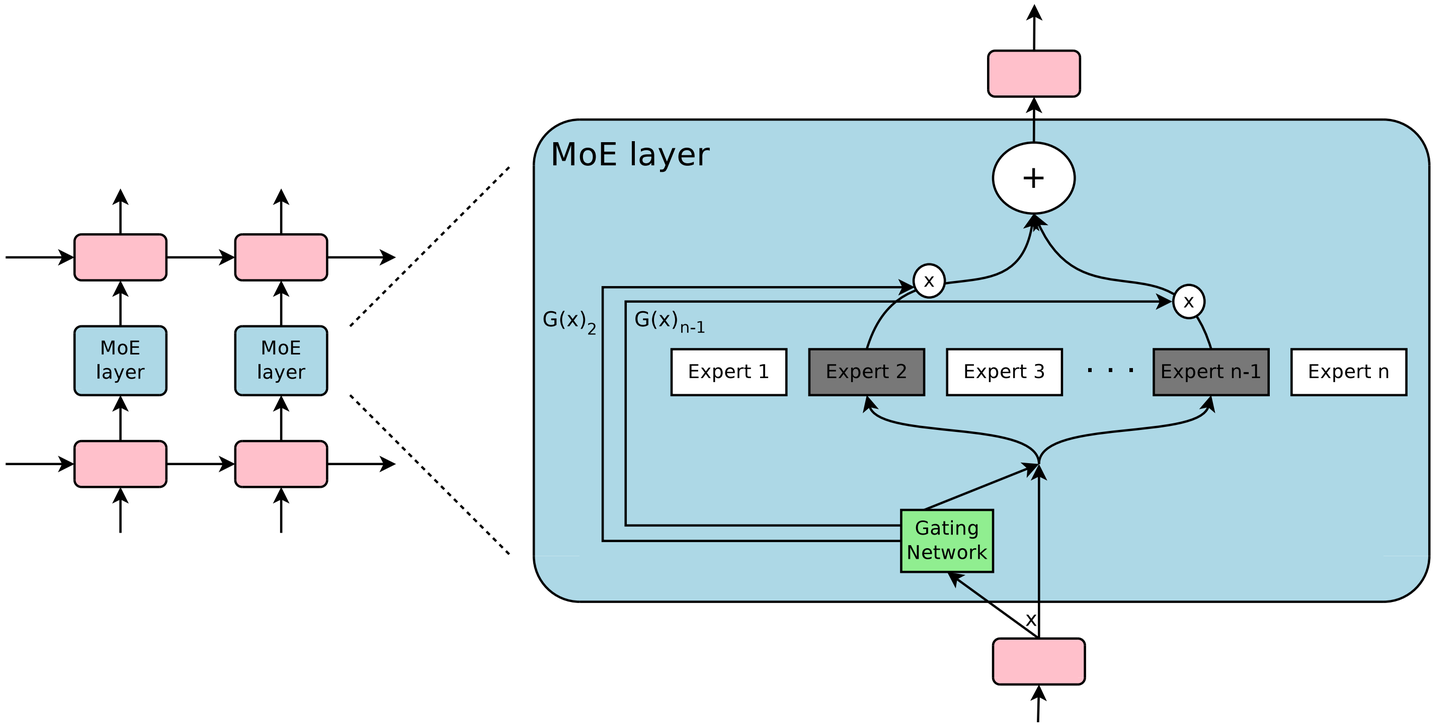
Pervasive Attention

- 2D convolutional sequence to sequence model
- Recomputes source sentence encoding based on target words generated so far
- Combined max pooling with self-attention
- Seems to work well on short sequences maybe not so well on longer ones



Mixtures of Experts

- Conditional computation to increase model capacity efficiently
- Noisy top-K gate: Sparse mixture of models (experts)
- Requires both data and model parallelism to work



Deliberation Networks

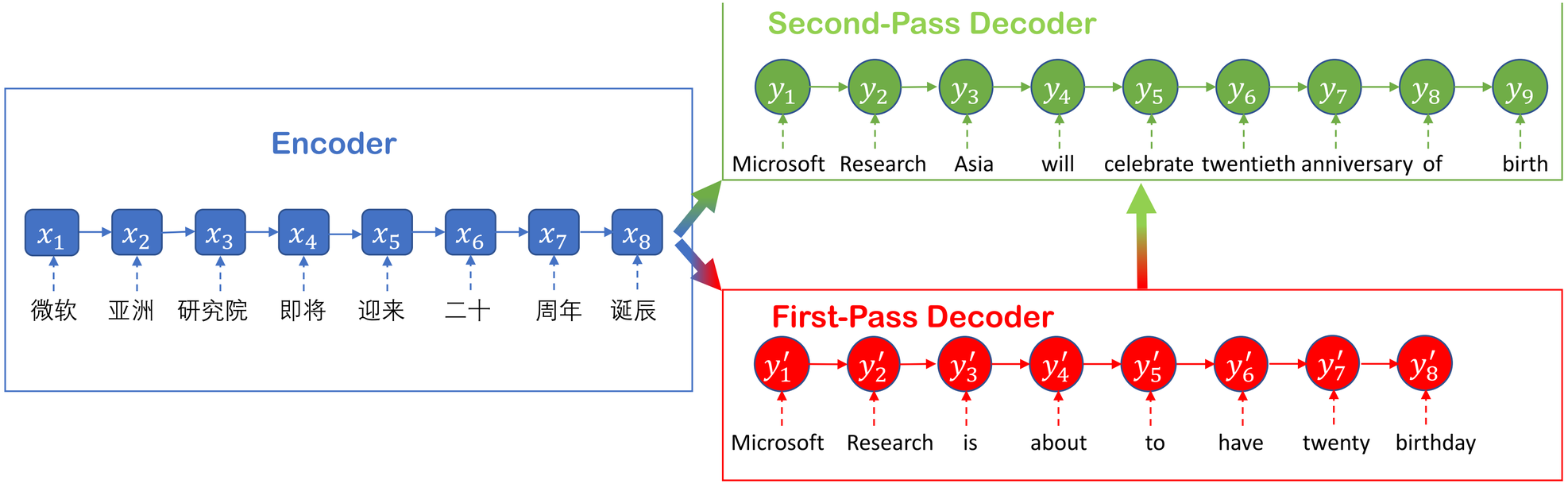
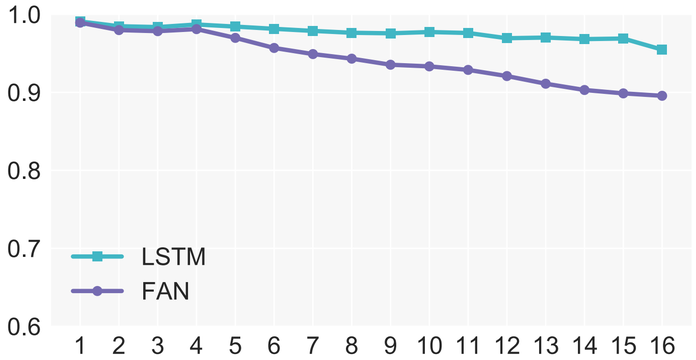


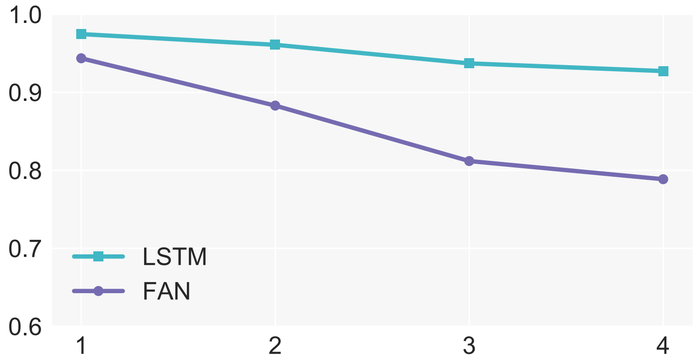
Figure 3: An example showing the decoding process of deliberation network.

Are RNNs Necessary?

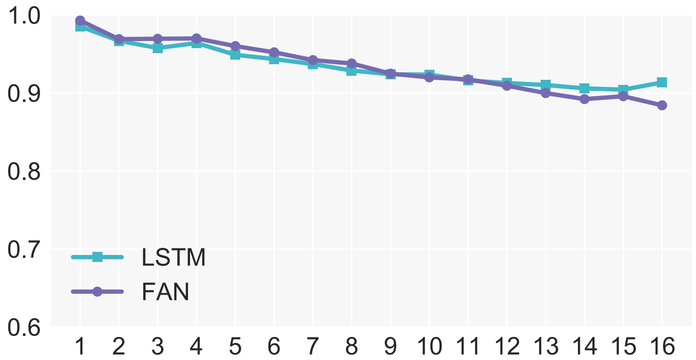
- Language is hierarchical (i.e. it allows embedding)
- Subject-verb agreement



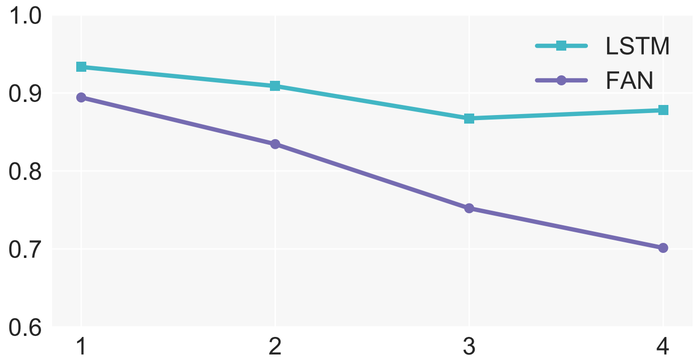
(a) Language model, breakdown by distance



(b) Language model, breakdown by # attractors



(c) Number prediction, breakdown by distance

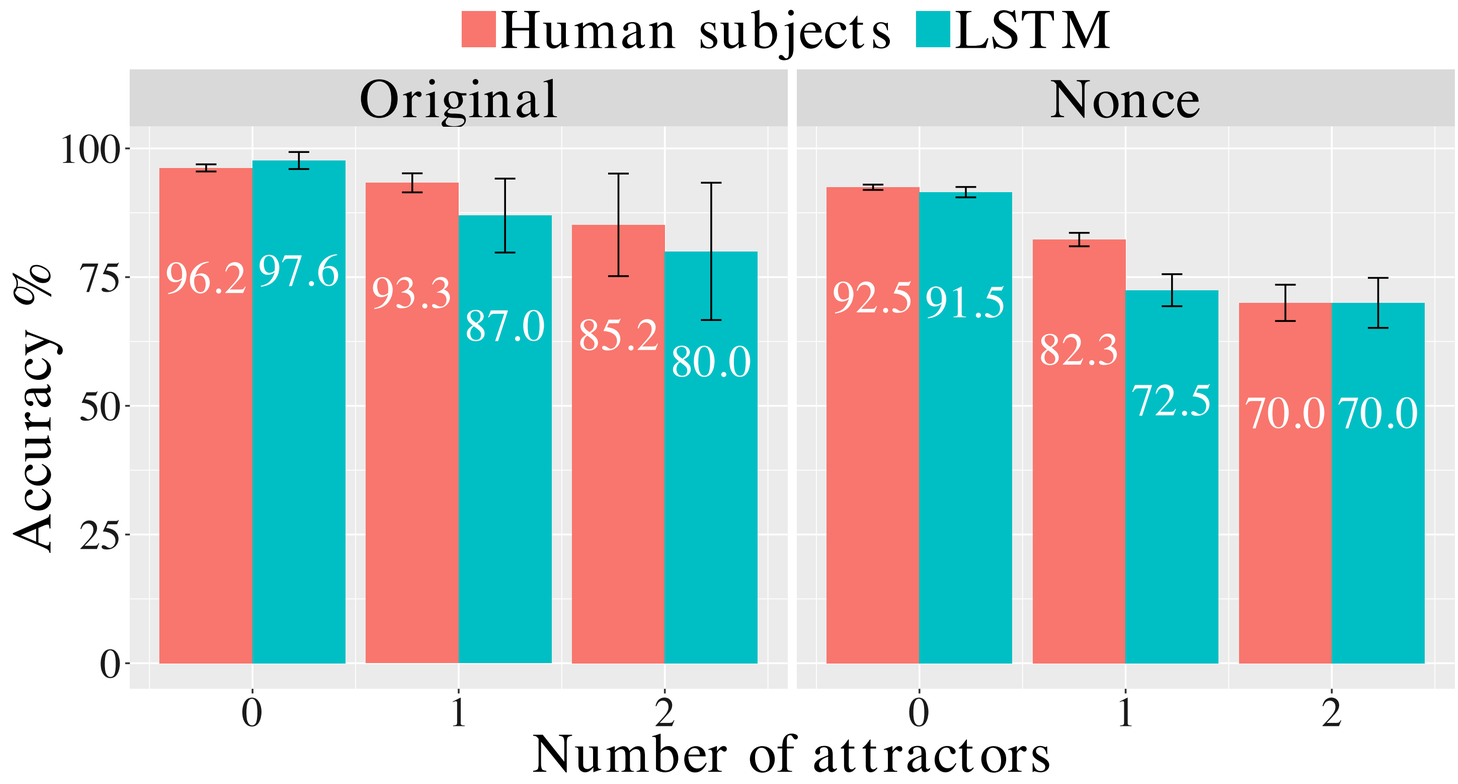


(d) Number prediction, breakdown by # attractors

Figure 2: Results of subject-verb agreement with different training objectives.

Are RNNs Necessary?

- As good as humans on non-sensical data :)



Maximum Path Length

Architecture	Complexity per layer	Sequential operations	Maximum path length
Self-Attention	O(n^2 d)	O(1)	O(1)
Recurrent	O(n d^2)	O(n)	O(n)
Convolutional	O(kn d^2)	O(1)	O(log_k(n))

How can monolingual data be used to improve NMT?

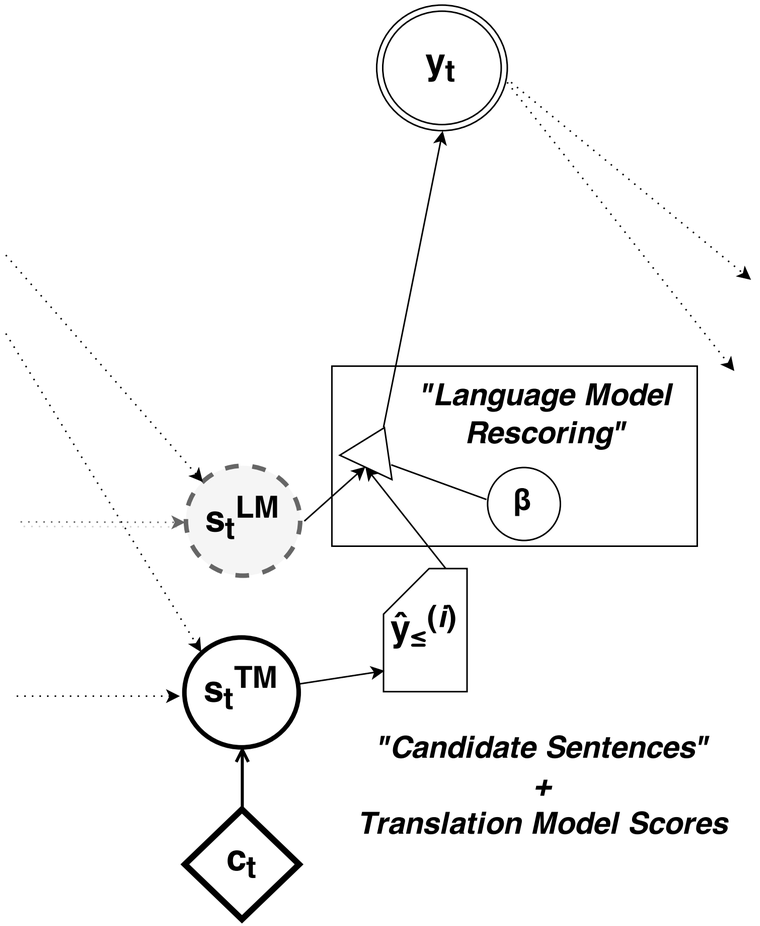
Language Models

- Shallow fusion

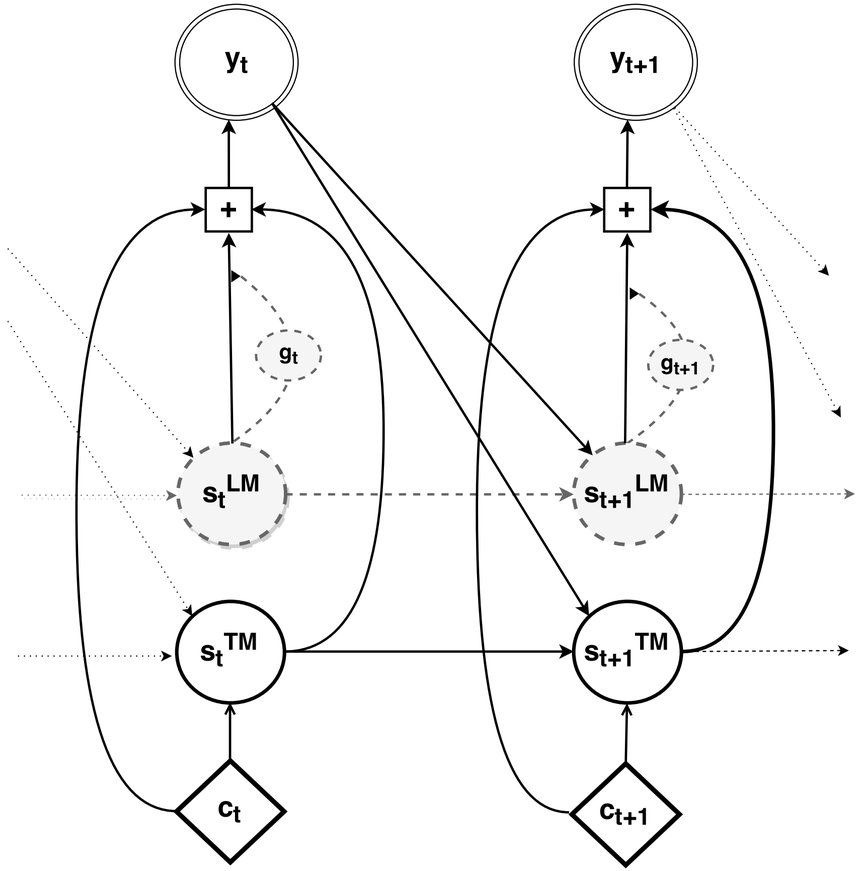
$$\log \Pr(e|f) = \log Pr_{tm}(e|f) + \beta \; log \; Pr_{lm}(e)$$

- Deep fusion integrates hidden state of RNNLM with NMT decoder
- Models are training separately, output layers parameters are fine-tuned
- Gate is added to learn when to pay attention to LM

Language Models



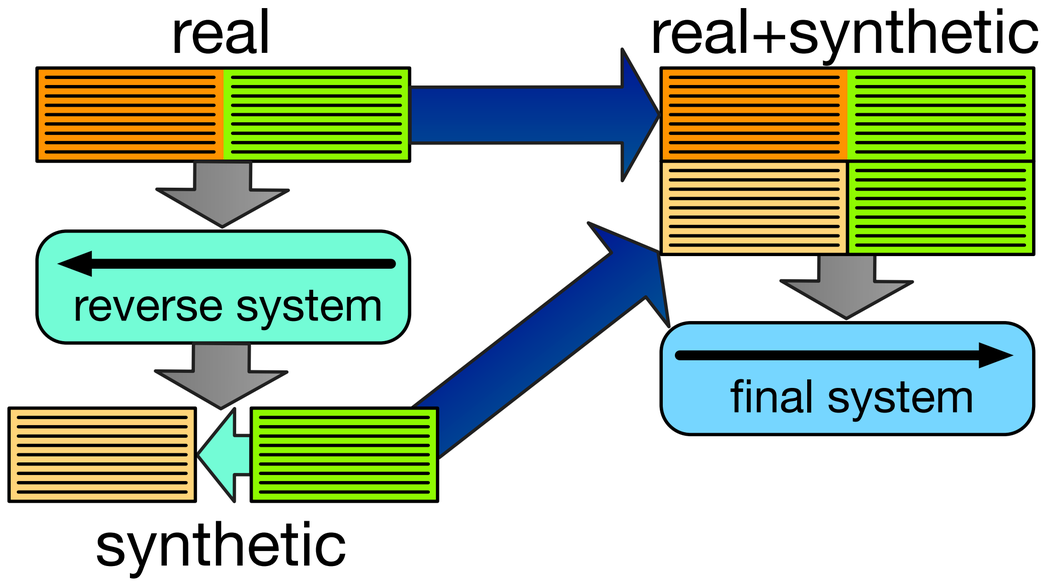
(a) Shallow Fusion (Sec. 4.1)



(b) Deep Fusion (Sec. 4.2)

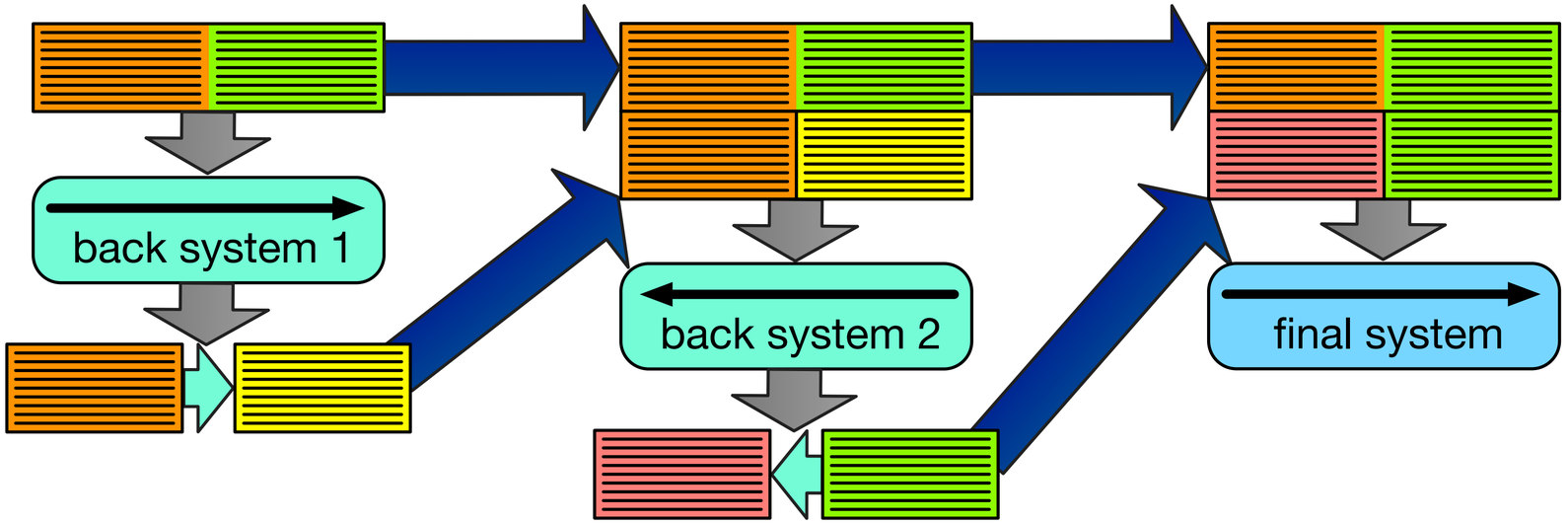
Back-translation

- Build a corpus by translating from target to source
- Why does it work better than self-training?
- Is it just domain adaptation?
- Quality of back-translation model matters
- Ratio of 1:3 real to synthetic data



Back-translation++

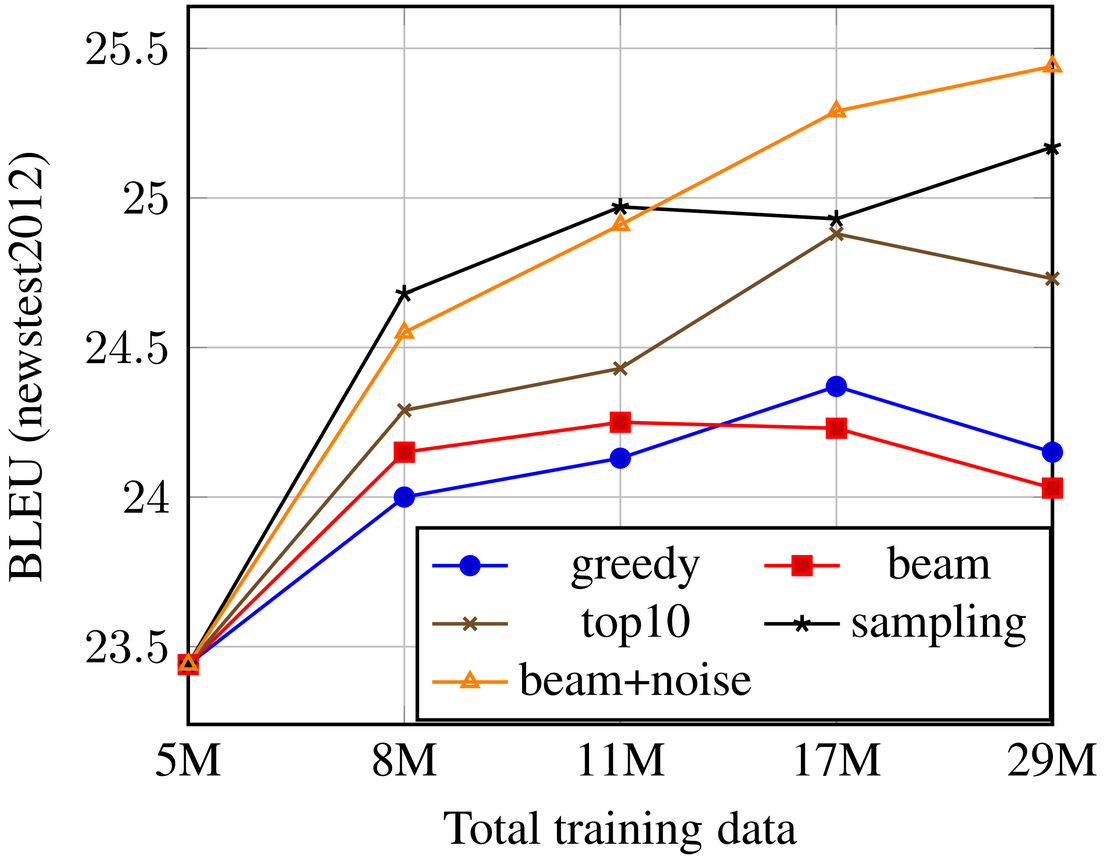
- Iterative back-translation as ‘co-training’
- Beat SOTA on ‘big’ data track WMT by 1+ BLEU
- Works also for PBMT (Moses)

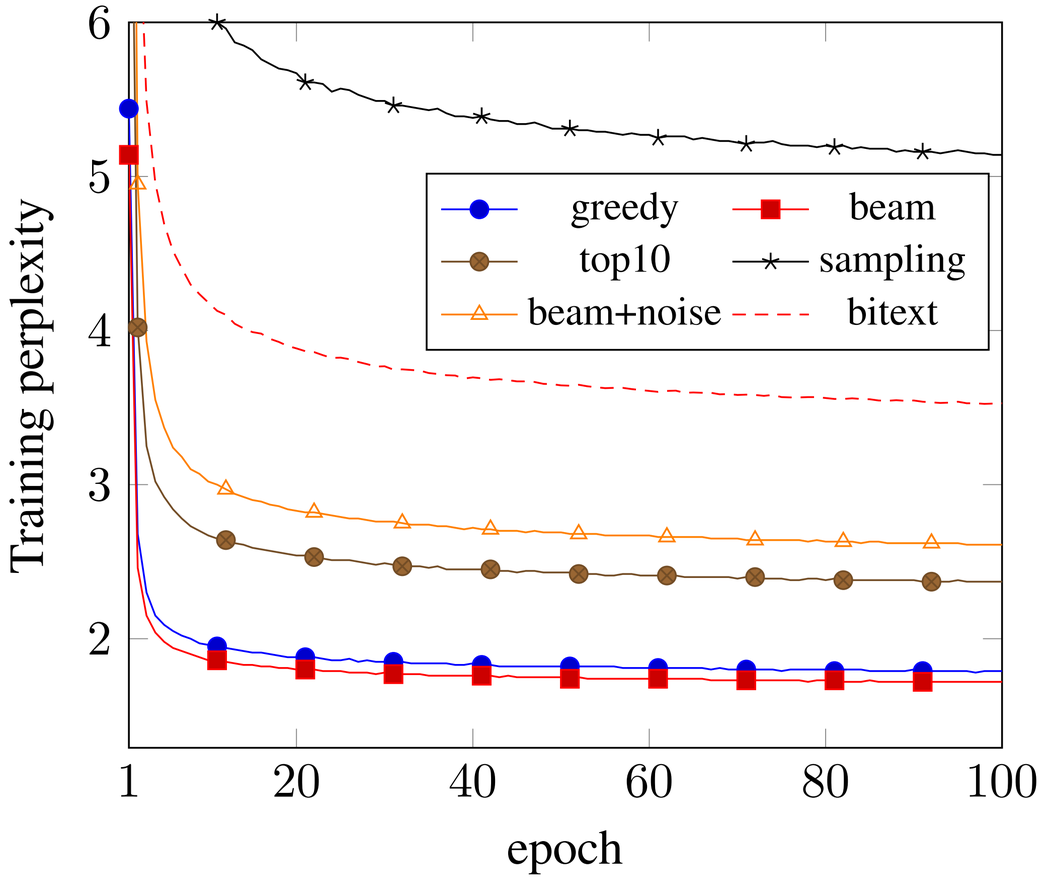


Understanding Back-Translation

- Sampling better than greedy/beam inference for back-translation
- MAP solution doesn't model true distribution well
- Compromise between MAP and unrestricted sampling:
  sample from top-**k** of beam
- Adding noise to input sentence (target) prior to back-translation

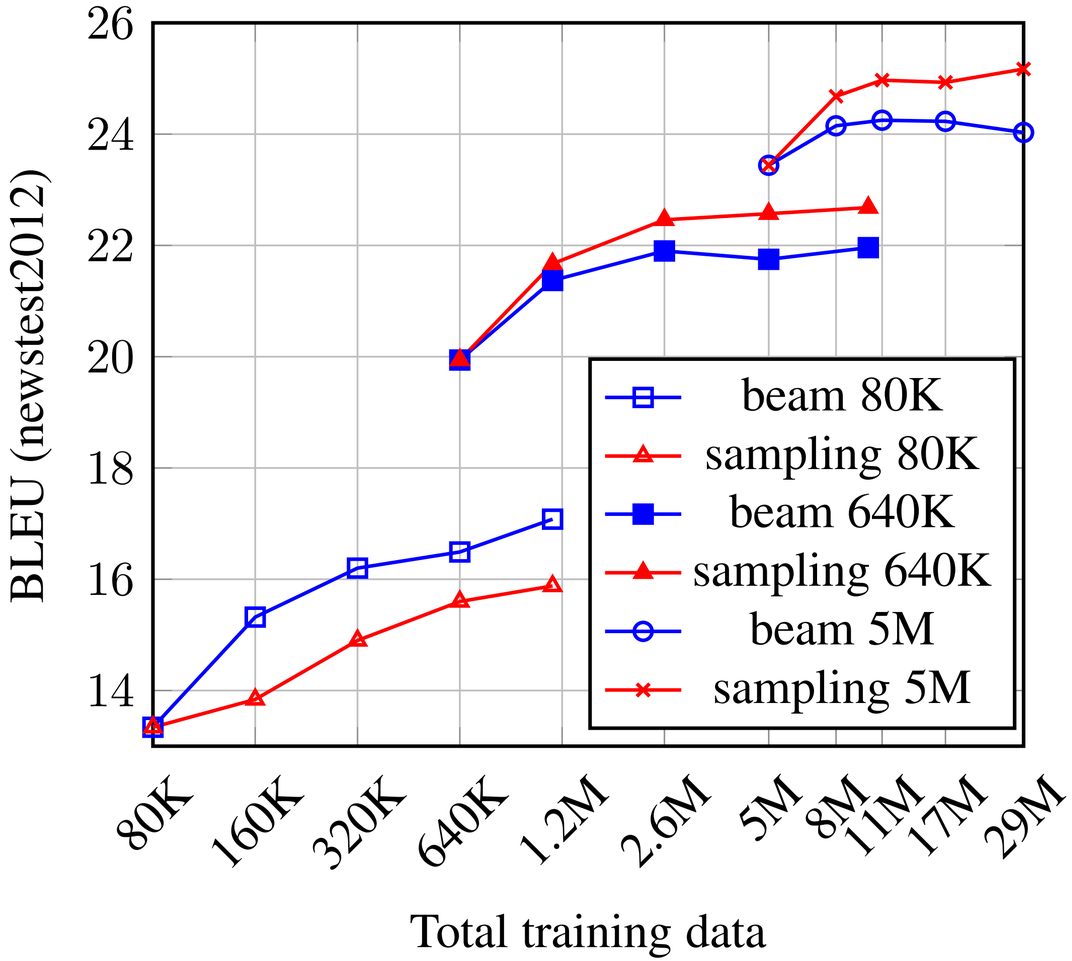
Understanding Back-Translation





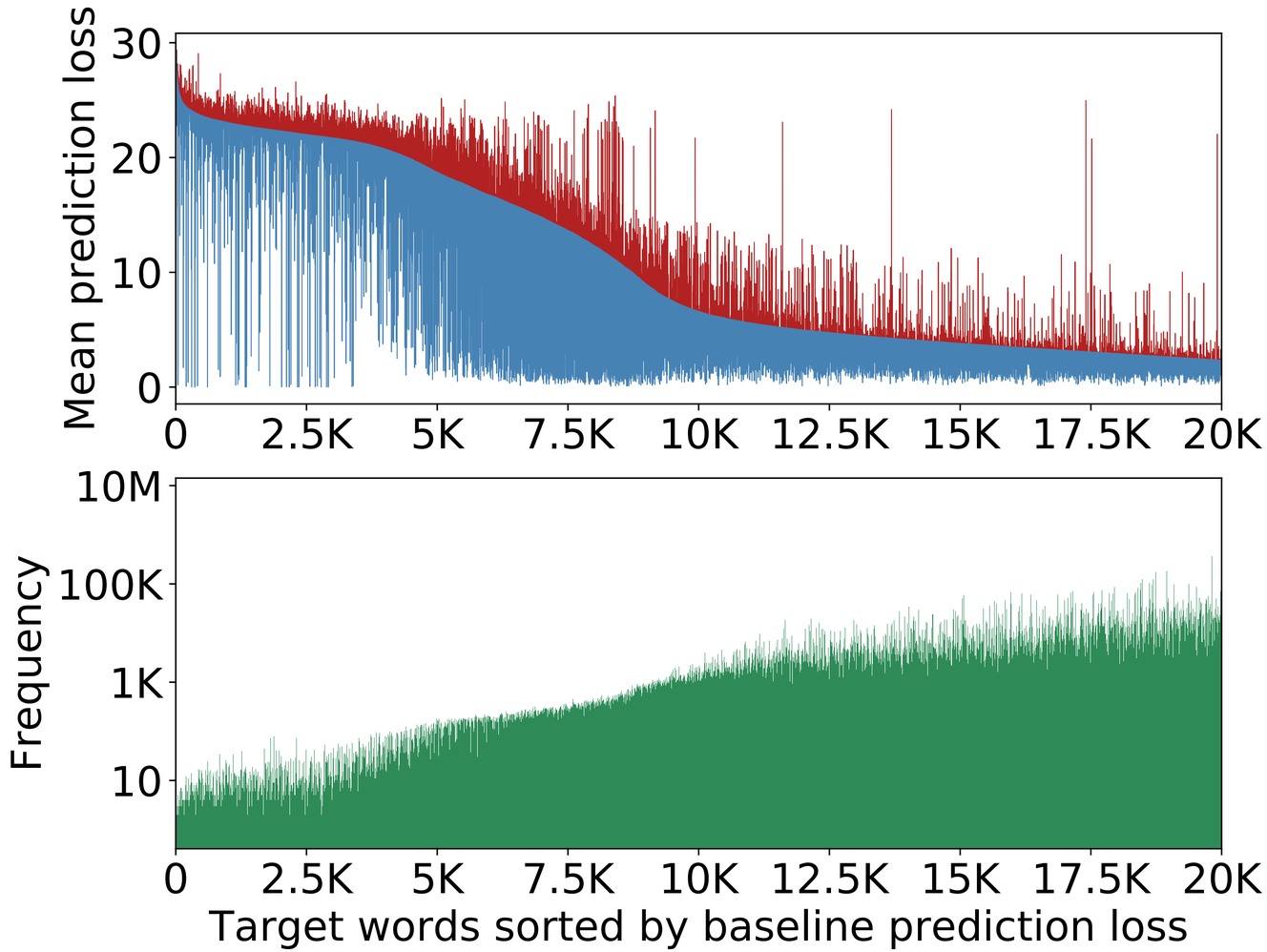
Understanding Back-Translation

- MAP data is too predictable
- Sampled data sometimes poor quality
- Diversity more important than accuracy unless in low-resource setting
- Up-sampling real data helps MAP methods



Back-Translation for Difficult Words

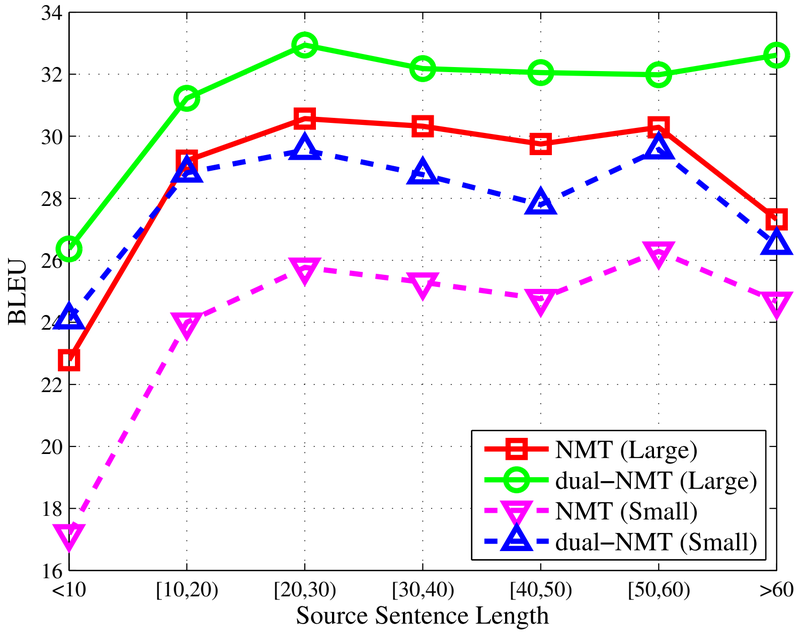
- Re-training on synthetic data redistributes loss more evenly
- So focus on rare/hard words
- Use mean prediction loss
- Best results sampled difficult words based on contexts close in embedding space



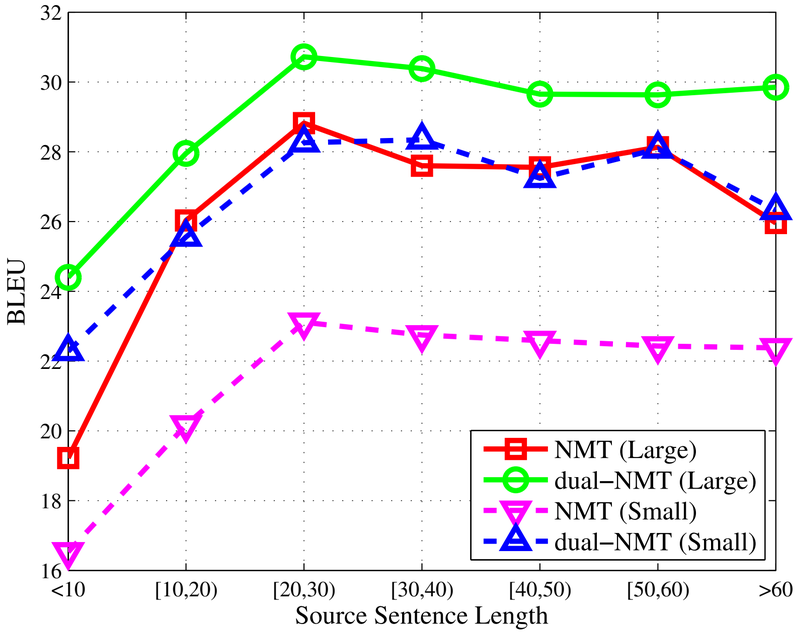
How can we build NMT systems for language pairs with very little parallel data?

Dual Learning

1. Translate from A to B with TM(B|A)
2. Reward = LM(B) + TM(A|B) (i.e. language model + reconstruction cost)
3. Optimize with policy gradient

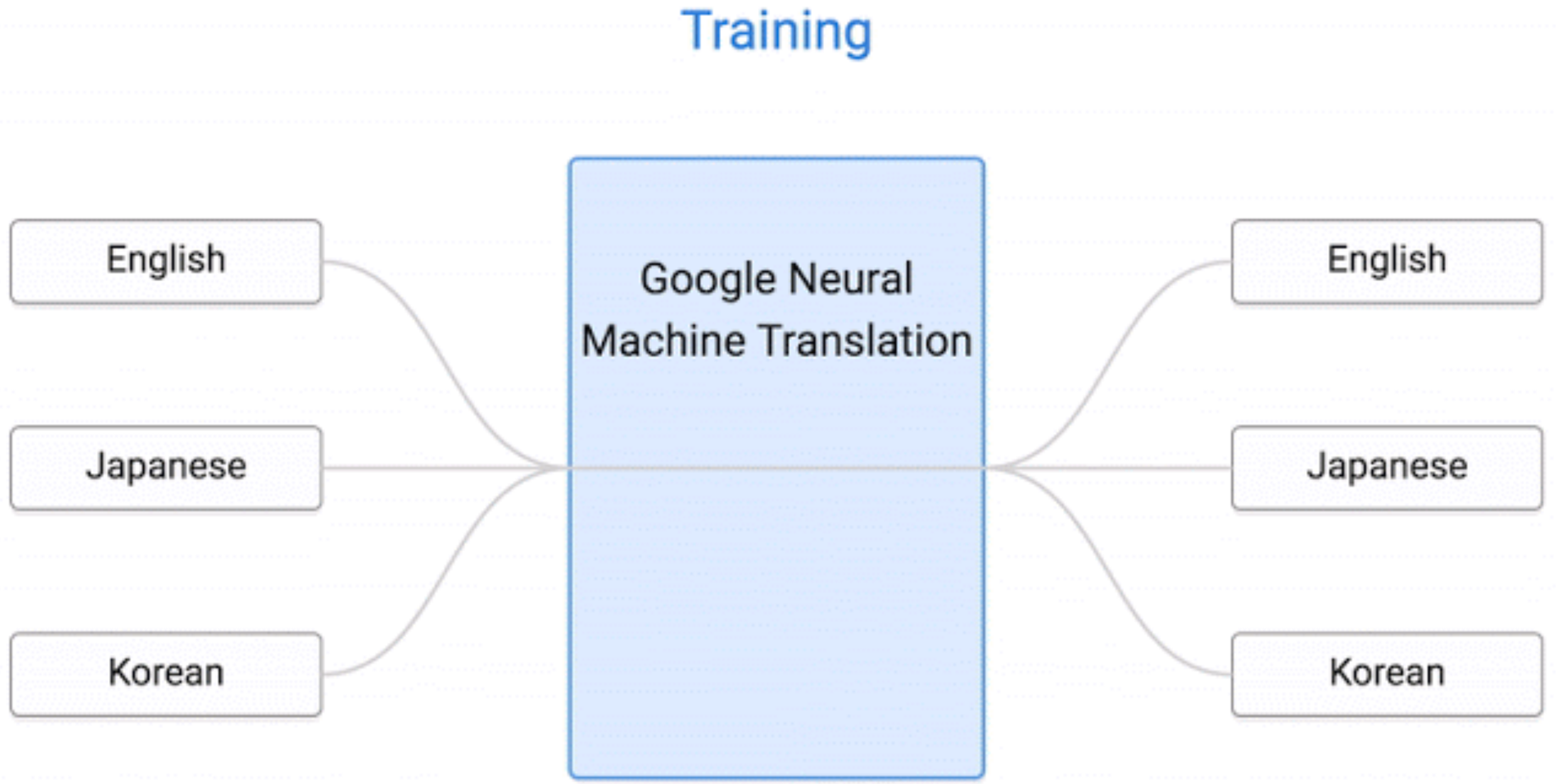


(a) En→Fr



(b) Fr→En

Zero Shot


Training
English
Japanese
Korean
Google Neural Machine Translation
English
Japanese
Korean

Zero Shot Dual Learning

- Start with zero-shot then improve with dual learning
- Within 2 BLEU of NMT trained on 1M sentences

	Phrase-based	**NMT-0**	**NMT-S**	**NMT-F**	**Dual-0**	**Dual-S**
Aligned Data	en-fr (11M) en-es (11M) es-fr (11M)	en-fr (1M) en-es (1M)	en-fr (1M) en-es (1M) es-fr (10k)	en-fr (1M) en-es (1M) es-fr (1M)	en-fr (1M) en-es (1M)	en-fr (1M) en-es (1M) es-fr (10k)
Monol. Data					es (0.5M) fr (0.5M)	es (0.5M) fr (0.5M)
en→es	61.26	49.00	43.33	44.06	37.05	38.74
es→en	59.89	49.67	40.17	18.24	32.84	32.03
en→fr	50.09	37.88	33.71	34.75	29.58	30.89
fr→en	52.22	42.12	34.17	13.58	27.95	26.00
es→fr	52.44	10.02	33.10	37.67	35.54	35.63
fr→es	49.79	6.25	38.33	40.85	38.83	39.00

Table 1: BLEU scores on the UN corpus test set. Each line reports the BLEU scores of the corresponding translation direction. The first column refers to the phrase-based model of Ziemski et al. [20]. All NMT and Dual models are trained on 1M en-es and 1M en-fr aligned sentences, used in both directions. The NMT-S (small) model is trained additionally on 10k es-fr aligned sentences, while NMT-F (full) is trained additionally on 1M es-fr sentences. The Dual-0 model does not use any es-fr aligned data, while Dual-S is trained starting from NMT-S.

Shared Attention

- One encoder and one decoder per language: O(L) rather than O(L^2)
- Language independent attention aids transfer learning also
- Add mappings to and from attention state
- Shared attention always better in low-resource setting

			Fr (39m)		Cs (12m)		De (4.2m)		Ru (2.3m)		Fi (2m)	
		Dir	→ En	En →	→ En	En →	→ En	En →	→ En	En →	→ En	En →
(a) BLEU	Dev	Single	27.22	26.91	21.24	15.9	24.13	20.49	21.04	18.06	13.15	9.59
		Multi	26.09	25.04	21.23	14.42	23.66	19.17	21.48	17.89	12.97	8.92
	Test	Single	27.94	**29.7**	20.32	**13.84**	24	**21.75**	22.44	**19.54**	12.24	**9.23**
		Multi	**28.06**	27.88	**20.57**	13.29	**24.20**	20.59	**23.44**	19.39	**12.61**	8.98
(b) LL	Dev	Single	-50.53	-53.38	-60.69	-69.56	-54.76	-61.21	-60.19	-65.81	-88.44	-91.75
		Multi	-50.6	-56.55	-54.46	-70.76	-54.14	-62.34	-54.09	-63.75	-74.84	-88.02
	Test	Single	-43.34	**-45.07**	-60.03	**-64.34**	-57.81	**-59.55**	-60.65	-60.29	-88.66	-94.23
		Multi	**-42.22**	-46.29	**-54.66**	-64.80	**-53.85**	-60.23	**-54.49**	**-58.63**	**-71.26**	**-88.09**

Table 3: (a) BLEU scores and (b) average log-probabilities for all the five languages from WMT'15.

Contextual Parameter Generator

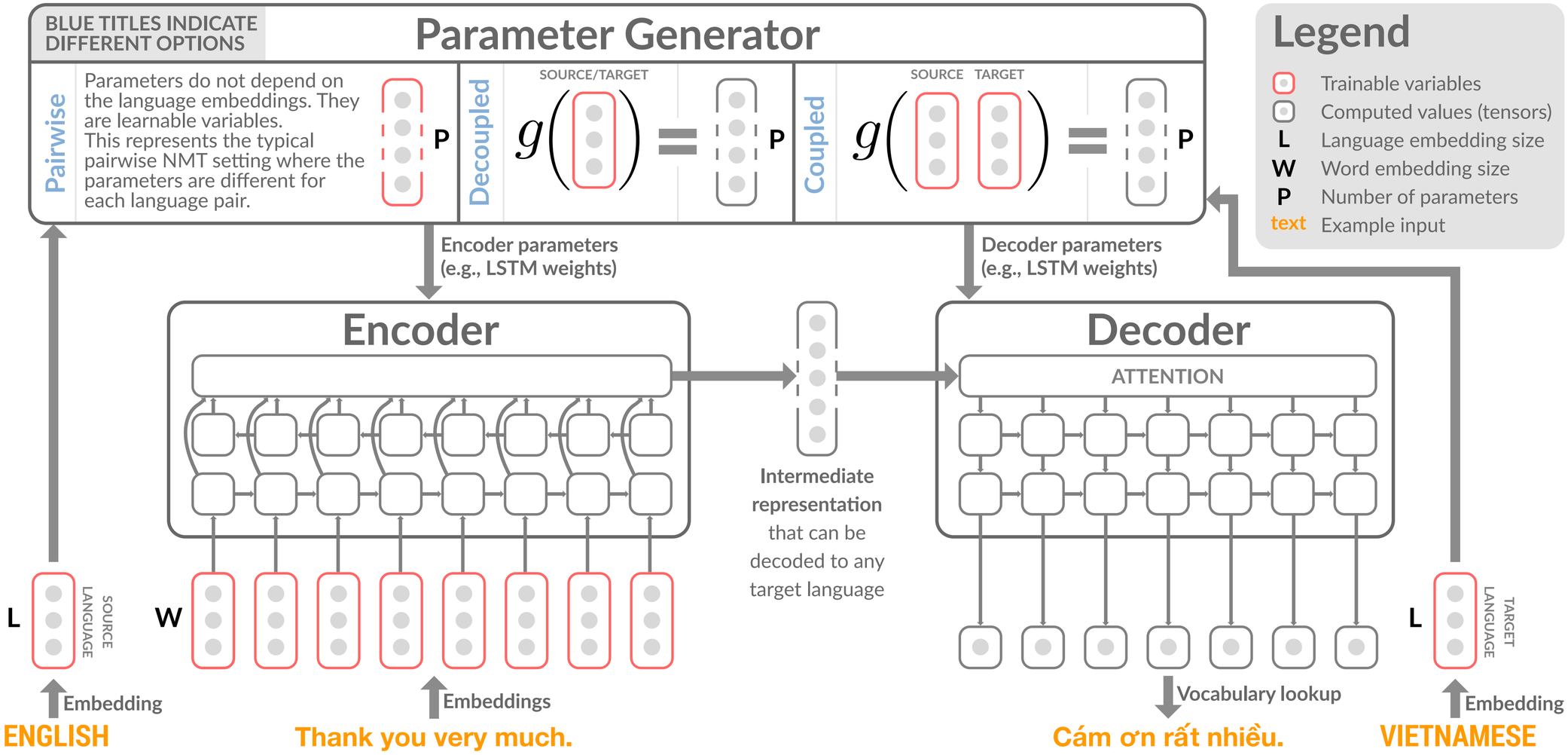
Given S source languages and T target languages

- Pairwise: standard NMT — O(ST)
- Per-language: shared attention — O(S+T)
- Universal — O(1)

Claims:

- Universal has high sample complexity (under-parameterized)
- Pairwise and per-language encoder/decoder are over-parameterized
- Embedding ‘language’ in same space as words is unintuitive

Contextual Parameter Generator



Contextual Parameter Generator

- Decoupled: generate encoder (and decoder) given source (target)
- Coupled: generate encoder and decoder given both source and target

$$g^{(enc)}(l_s) \equiv \boldsymbol{W}^{(enc)} l_s$$

- Low rank approximation of $\boldsymbol{W}^{(enc)}$ for groups j

$$\theta_j^{(enc)}(l_t) \equiv \boldsymbol{W}_j^{(enc)} \boldsymbol{P}_j^{(enc)} l_s$$

- Use monolingual data via auto-encoding based on parameter sharing

Contextual Parameter Generator

		BLEU				Meteor			
		PNMT	GML	CPG*	CPG	PNMT	GML	CPG*	CPG
100% Parallel Data	En→Cs	14.89	15.92	16.88	**17.22**	19.72	20.93	21.51	**21.72**
	Cs→En	24.43	25.25	26.44	**27.37**	27.29	27.46	28.16	**28.52**
	En→De	25.99	15.92	26.41	**26.77**	44.72	42.97	45.97	**46.30**
	De→En	30.93	29.60	31.24	**31.77**	30.73	29.90	30.95	**31.13**
	En→Fr	38.25	34.40	38.10	**38.32**	57.43	53.86	57.42	**57.68**
	Fr→En	37.40	35.14	37.11	**37.89**	34.83	33.14	34.34	**34.89**
	En→Th	23.62	22.22	26.03	**26.33**	-	-	-	-
	Th→En	15.54	14.03	16.54	**26.77**	21.58	21.02	22.78	**23.05**
	En→Vi	27.47	25.54	28.33	**29.03**	-	-	-	-
	Vi→En	24.03	23.19	25.91	**26.38**	27.59	26.96	28.23	**28.79**
	Mean	26.26	24.12	27.30	**27.80**	32.98	32.03	33.67	**34.01**
10% Parallel Data	En→Cs	5.71	8.18	8.40	**9.49**	12.18	14.97	15.25	**15.90**
	Cs→En	6.64	14.56	14.81	**15.38**	13.02	20.04	19.98	**20.87**
	En→De	11.70	14.60	15.09	**16.03**	29.98	33.74	34.88	**36.19**
	De→En	18.10	19.02	19.77	**20.25**	22.57	23.27	23.65	**24.40**
	En→Fr	24.47	25.15	24.00	**25.79**	44.10	44.84	44.95	**46.22**
	Fr→En	23.79	25.02	24.55	**27.12**	26.28	26.61	26.20	**28.18**
	En→Th	7.86	15.58	**18.41**	17.65	-	-	-	-
	Th→En	7.13	9.11	**10.19**	10.14	13.91	16.32	16.78	**16.92**
	En→Vi	18.01	17.51	**18.92**	18.90	-	-	-	-
	Vi→En	6.69	16.00	16.28	**16.86**	13.39	21.01	21.34	**22.28**
	Mean	13.01	16.47	17.04	**17.76**	21.93	25.10	25.38	**26.37**
1% Parallel Data	En→Cs	0.49	1.25	1.57	**2.38**	4.60	6.24	6.28	**8.38**
	Cs→En	1.10	1.76	1.87	**4.60**	6.29	7.13	7.08	**11.15**
	En→De	1.22	4.13	4.06	**6.46**	12.23	18.29	17.61	**23.83**
	De→En	1.46	3.42	3.86	**7.49**	7.58	8.79	8.95	**13.73**
	En→Fr	2.88	7.74	7.41	**12.45**	13.88	21.29	21.80	**30.36**
	Fr→En	4.05	5.22	5.06	**11.39**	9.58	9.86	9.83	**16.34**
	En→Th	1.22	5.72	8.01	**9.26**	-	-	-	-
	Th→En	1.42	1.66	1.65	**3.37**	6.08	7.22	5.89	**8.74**
	En→Vi	5.35	5.61	5.48	**8.00**	-	-	-	-
	Vi→En	2.01	3.57	3.64	**6.43**	7.86	8.76	8.48	**12.04**
	Mean	2.12	4.01	4.26	**7.18**	8.51	10.95	10.74	**15.58**

Contextual Parameter Generator

		BLEU							
		PNMT	GML	CPG^{8}	CPG^{8}_{C4}	CPG^{8}_{C2}	CPG^{8}_{C1}	CPG^{64}_{C8}	CPG^{512}_{C8}
Supervised	De→En	21.78	21.25	**22.56**	20.78	22.09	21.23	21.50	22.38
	De→It	13.16	13.84	**14.73**	14.34	14.43	13.84	14.34	14.11
	De→Ro	10.85	11.95	12.24	12.37	**12.72**	10.37	11.32	11.94
	En→De	**19.75**	17.06	19.41	19.04	18.42	17.04	17.46	19.29
	En→It	27.70	25.74	27.57	27.11	**28.21**	26.26	27.26	27.48
	En→Nl	24.41	22.46	24.47	**25.15**	24.64	23.94	24.48	24.50
	En→Ro	19.23	18.60	20.83	**20.96**	18.69	17.23	20.20	20.86
	It→De	14.39	12.76	14.61	**15.06**	14.15	13.12	14.18	14.69
	It→En	29.84	27.96	**30.62**	30.10	29.44	29.22	29.56	30.18
	It→Nl	16.74	16.27	17.99	**18.11**	18.05	17.13	17.71	17.99
	Nl→En	26.30	24.78	26.31	26.17	25.74	26.15	**26.33**	26.20
	Nl→It	16.03	16.10	16.81	**17.50**	17.03	16.81	16.89	17.09
	Nl→Ro	12.84	12.48	14.01	**14.44**	12.56	11.79	12.38	13.66
	Ro→De	12.75	12.21	13.58	**13.66**	13.02	12.62	12.96	13.63
	Ro→En	24.33	22.88	23.83	23.88	24.20	23.58	**24.65**	23.57
	Ro→Nl	13.70	14.11	15.34	**15.51**	15.11	14.65	15.29	15.19
	Mean	18.99	18.15	19.68	**19.75**	19.28	18.44	19.16	19.74
Zero-Shot	De→Nl	12.75	12.50	12.74	**12.80**	11.65	12.41	12.67	12.75
	It→Ro	9.97	9.57	10.57	10.17	10.42	9.65	**10.69**	10.32
	Nl→De	11.32	10.47	11.52	11.20	11.28	10.89	**11.63**	11.45
	Ro→It	11.69	10.82	11.51	11.40	11.66	11.42	**11.78**	11.27
	Mean	11.43	10.84	11.59	11.39	11.25	11.09	**11.69**	11.44

Contextual Parameter Generator

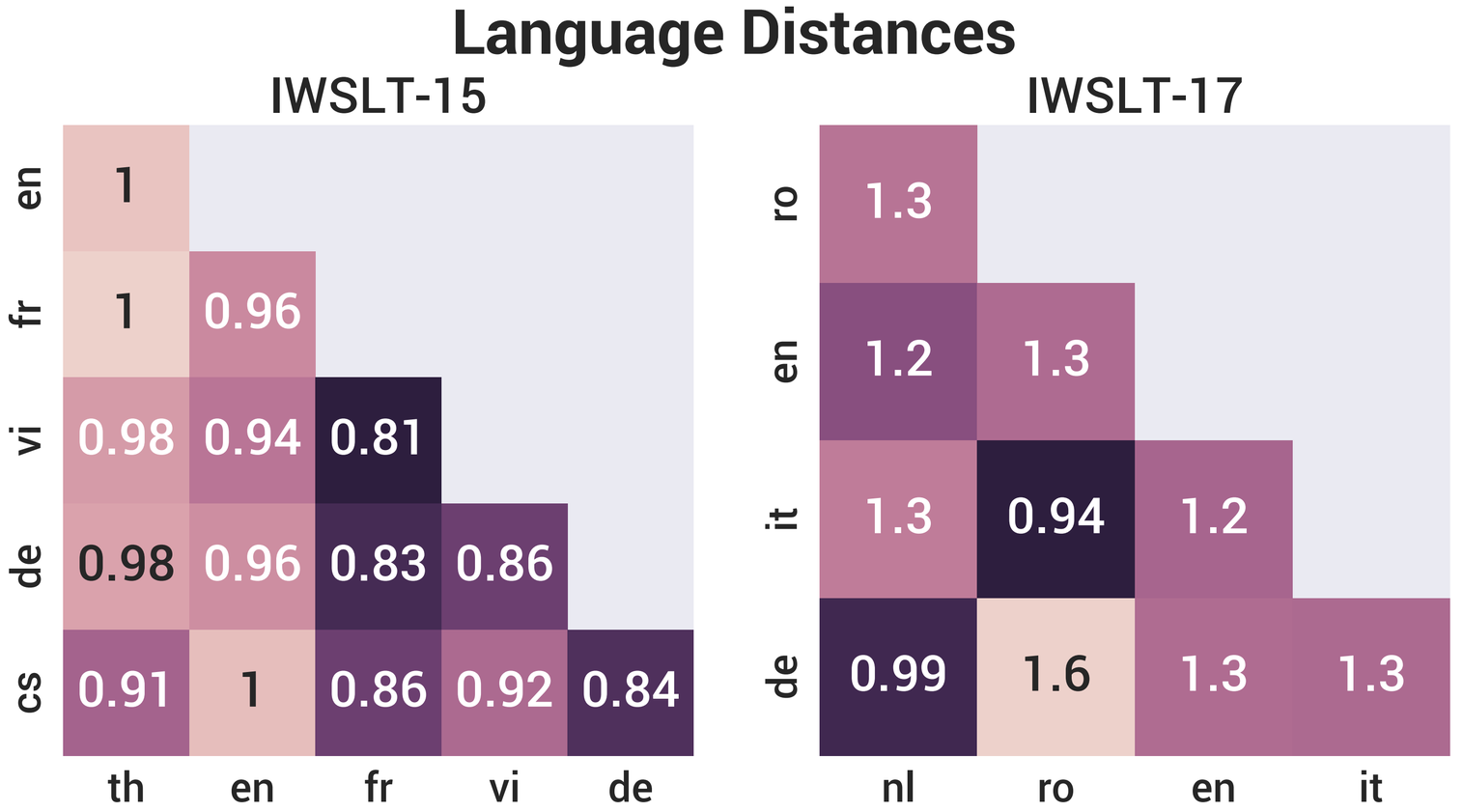


Figure 2: Pairwise cosine distance for all language pairs in the IWSLT-15 and IWSLT-17 datasets. **Darker** colors represent more similar languages.

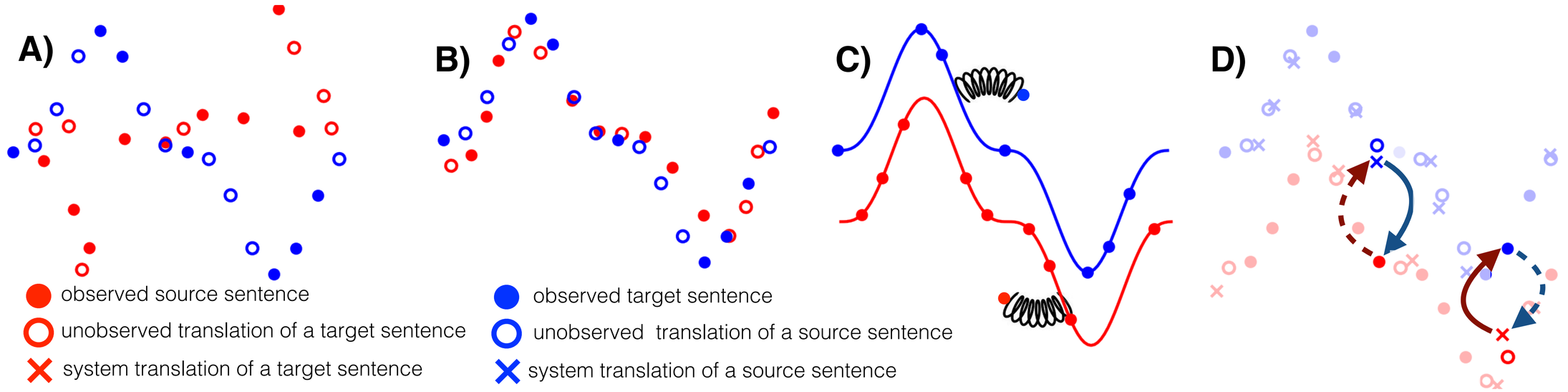
Unsupervised NMT and PBMT

(A) Monolingual corpora

(B) Initialization of translation model (joint BPE, aligned embeddings?)

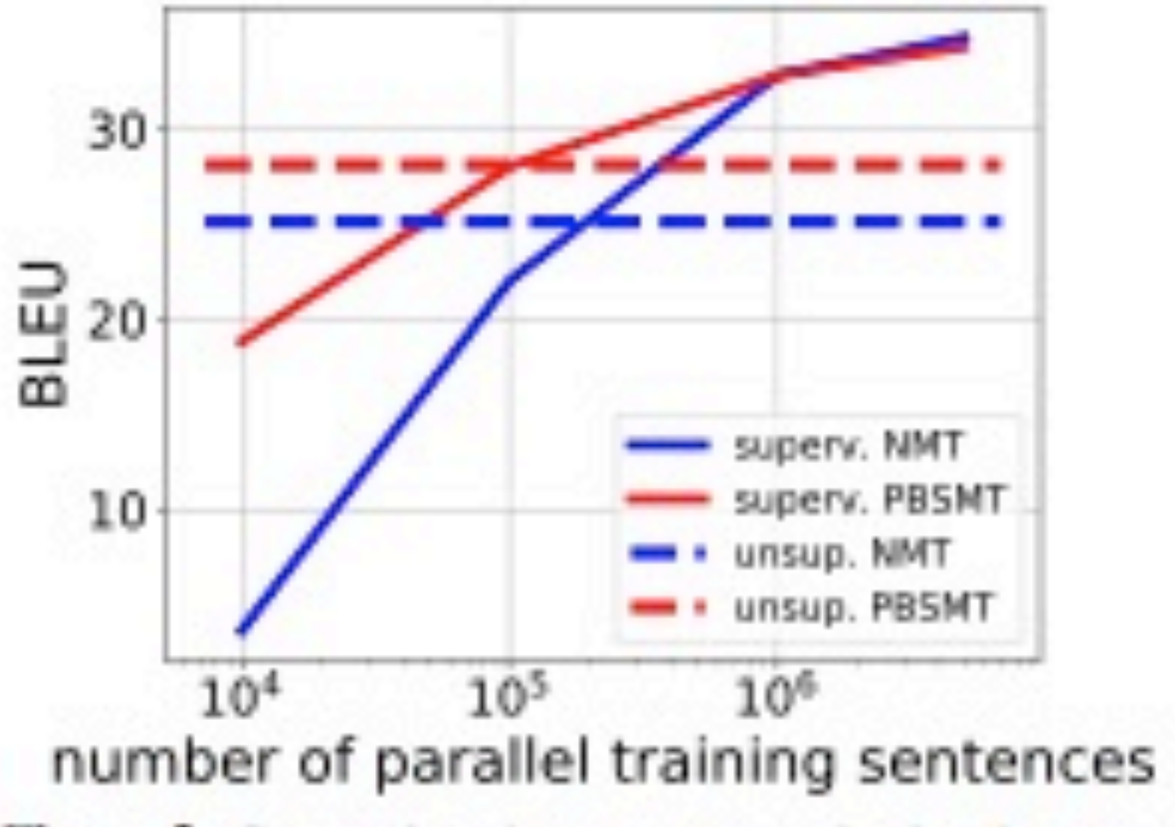
(C) Language modelling to ‘de-noise’ weak translations

(D) Iterative back-translation



Unsupervised NMT and PBMT

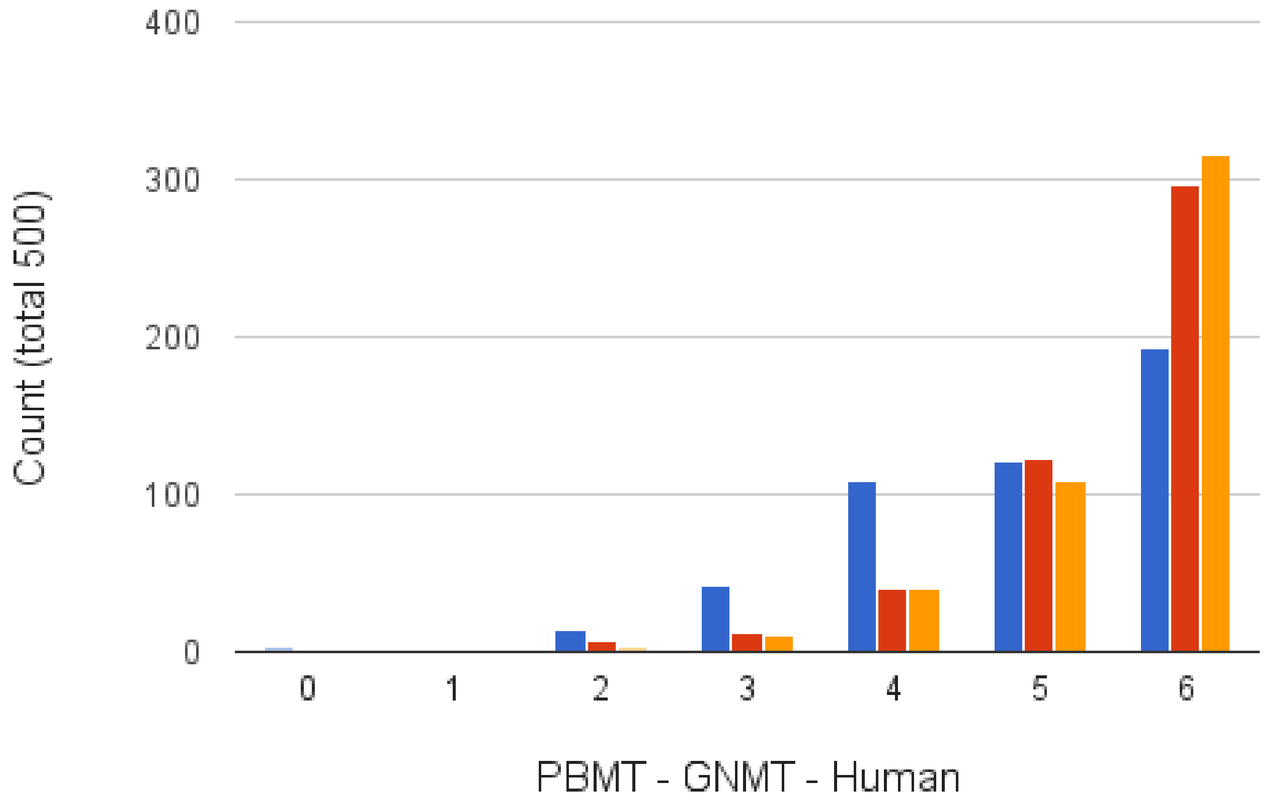
- Round trip de-noising auto-encoders
- Ignore discrete sampling steps in back-prop (i.e. no RL)
- Shared encoder parameters
- Phrase table built from frequent monolingual phrases scored by word embedding model
- PBMT not bad at all on low resource
- Hybrid PBMT+NMT is best



Has NMT really bridged the gap between MT and human translation?

Claims of ‘Human Parity’

- What problems are there with this evaluation?
- How could we improve it?



Challenges for NMT

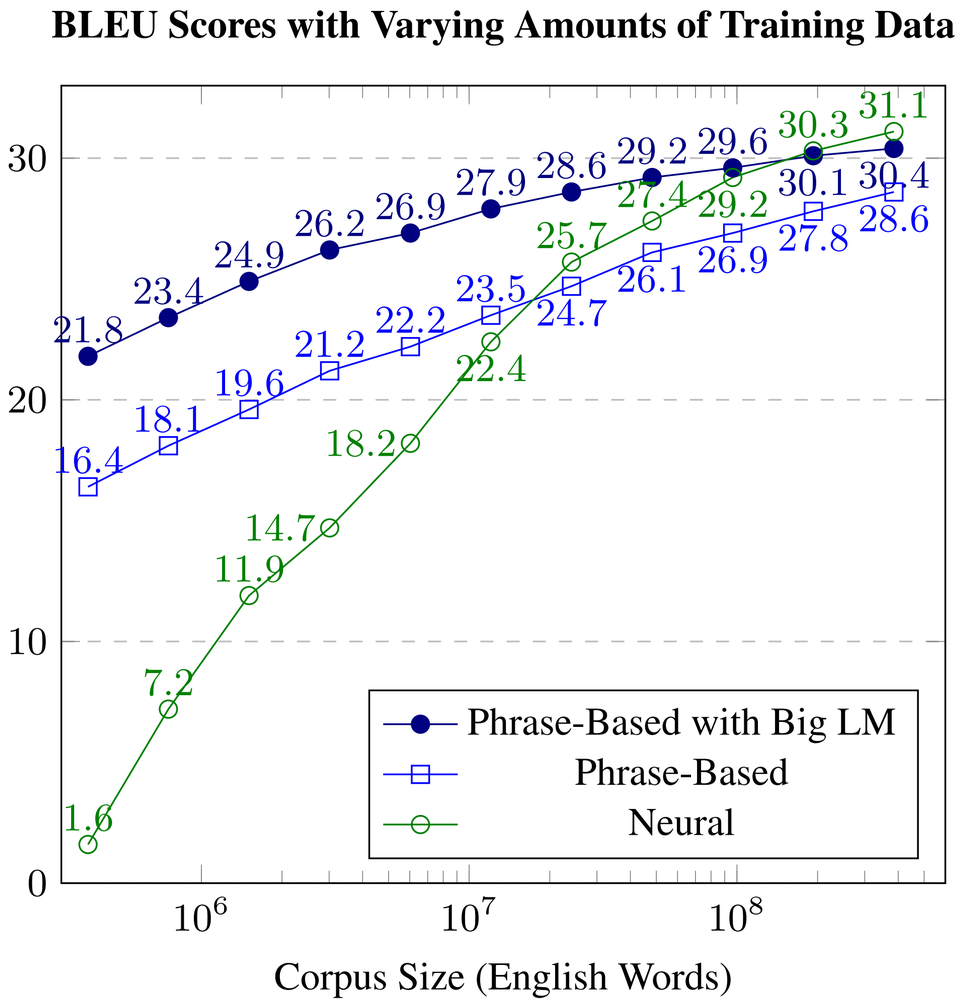
- Sensitive to domain mismatch
- Slower initial learning curve
- BPE helps on rare words but not on rare inflected words (verbs?)
- Worse on very long sentences
- Attention is not word alignment (?)
- Increasing beam doesn’t help (short hypotheses end up winning too often even with length normalization)

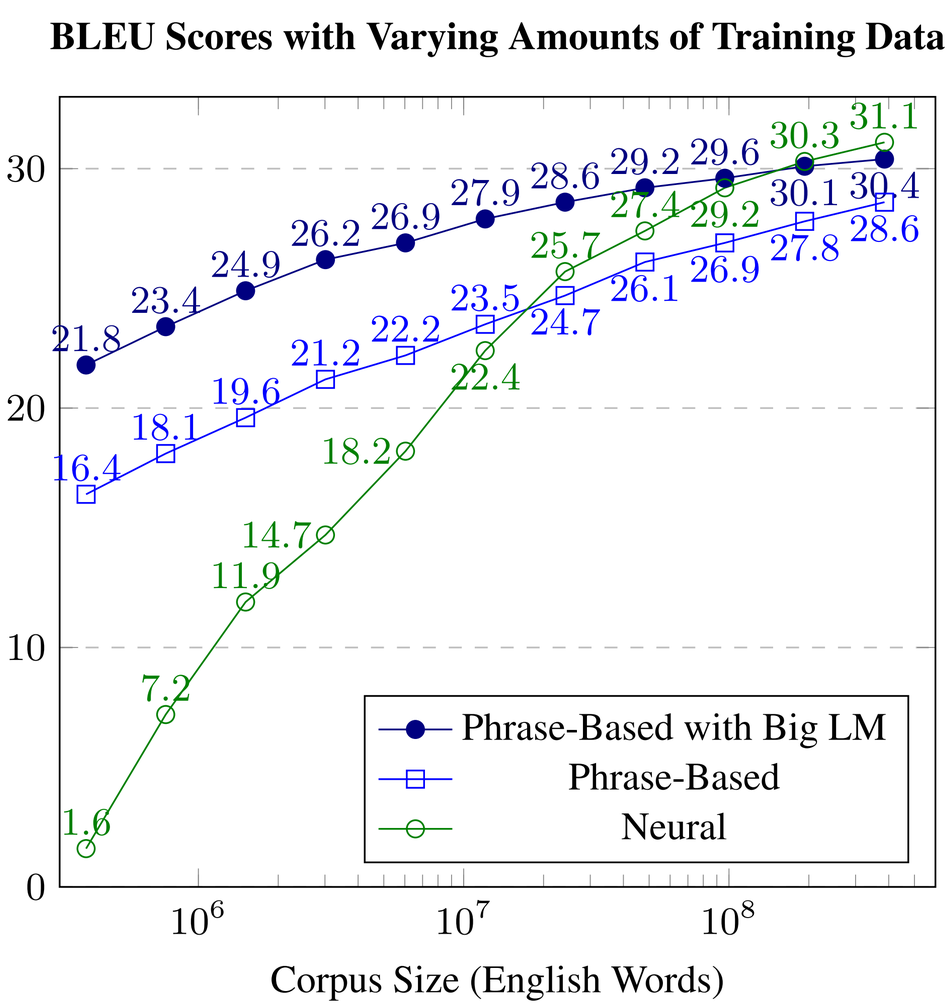
Challenges for NMT

System ↓	Law		Medical		IT		Koran		Subtitles	
All Data	30.5	32.8	45.1	42.2	35.3	44.7	17.9	17.9	26.4	20.8
Law	31.1	34.4	12.1	18.2	3.5	6.9	1.3	2.2	2.8	6.0
Medical	3.9	10.2	39.4	43.5	2.0	8.5	0.6	2.0	1.4	5.8
IT	1.9	3.7	6.5	5.3	42.1	39.8	1.8	1.6	3.9	4.7
Koran	0.4	1.8	0.0	2.1	0.0	2.3	15.9	18.8	1.0	5.5
Subtitles	7.0	9.9	9.3	17.8	9.2	13.6	9.0	8.4	25.9	22.1

Figure 1: Quality of systems (BLEU), when trained on one domain (rows) and tested on another domain (columns). Comparably, NMT systems (left bars) show more degraded performance out of domain.

Challenges for NMT





Challenges for NMT

- NMT translates words seen once worse than OOVs
- PBMT slightly worse performance on rare words

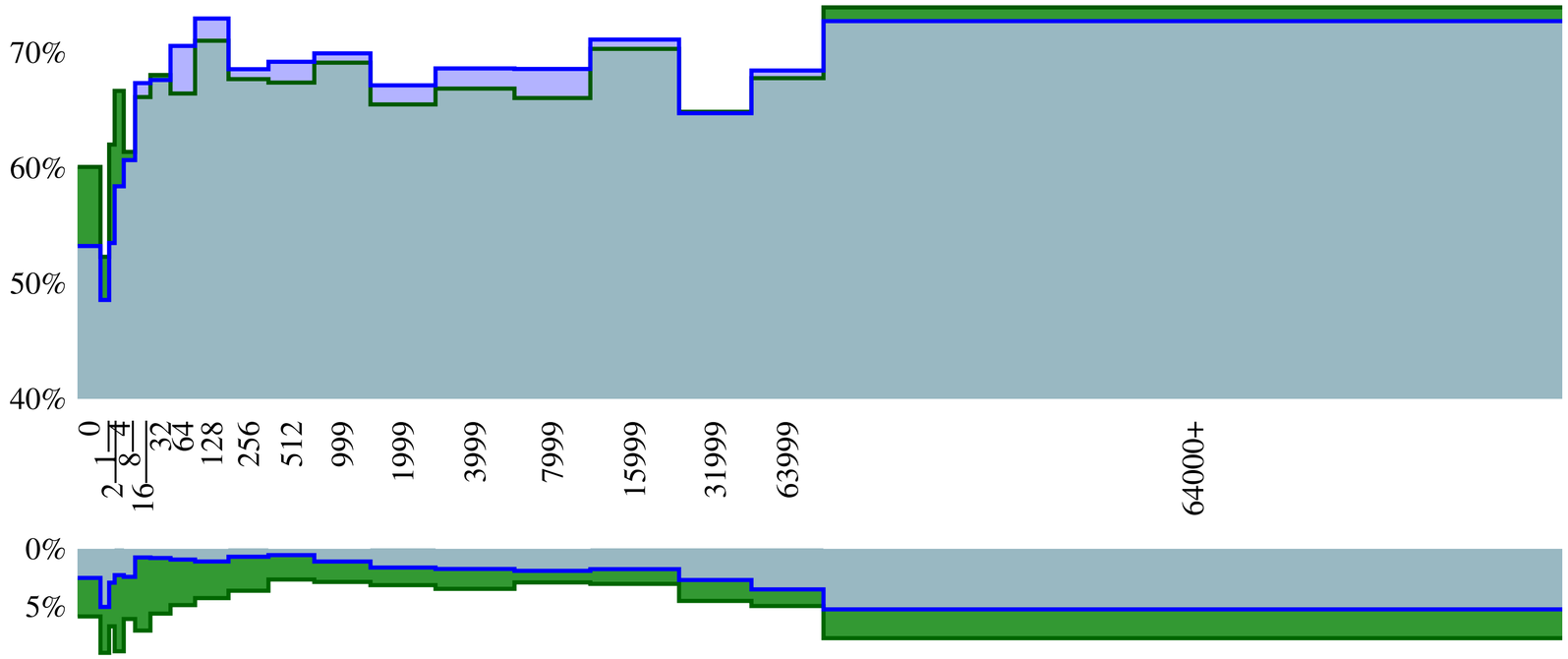
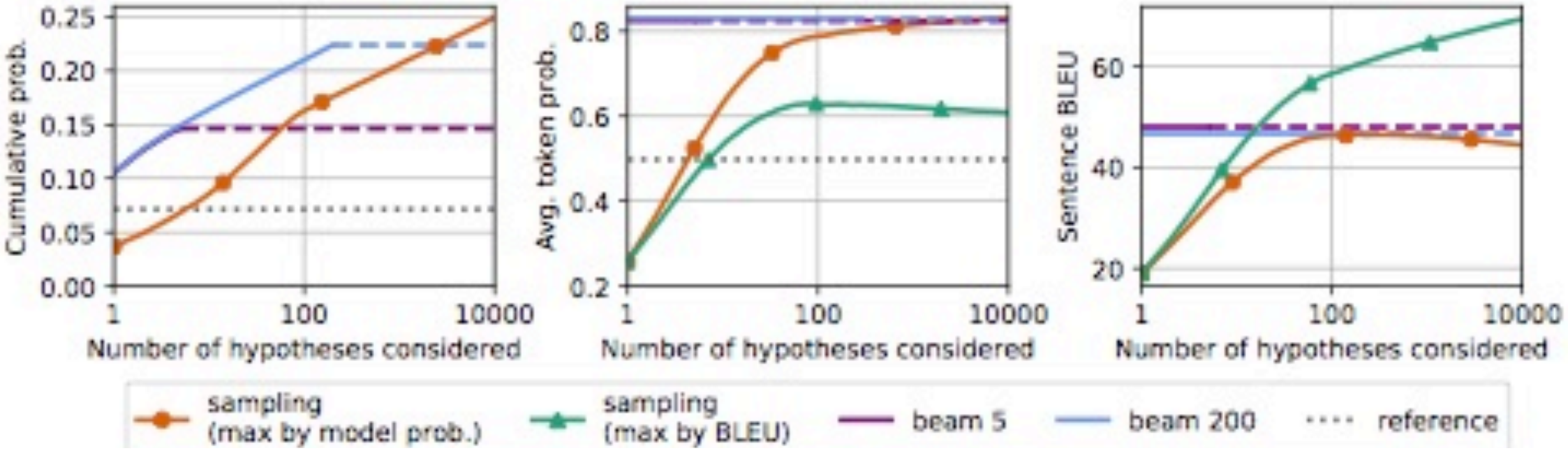


Figure 5: Precision of translation and deletion rates by source words type. SMT (light blue) and NMT

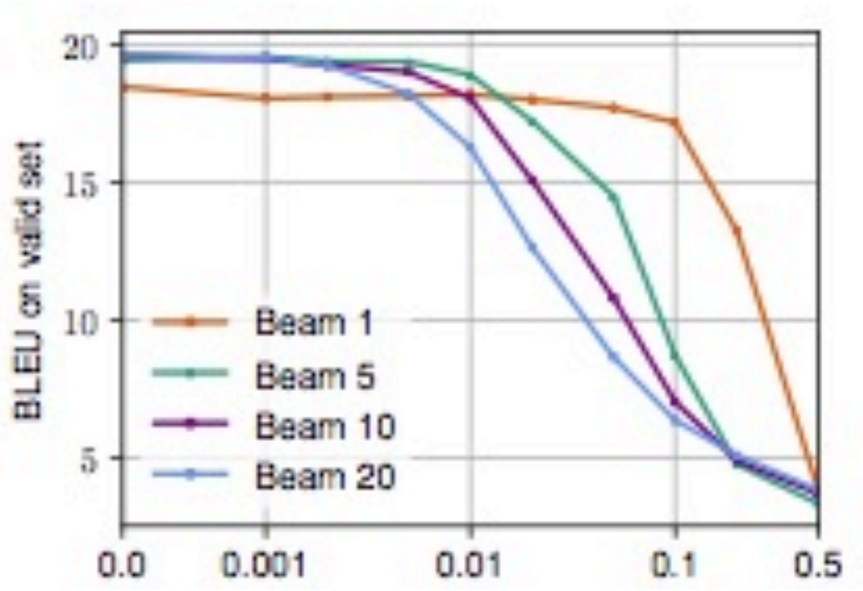
Analyzing Uncertainty

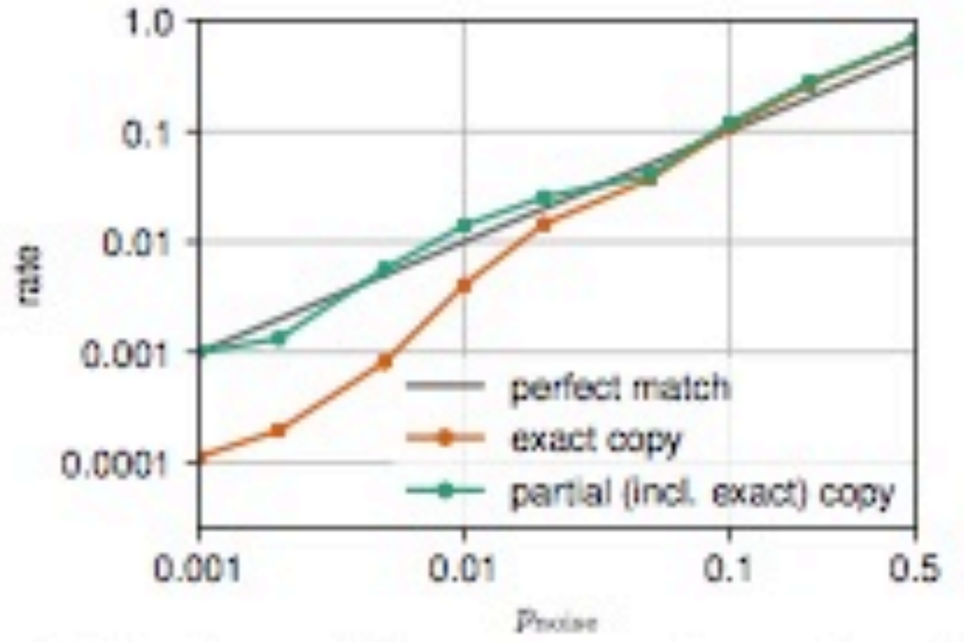
- Beam is efficient but only covers 25% of probabilty
- BLEU imperfectly correlated with model probability
- Copies overrepresented in the beam – hence large beams are bad



Analyzing Uncertainty

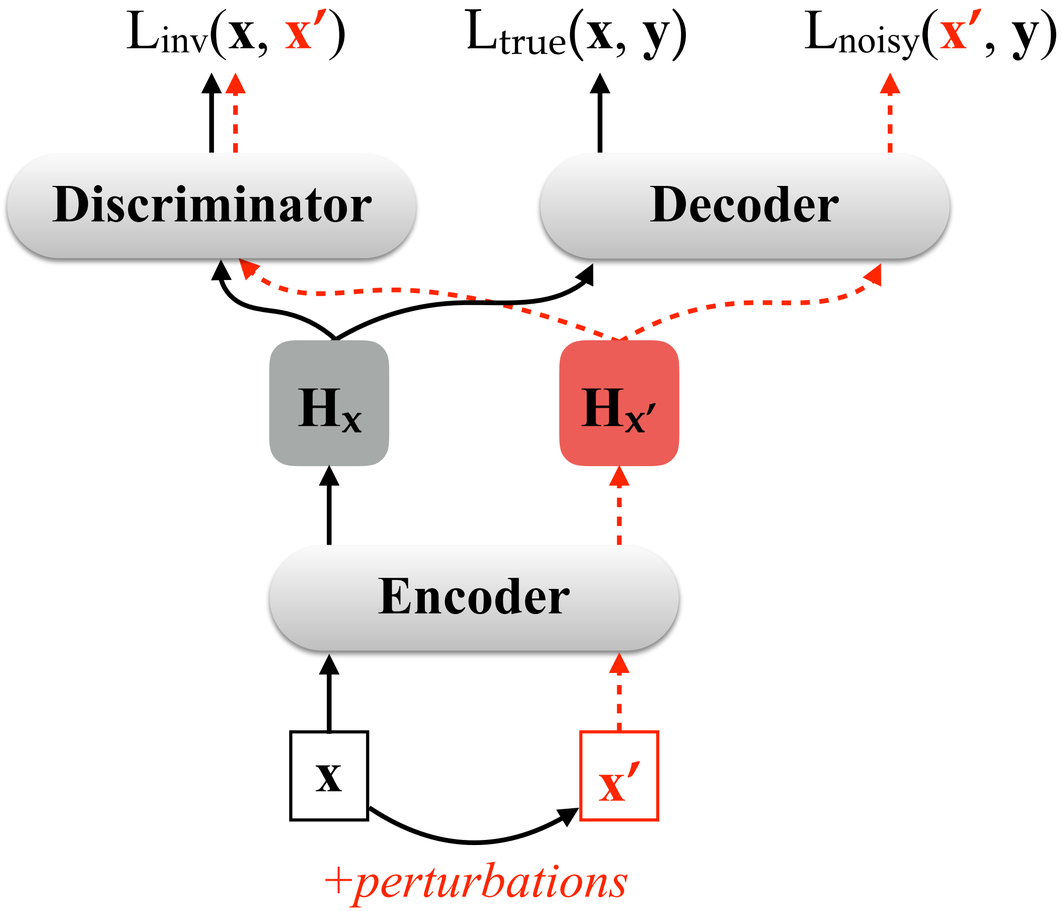
- Beam explores only high probability tokens
- Noise in training data has more impact on larger beams
- Rare words not matched by beam search
- Control: add copies as noise
- Conclusion: model is too smooth: it assigns mass to unobserved partial copies



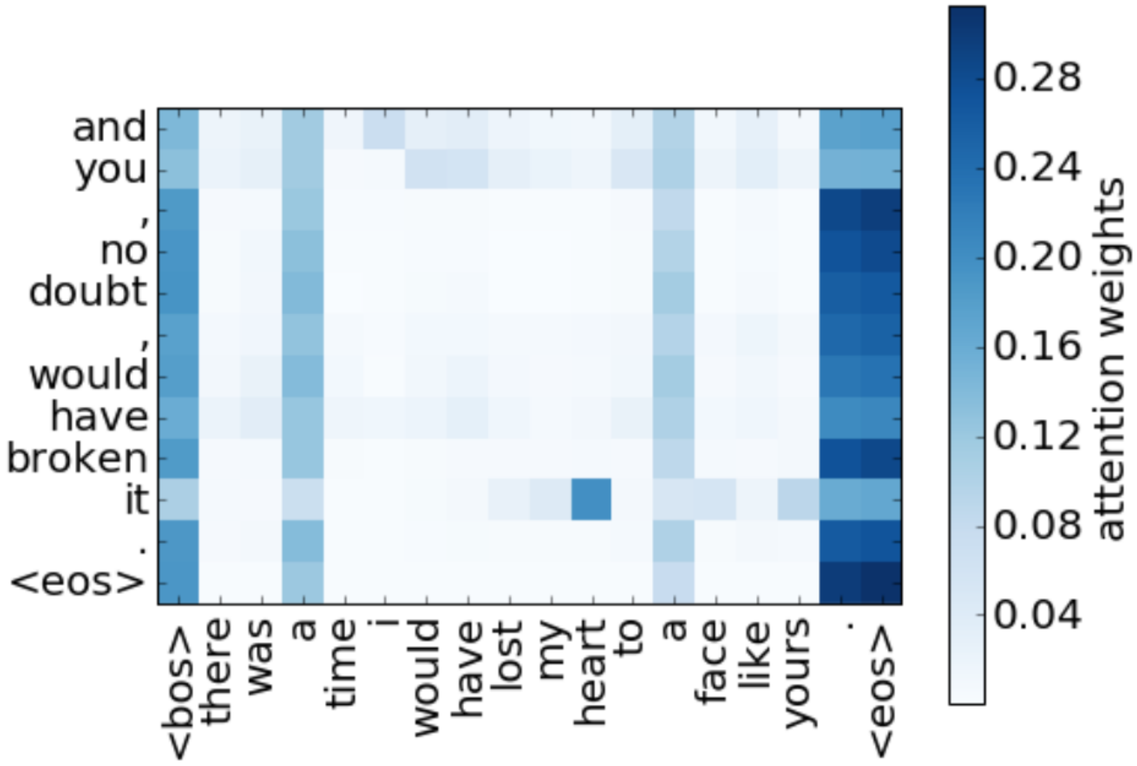


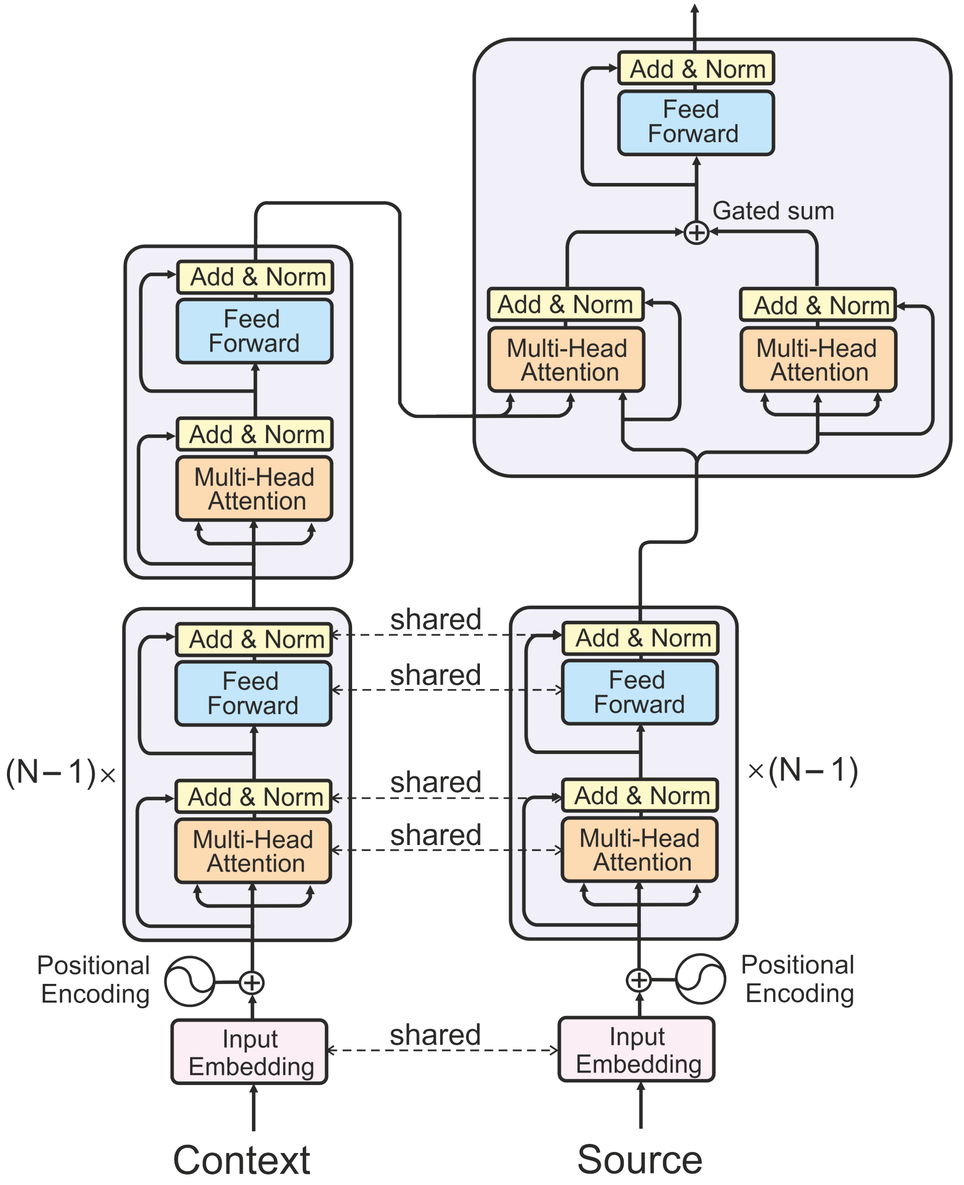
Perturbation-Invariant Encoder

- Crazy translations are common
- NMT is sensitive to noise in input
- Adversarial stability training:
  train noise-invariant encoder and decoder
- Discriminator tries to distinguish noised from real data
- Noise types: lexical (words) vs. feature (embeddings)
- Feature noise works best on clean test
- Ablation study



Context Aware





Context Aware

- Choice of baselines
- Interpretable architecture

model	BLEU
baseline	29.46
concatenation (previous sentence)	29.53
context encoder (previous sentence)	**30.14**
context encoder (next sentence)	29.31
context encoder (random context)	29.69

Table 1: Automatic evaluation: BLEU. Significant differences at $p < 0.01$ are in bold.

pronoun	agreement (in %)			
	random	first	last	attention
it	40	36	52	58
you	42	63	29	67
I	39	56	35	62

Table 7: Agreement with CoreNLP for test sets of pronouns having a nominal antecedent in context sentence (%). Examples with $\geq$1 noun in context sentence.

	agreement (in %)
CoreNLP	77
attention	72
last noun	54

Table 8: Performance of CoreNLP and our model's attention mechanism compared to human assessment. Examples with $\geq$1 noun in context sentence.