курс «Глубокое обучение»

Машинный перевод с помощью НС (Neural Machine Translation)

Александр Дьяконов

20 апреля 2020 года

План

Нейросетевой перевод

Метрика качества в переводе и саммаризации:
BLEU, GLEU, ROUGE, METEOR, TER, SARI

Подходы к NMT: Seq2seq, Attention, The Transformer model
ConvS2S, GNMT, RNMT+

Языковая модель в NMT
Использование нескольких языков
Редкие пары языков
Проблемы переводчика

Машинный перевод (Machine Translation)

самая естественная задача для ИИ – «понимание»

Анализ + генерация

коммерческое использование: Google Translate 100 млрд слов / день

Раньше: на правилах, статистика
Сейчас: глубокое обучение

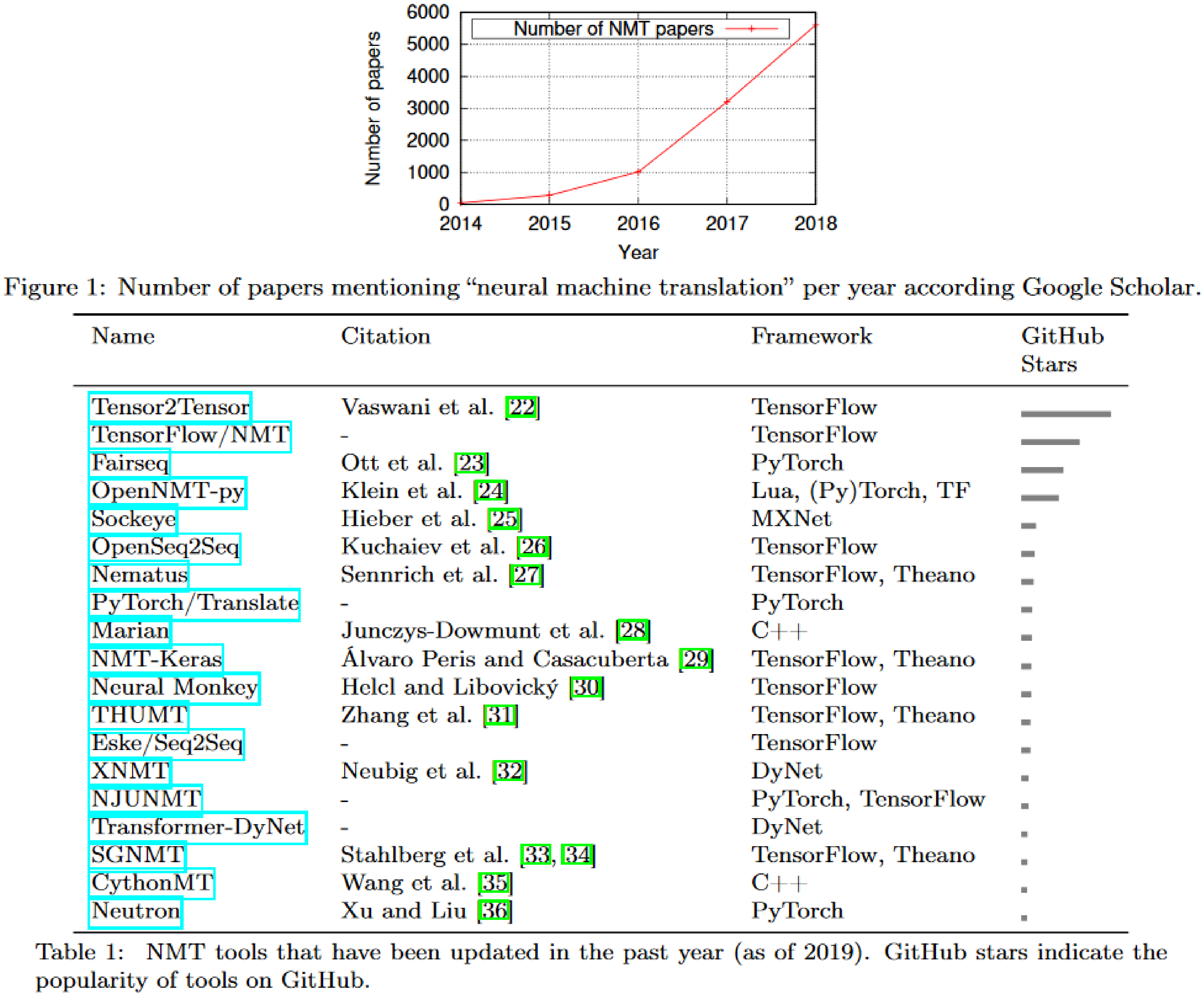

Number of NMT papers
Number of papers
6000
5000
4000
3000
2000
1000
0
2014
2015
2016
2017
2018
Year

Figure 1: Number of papers mentioning “neural machine translation” per year according Google Scholar.

Name	Citation	Framework	GitHub Stars
Tensor2Tensor	Vaswani et al. [22]	TensorFlow	▬▬▬▬▬▬▬▬▬▬▬▬
TensorFlow/NMT	-	TensorFlow	▬▬▬▬▬▬▬▬
Fairseq	Ott et al. [23]	PyTorch	▬▬▬▬▬▬
OpenNMT-py	Klein et al. [24]	Lua, (Py)Torch, TF	▬▬▬▬▬
Sockeye	Hieber et al. [25]	MXNet	▬▬
OpenSeq2Seq	Kuchaiev et al. [26]	TensorFlow	▬▬
Nematus	Sennrich et al. [27]	TensorFlow, Theano	▬▬
PyTorch/Translate	-	PyTorch	▬▬
Marian	Junczys-Dowmunt et al. [28]	C++	▬▬
NMT-Keras	Álvaro Peris and Casacuberta [29]	TensorFlow, Theano	▬
Neural Monkey	Helcl and Libovický [30]	TensorFlow	▬
THUMT	Zhang et al. [31]	TensorFlow, Theano	▬
Eske/Seq2Seq	-	TensorFlow	▬
XNMT	Neubig et al. [32]	DyNet	▪
NJUNMT	-	PyTorch, TensorFlow	▪
Transformer-DyNet	-	DyNet	▪
SGNMT	Stahlberg et al. [33, 34]	TensorFlow, Theano	▪
CythonMT	Wang et al. [35]	C++	▪
Neutron	Xu and Liu [36]	PyTorch	▪

Table 1: NMT tools that have been updated in the past year (as of 2019). GitHub stars indicate the popularity of tools on GitHub.

Особенности нейросетевого перевода

+ лучше качество

быстрее, используют контекст

+ одна архитектура для всех пар языков

не надо вручную генерировать признаки

– плохо интерпретируема

сложно отлаживать, влиять на перевод

текущие вызовы

Out-of-vocabulary

Другая специфика (Domain mismatch)

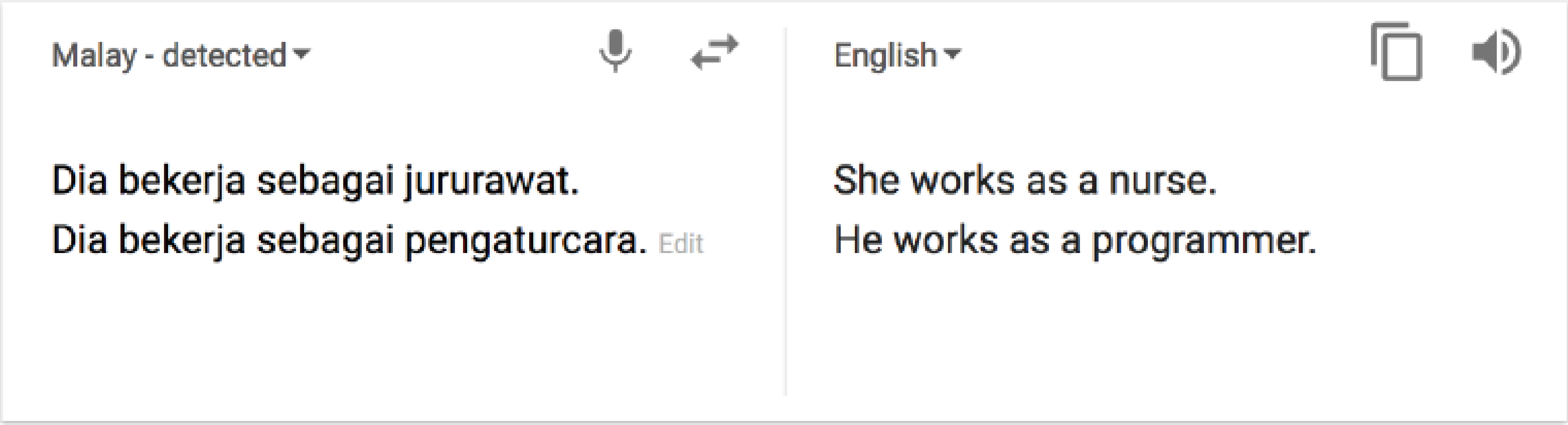
Длинный контекст

Редкие языковые пары

идиомы

статистические смещения

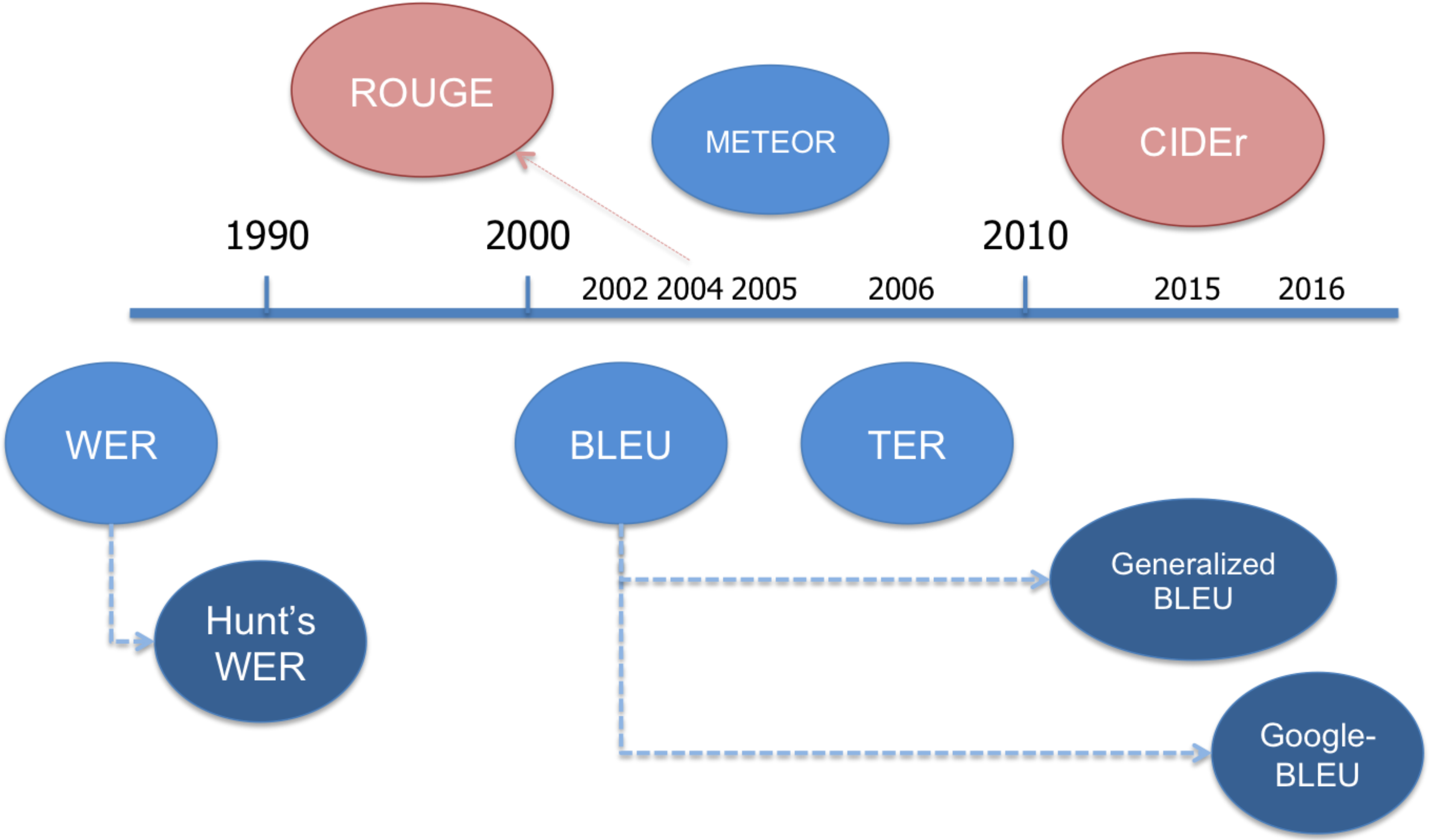
Статистические смещения



Dia bekerja sebagai jururawat.
Dia bekerja sebagai pengaturcara.

She works as a nurse.
He works as a programmer.

Метрика качества в переводе и саммаризации



https://github.com/gcunhase/NLPMetrics

Стандартная метрика качества: BLEU (Bilingual Evaluation Understudy)

число n-грамм ответа, которые есть в правильном ответе – N_n

BLEU = $\sqrt[4]{N_1 N_2 N_3 N_4} \cdot \underbrace{\min(1, |a| / |y|)}_{\text{BP}}$

BP – Brevity penalty – штраф за краткость

Пример

y **= «The best thing is my»**
a **= «My best thing»**

$N_1 = 3$ **(my, best, thing)**
$N_2 = 1$ **(best thing)**
$N_3 = N_4 = 0$

$$\sqrt[4]{3 \cdot 1 \cdot 0 \cdot 0} \cdot \min(1, 3/5) = 0$$

Хороший перевод может иметь маленькую BLEU-оценку

Papineni et al «BLEU: a Method for Automatic Evaluation of Machine Translation», 2002 // http://aclweb.org/anthology/P02-1040

GLEU (Google-BLEU)

для n-грамм n=1,2,3,4 берём минимум между точностью и полнотой

ROUGE: Recall-Oriented Understudy for Gisting Evaluation

обычно для сравнений саммари (в отличие от BLUE больше ориентирован на полноту)

ROUGE_n **= число общих n-грамм / число n-грамм в правильном ответе**

$$\mathrm{ROUGE}_L = \frac{(1+\beta^2)R_{\mathrm{LCS}}P_{\mathrm{LCS}}}{R_{\mathrm{LCS}}+\beta^2 P_{\mathrm{LCS}}}$$

$$R_{\mathrm{LCS}} = \frac{\mathrm{LCS}}{|y|},\ P_{\mathrm{LCS}} = \frac{\mathrm{LCS}}{|a|}$$

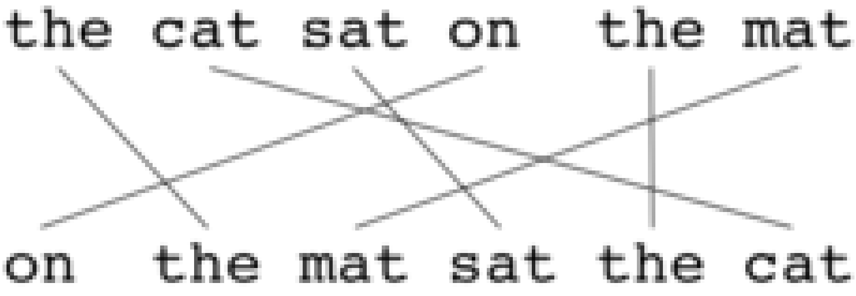
LCS = longest common subsequence

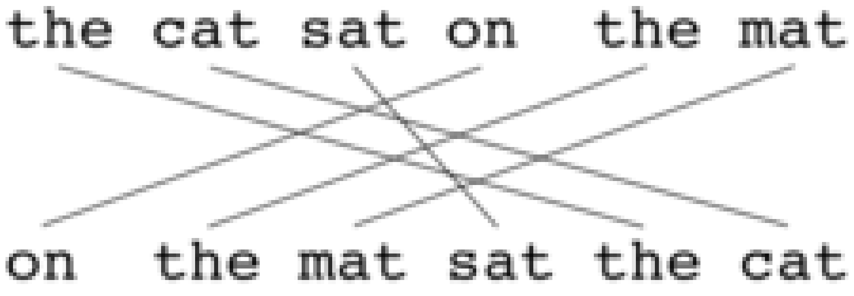
ROUGE_W **– взвешанная LSC-статистика**

https://github.com/pltrdy/rouge

METEOR: Metric for Evaluation of Translation with Explicit ORdering

сначала находят соответствия слов





каждое слово – чанк

on the mat, sat, the cat

+ формируем чанки – непрерывные последовательности, которые отображаются

$$\text{METEOR} = \frac{10RP}{R+9P}\left(1-0.5\left(\frac{\#\text{chanks}}{\#\text{owords}}\right)^3\right)$$

P = число общих слов / число слов в ответе

R = число общих слов / число слов в правильном ответе

chunks = число чанков

owords = число общих слов

https://www.cs.cmu.edu/~alavie/METEOR/pdf/Banerjee-Lavie-2005-METEOR.pdf

TER (Translation Edit Rate)

Считаем, сколько исправлений надо сделать, чтобы ответ стал поход на правильный (или один из правильных):

- вставка слова
- удаление слова
- замена слова
- перенос отрезка

TER = число исправлений / число слов в правильном ответе

Если правильных ответов несколько – в знаменателе берём среднее по ним

Matthew Snover and Bonnie Dorr «A Study of Translation Edit Rate with Targeted Human Annotation» // https://www.cs.umd.edu/~snover/pub/amta06/ter_amta.pdf

SARI: System output against references and against the input sentence

$$SARI = d_1 F_{add} + d_2 F_{keep} | + d_3 P_{del}$$

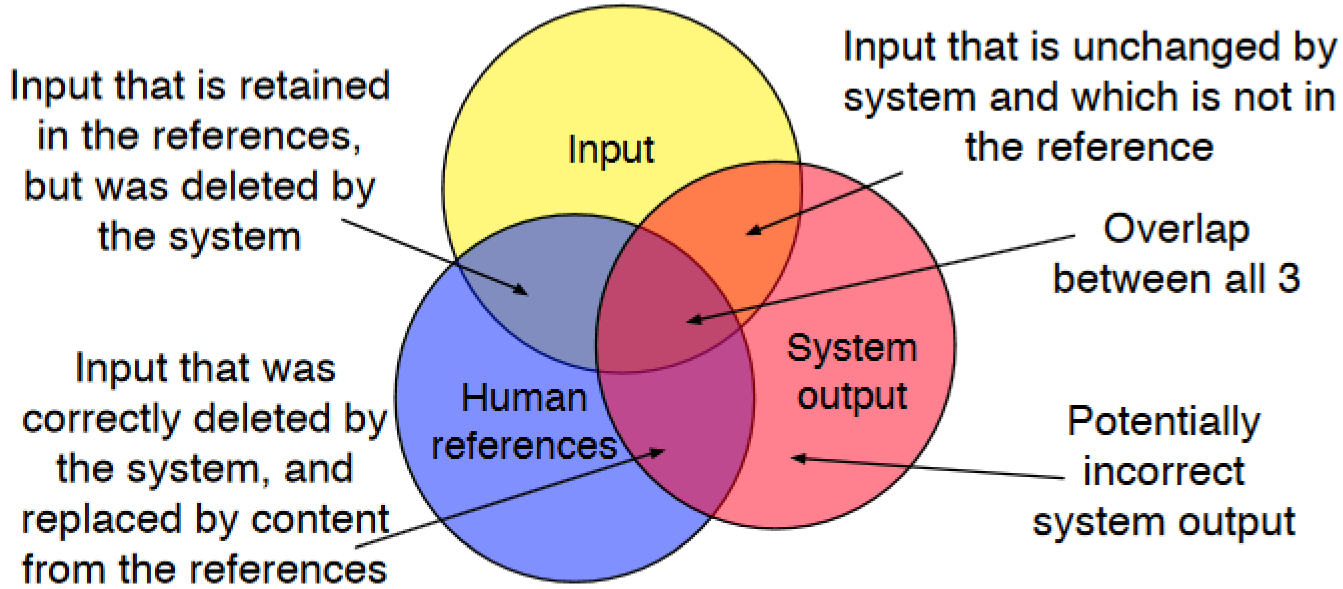

Input that is retained in the references, but was deleted by the system
Input
Input that is unchanged by system and which is not in the reference
Overlap between all 3
Input that was correctly deleted by the system, and replaced by content from the references
Human references
System output
Potentially incorrect system output

Figure 1: Metrics that evaluate the output of monolingual text-to-text generation systems can compare system output against references and against the input sentence, unlike in MT metrics which do not compare against the (foreign) input sentence. The different regions of this Venn diagram are treated differently with our SARI metric.

Больше для саммаризации

Учитываем добавления, удаления и сохранения слов

Wei Xu «Optimizing Statistical Machine Translation for Text Simplification» // https://cocoxu.github.io/publications/tacl2016-smt-simplification.pdf

Neural Machine Translation – уже почти изучили все этапы

seq2seq models

(Kalchbrenner and Blunsom, 2013; Sutskeveret al., 2014; Cho et al., 2014)

Recurrent Neural Networks (RNNs) interacting via a soft-attention mechanism

(Bahdanau et al., 2015)

Convolutional neural network based approaches

(LeCun and Bengio, 1998)

ByteNet (Kalchbrenner et al., 2016) 🟥

ConvS2S (Gehring et al., 2017)

Transformer model (Vaswaniet al., 2017)

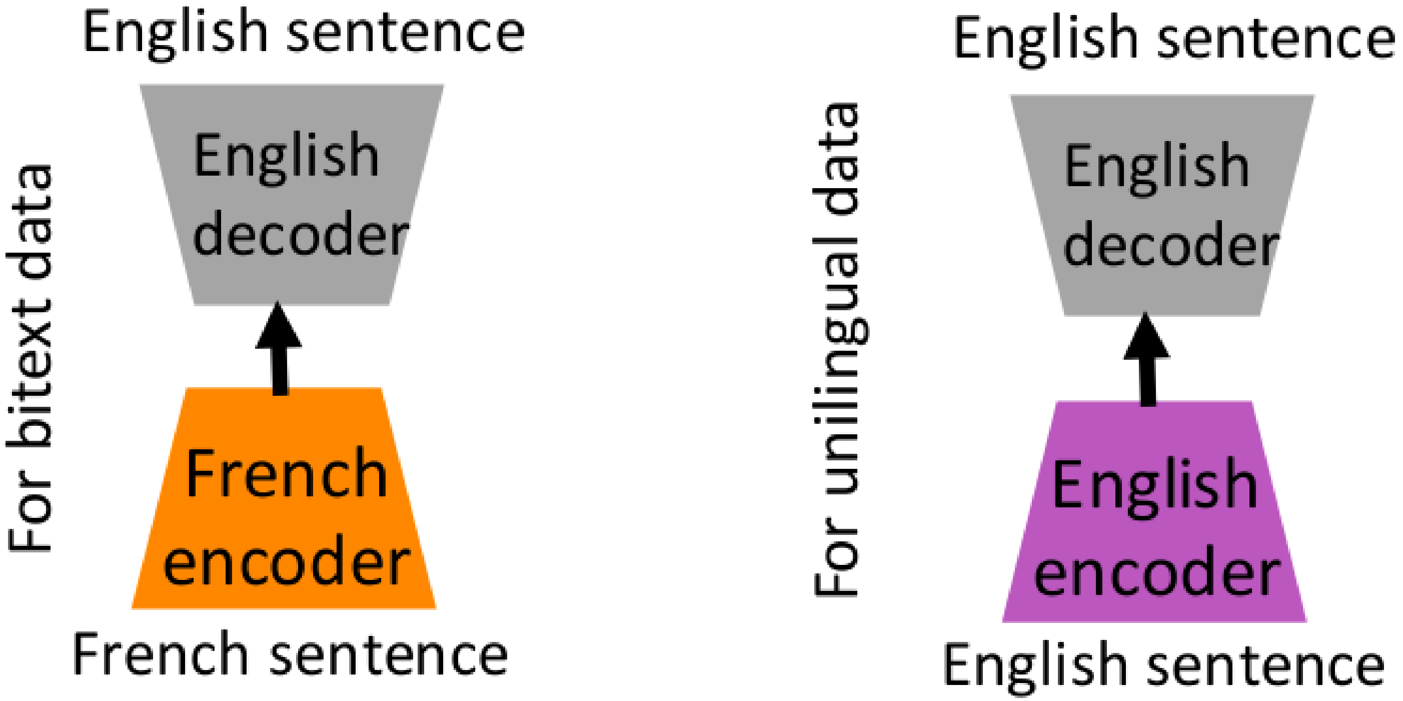
Self-attention mechanism (Parikh et al., 2016)

https://arxiv.org/pdf/1606.01933.pdf

Seq2seq

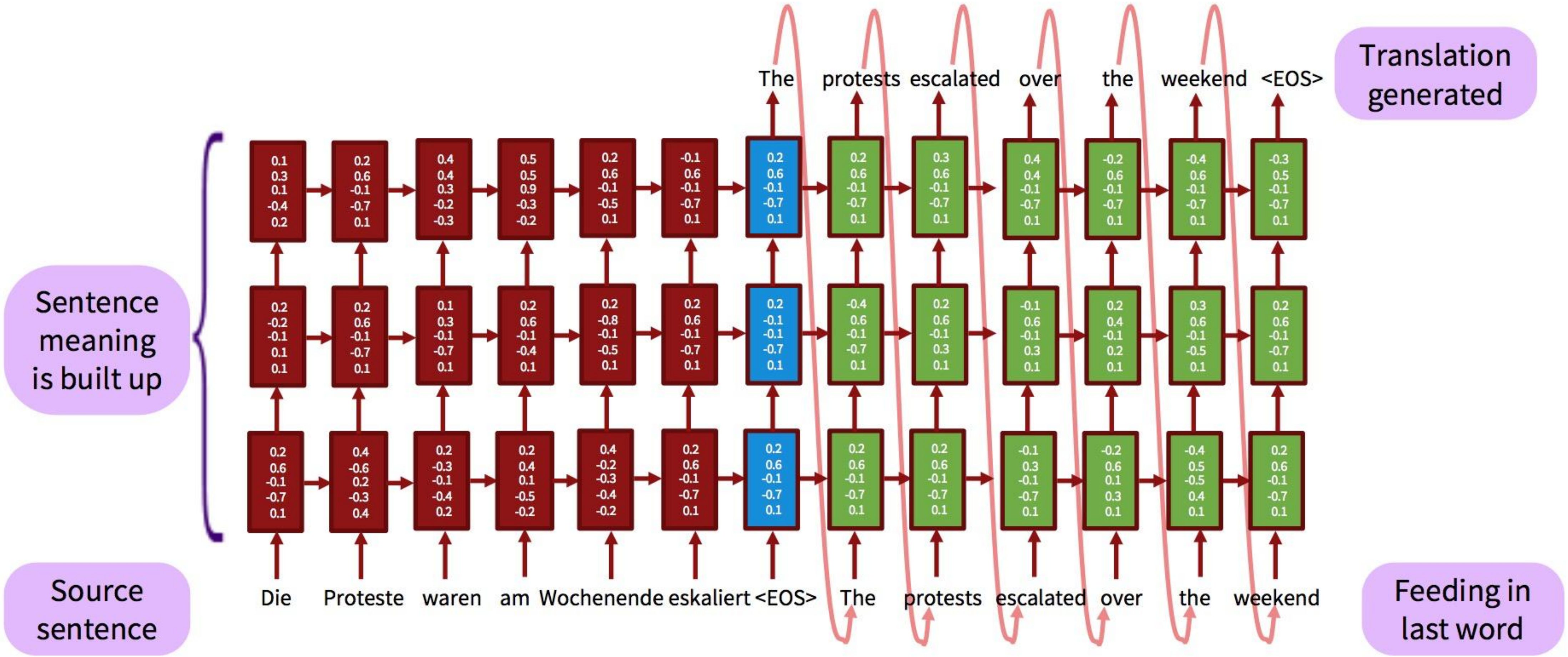
кодировщик – декодировщик

вход – векторные представления слов



[Kalchbrenner and Blunsom, 2013]

Seq2seq


Translation generated
The protests escalated over the weekend <EOS>
Sentence meaning is built up
Source sentence
Die Proteste waren am Wochenende eskaliert <EOS> The protests escalated over the weekend
Feeding in last word

пробовали разные виды RNN

Cho, Kyunghyun, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation // arXiv preprint arXiv:1406.1078 (2014).

Fully Character-Level Neural Machine Translation without Explicit Segmentation

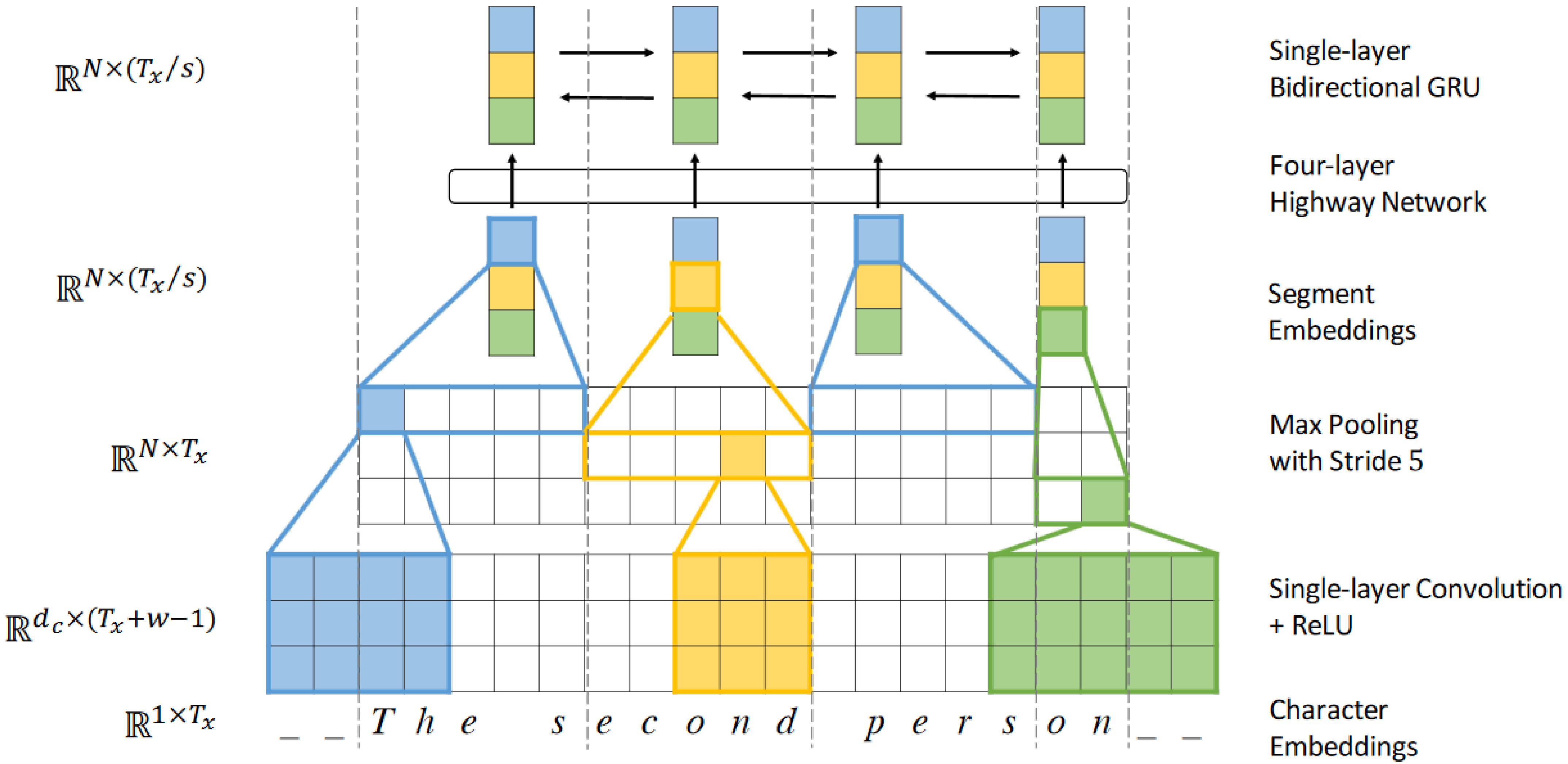

$\mathbb{R}^{N\times(T_x/s)}$
Single-layer Bidirectional GRU
Four-layer Highway Network
$\mathbb{R}^{N\times(T_x/s)}$
Segment Embeddings
$\mathbb{R}^{N\times T_x}$
Max Pooling with Stride 5
$\mathbb{R}^{d_c\times(T_x+w-1)}$
Single-layer Convolution + ReLU
$\mathbb{R}^{1\times T_x}$
_ _ T h e s e c o n d p e r s o n _ _
Character Embeddings

Figure 1: Encoder architecture schematics. Underscore denotes padding. A dotted vertical line delimits each segment. The stride of pooling s is 5 in the diagram.

Fully Character-Level Neural Machine Translation without Explicit Segmentation

последовательность символов → последовательность символов

без сегментации (нет формализации границ слов)

свёрточная сеть в кодировщике

можно один кодировщик для разных пар задач

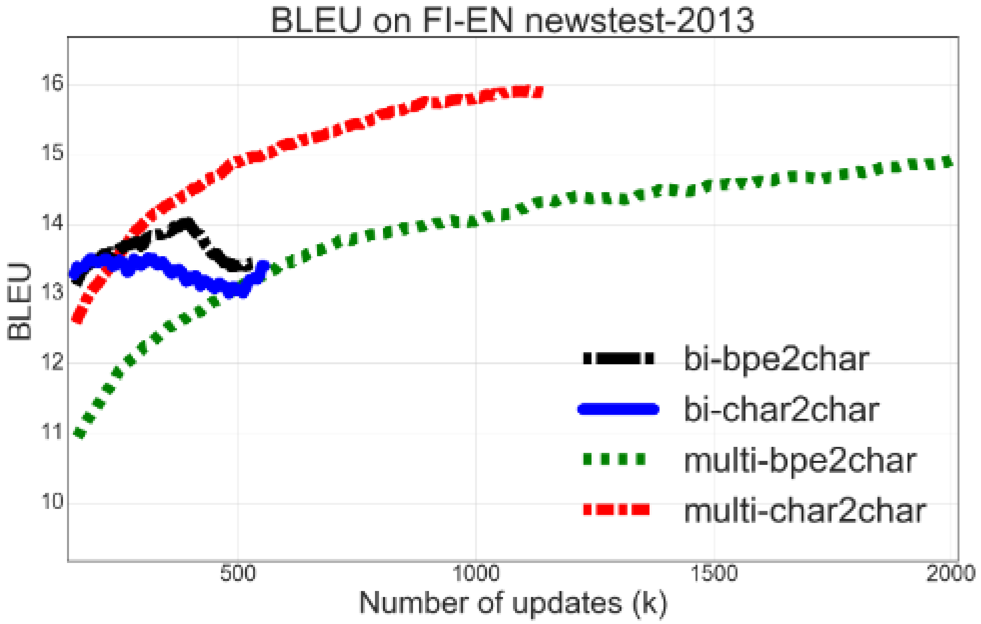
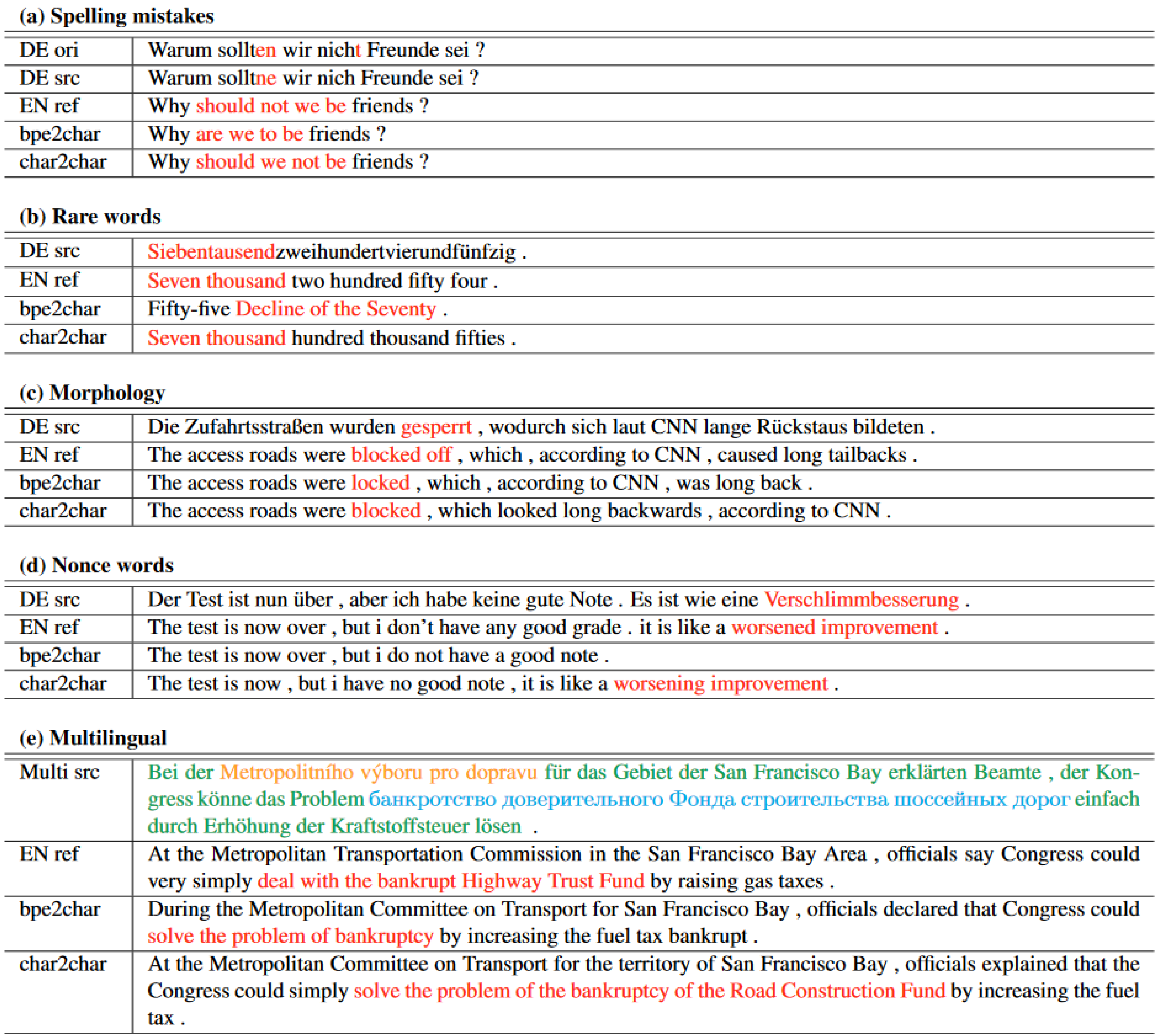


Figure 2: Multilingual models overfit less than bilingual models on low-resource language pairs.

Jason Lee, Kyunghyun Cho, Thomas Hoffmann «Fully Character-Level Neural Machine Translationwithout Explicit Segmentation». // https://arxiv.org/pdf/1610.03017.pdf



(a) Spelling mistakes

DE ori	Warum sollten wir nicht Freunde sei ?
DE src	Warum solltne wir nich Freunde sei ?
EN ref	Why should not we be friends ?
bpe2char	Why are we to be friends ?
char2char	Why should we not be friends ?

(b) Rare words

DE src	Siebentausendzweihundertvierundfünfzig .
EN ref	Seven thousand two hundred fifty four .
bpe2char	Fifty-five Decline of the Seventy .
char2char	Seven thousand hundred thousand fifties .

(c) Morphology

DE src	Die Zufahrtsstraßen wurden gesperrt , wodurch sich laut CNN lange Rückstaus bildeten .
EN ref	The access roads were blocked off , which , according to CNN , caused long tailbacks .
bpe2char	The access roads were locked , which , according to CNN , was long back .
char2char	The access roads were blocked , which looked long backwards , according to CNN .

(d) Nonce words

DE src	Der Test ist nun über , aber ich habe keine gute Note . Es ist wie eine Verschlimmbesserung .
EN ref	The test is now over , but i don't have any good grade . it is like a worsened improvement .
bpe2char	The test is now over , but i do not have a good note .
char2char	The test is now , but i have no good note , it is like a worsening improvement .

(e) Multilingual

Multi src	Bei der Metropolitního výboru pro dopravu für das Gebiet der San Francisco Bay erklärten Beamte , der Kongress könne das Problem банкротство доверительного Фонда строительства шоссейных дорог einfach durch Erhöhung der Kraftstoffsteuer lösen .
EN ref	At the Metropolitan Transportation Commission in the San Francisco Bay Area , officials say Congress could very simply deal with the bankrupt Highway Trust Fund by raising gas taxes .
bpe2char	During the Metropolitan Committee on Transport for San Francisco Bay , officials declared that Congress could solve the problem of bankruptcy by increasing the fuel tax bankrupt .
char2char	At the Metropolitan Committee on Transport for the territory of San Francisco Bay , officials explained that the Congress could simply solve the problem of the bankruptcy of the Road Construction Fund by increasing the fuel tax .

Stronger character results with depth in LSTM seq2seq model

character level

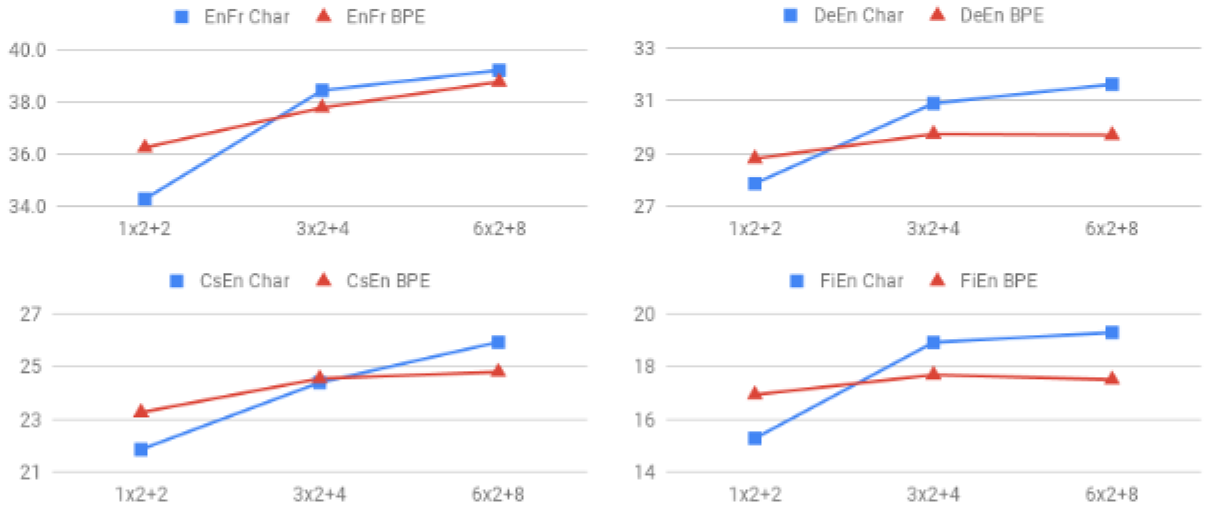


Figure 1: Test BLEU for character and BPE translation as architectures scale from 1 BiLSTM encoder layer and 2 LSTM decoder layers (1×2+2) to our standard 6×2+8. The y-axis spans 6 BLEU points for each language pair.

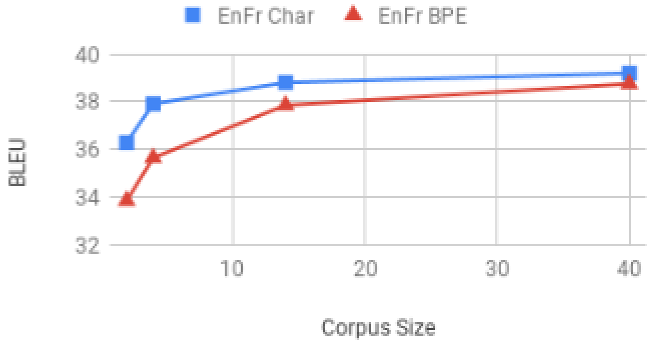


Figure 2: BLEU versus training corpus size in millions of sentence pairs, for the EnFr language-pair.

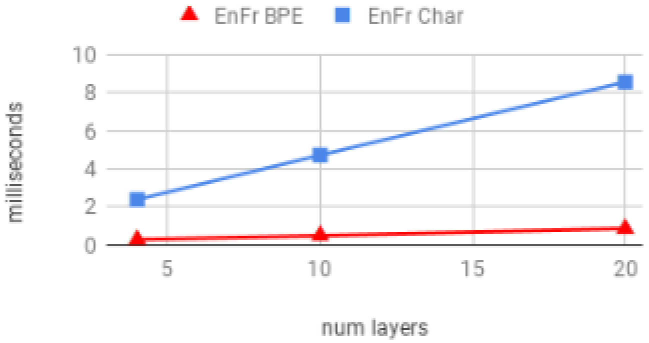
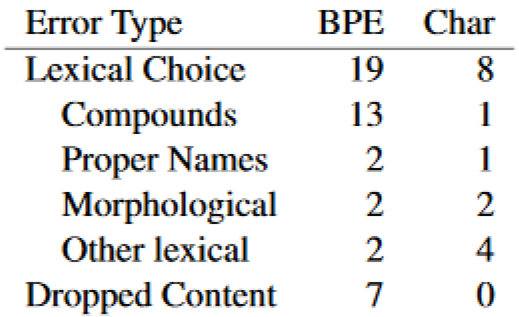


Figure 3: Training time per sentence versus total number of layers (encoder plus decoder) in the model.



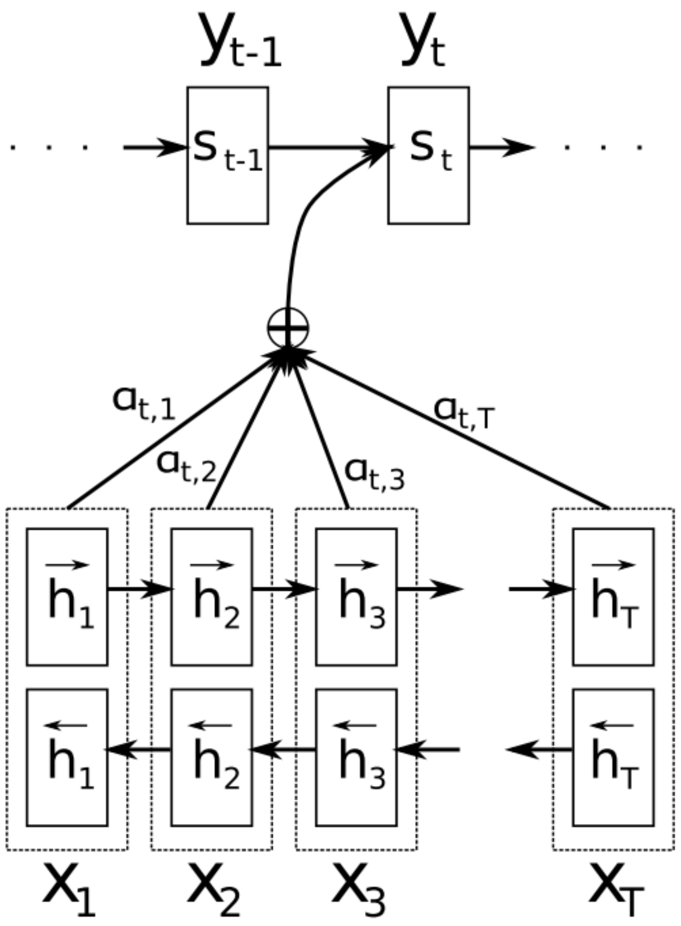
Error Type	BPE	Char
Lexical Choice	19	8
Compounds	13	1
Proper Names	2	1
Morphological	2	2
Other lexical	2	4
Dropped Content	7	0

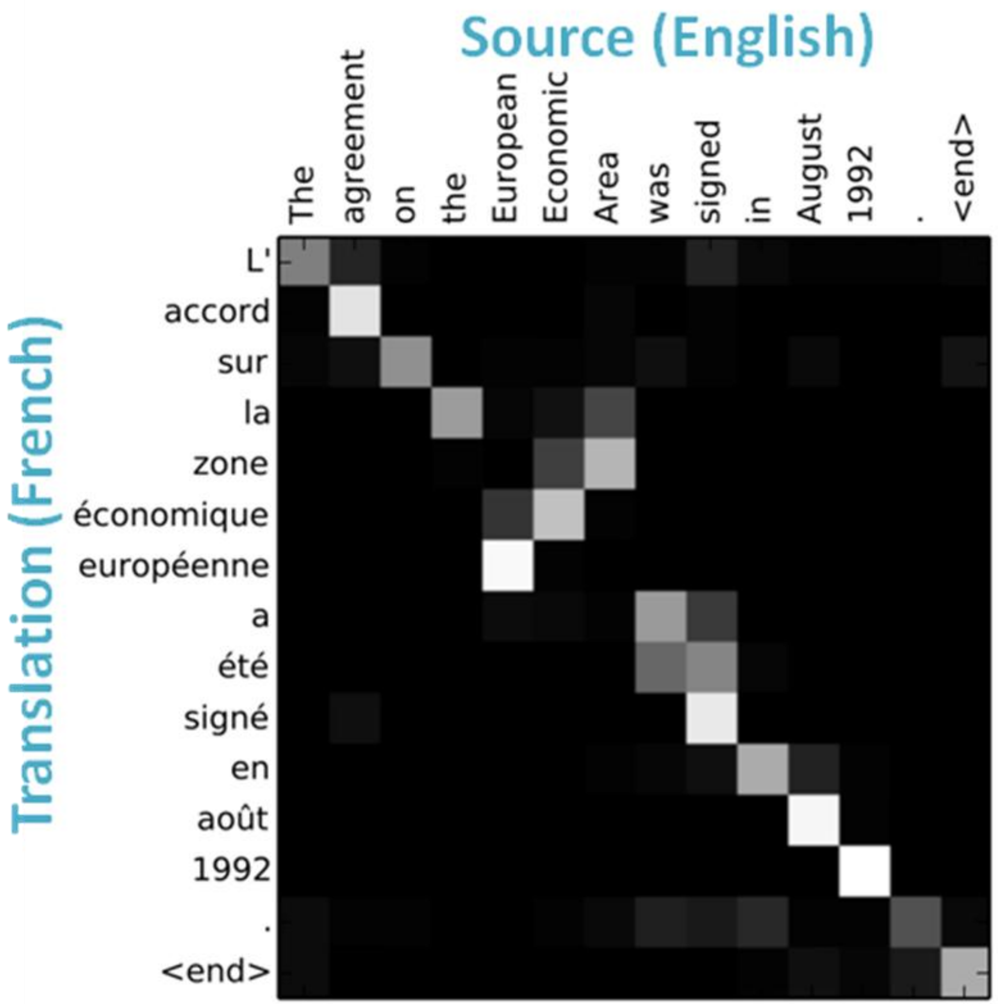
Table 4: Error counts out of 100 randomly sampled examples from the DeEn test set.

2018. Cherry, Foster, Bapna, Firat, Macherey, Google AI

https://www.aclweb.org/anthology/D18-1461.pdf

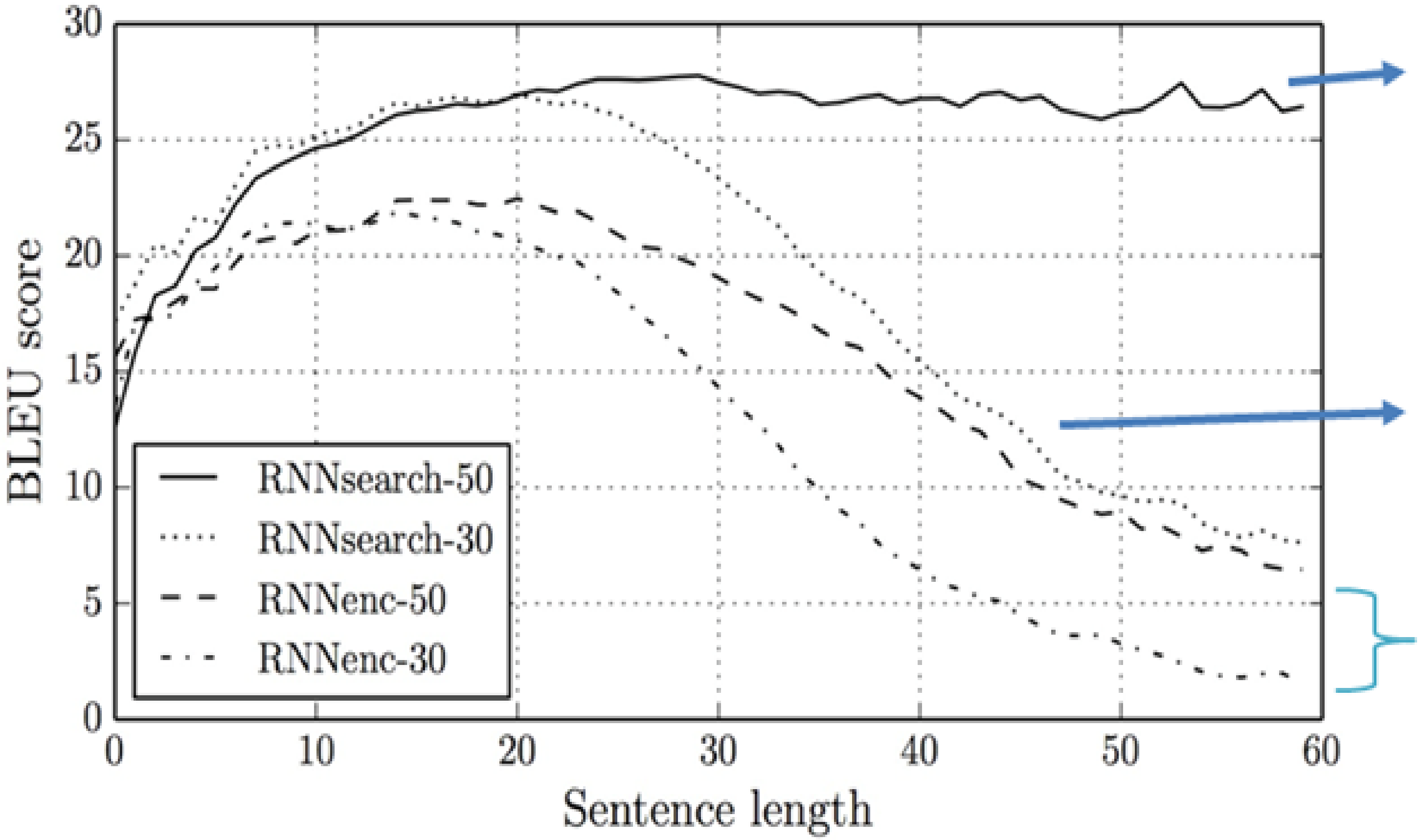
Attention





D. Bahdanau, K. Cho, Y. Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. ICLR 2015

Attention vs seq2seq



With attention

(trained with sentence length <= 50)

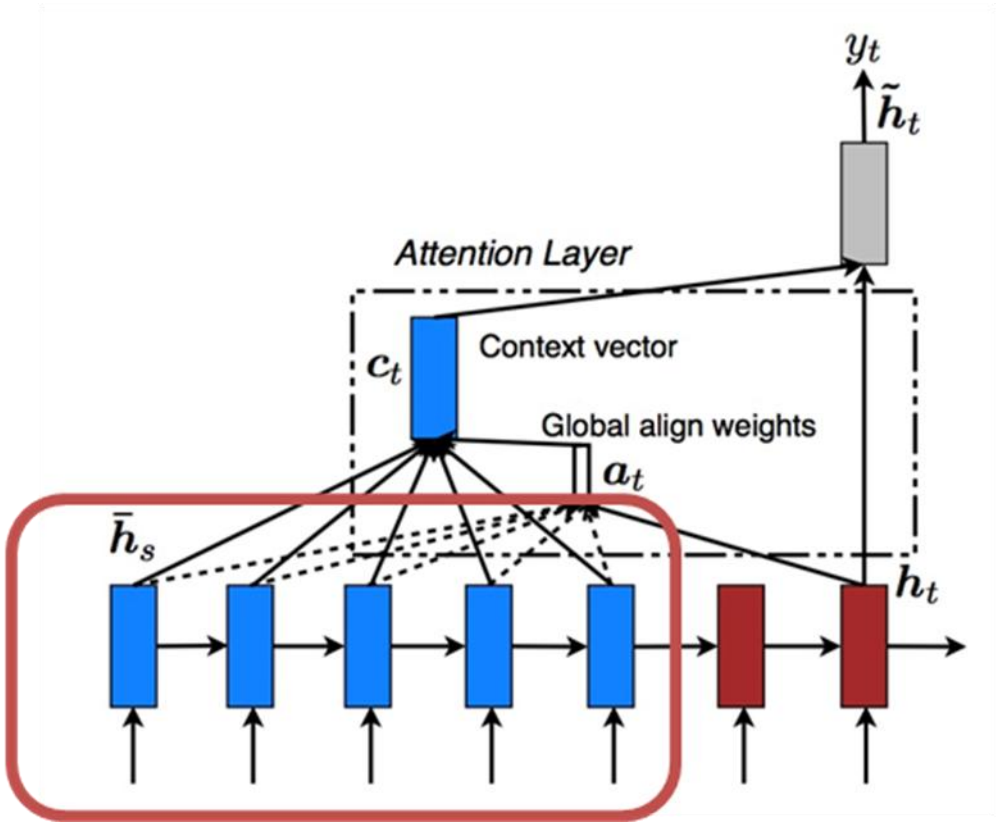
With attention

(trained with sentence length <= 30)

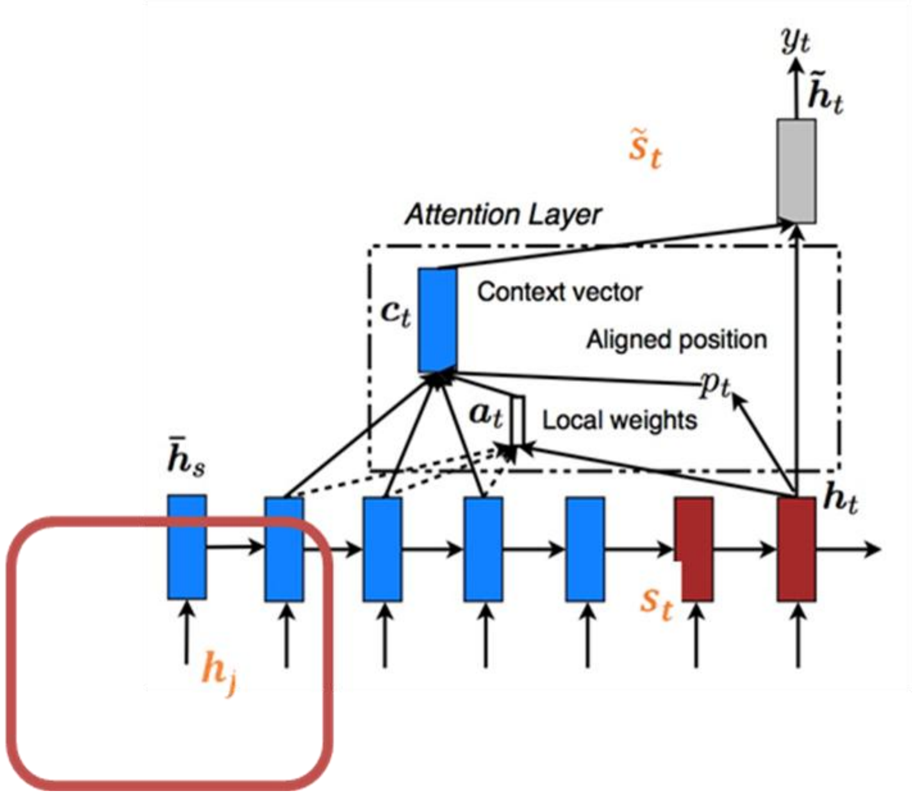
No attention

Attention+

Global Attention



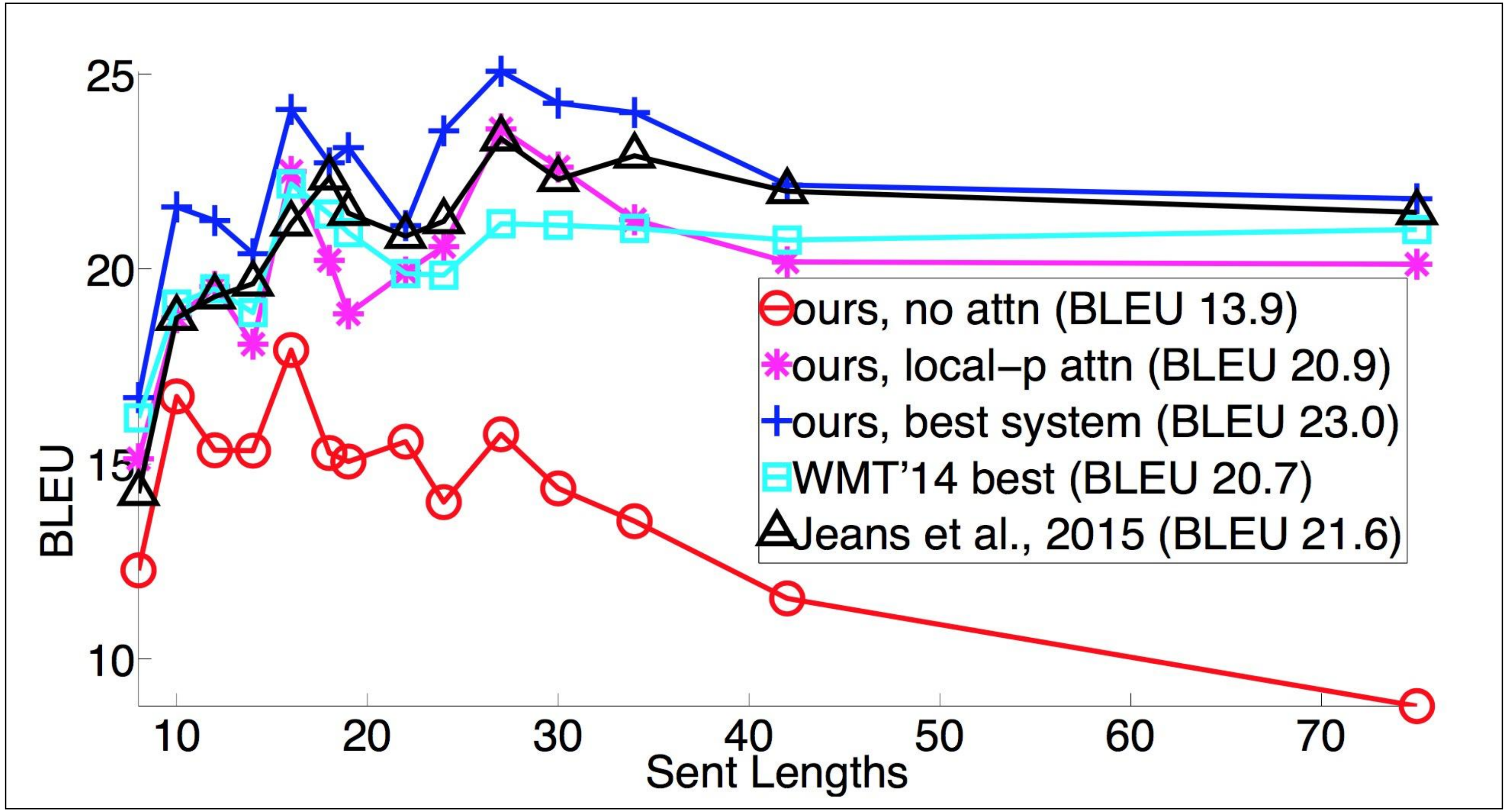
Local Attention



не фокусируемся сразу на всём (можно предсказывать на сколько смещаться)

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. EMNLP 2015

Attention+


25
20
15
10
BLEU
10
20
30
40
50
60
70
Sent Lengths
ours, no attn (BLEU 13.9)
ours, local–p attn (BLEU 20.9)
ours, best system (BLEU 23.0)
WMT'14 best (BLEU 20.7)
Jeans et al., 2015 (BLEU 21.6)

Convolutional Sequence to Sequence Learning (ConvS2S)

кодировщик и декодировщик – несколько свёрточных слоёв

Каждый слой – 1D-свёртка + GLU

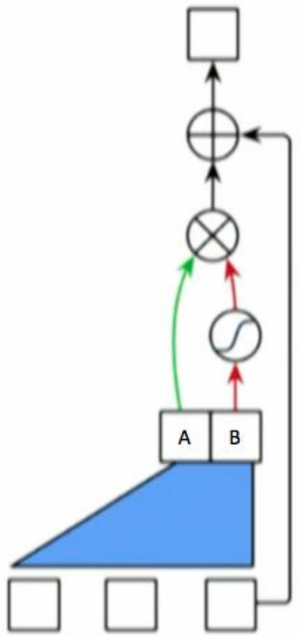
GLU хороша в MT

Каждый слой декодировщика вычисляет внимание (текущий выход декодировщика, финальный выход кодировщика)

Используется «Positional embeddings»

В 2020 до сих пор конкурентна...

Gated Linear Unit (GLU) с прокидыванием связей



свёртка с окном k=3 → два выхода A и B

A · sigmoid(B) + residual

[Gehring et al., 2017 https://arxiv.org/pdf/1705.03122.pdf]

https://github.com/facebookresearch/fairseq-py

Convolutional Sequence to Sequence Learning (ConvS2S)

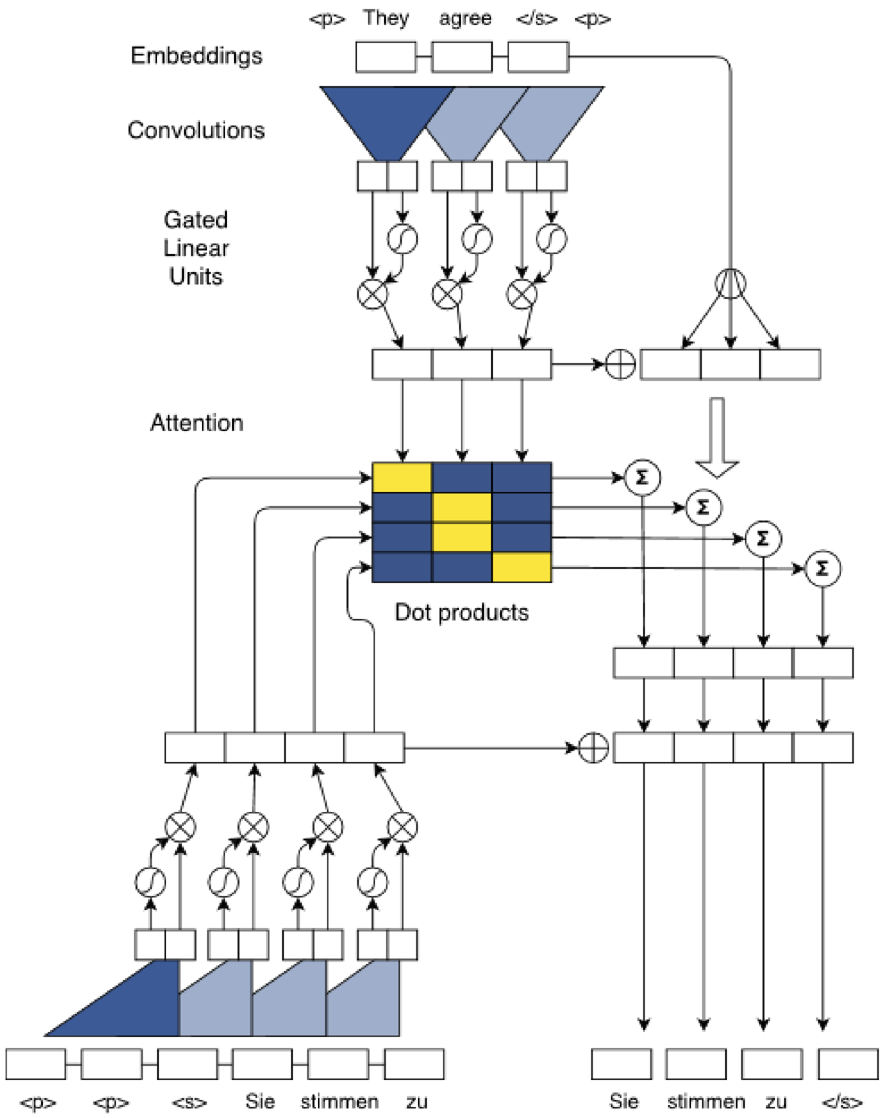


Figure 1. Illustration of batching during training. The English source sentence is encoded (top) and we compute all attention values for the four German target words (center) simultaneously. Our attentions are just dot products between decoder context representations (bottom left) and encoder representations. We add the conditional inputs computed by the attention (center right) to the decoder states which then predict the target words (bottom right). The sigmoid and multiplicative boxes illustrate Gated Linear Units.

	PPL	BLEU
ConvS2S	6.64	21.7
-source position	6.69	21.3
-target position	6.63	21.5
-source & target position	6.68	21.2

Table 4. Effect of removing position embeddings from our model in terms of validation perplexity (valid PPL) and BLEU.

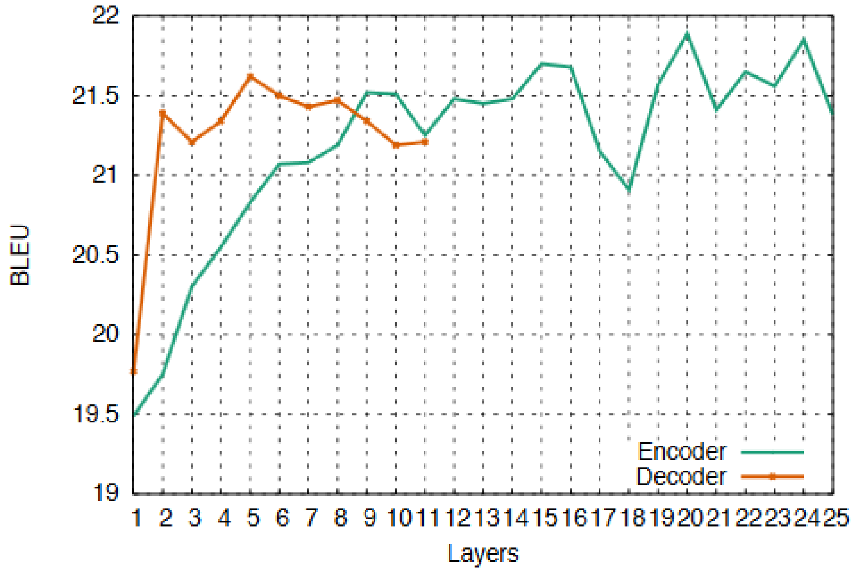


Figure 2. Encoder and decoder with different number of layers.

Attn Layers	PPL	BLEU
1,2,3,4,5	6.65	21.63
1,2,3,4	6.70	21.54
1,2,3	6.95	21.36
1,2	6.92	21.47
1,3,5	6.97	21.10
1	7.15	21.26
2	7.09	21.30
3	7.11	21.19
4	7.19	21.31
5	7.66	20.24

Table 5. Multi-step attention in all five decoder layers or fewer layers in terms of validation perplexity (PPL) and test BLEU.

GNMT (Google’s Neural Machine Translation System)

кодировщик (8 слоёв)-декодировщик (8 слоёв)

нижний уровень кодировщика двунаправленный (дальше конкатенация),

остальные слева-направо (там прокидывание связей и сумма)

модель распределяется по разным GPU, Wordpiece

Во время вывода – low-precision arithmetic

Beam search + length-normalization + coverage penalty (эмпирически хороша)

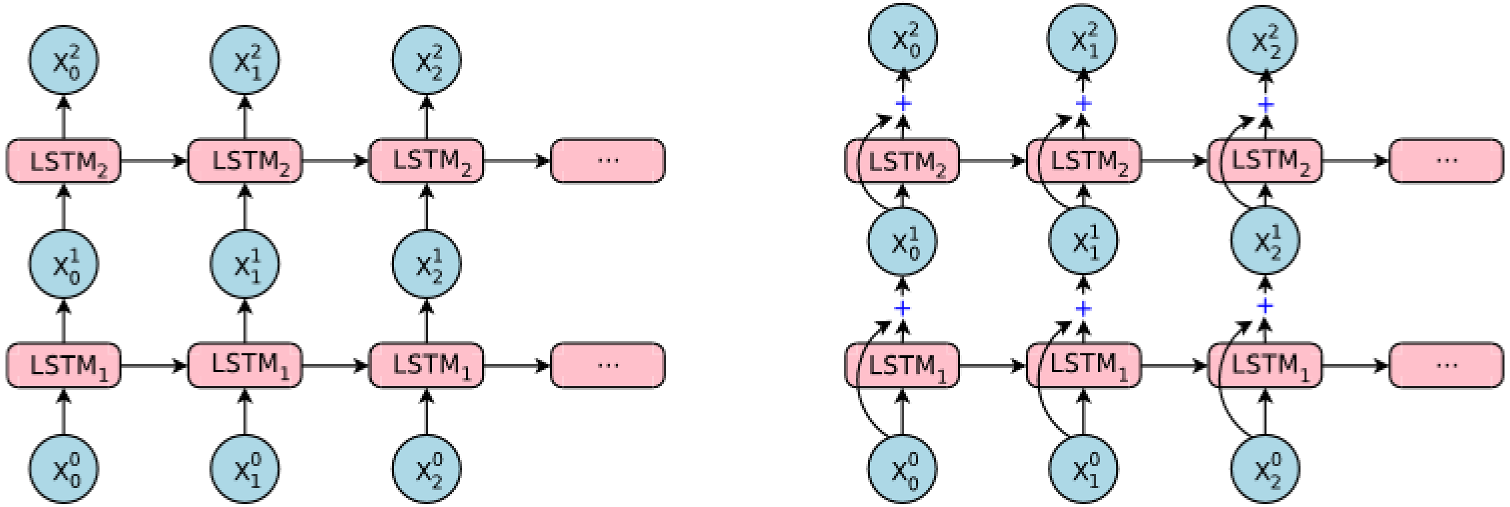


Figure 2: The difference between normal stacked LSTM and our stacked LSTM with residual connections. On the left: simple stacked LSTM layers [41]. On the right: our implementation of stacked LSTM layers with residual connections. With residual connections, input to the bottom LSTM layer ($\mathbf{x_i^0}$’s to LSTM$_1$) is element-wise added to the output from the bottom layer ($\mathbf{x_i^1}$’s). This sum is then fed to the top LSTM layer (LSTM$_2$) as the new input.

Wu, Y. et al «Google’s neural machine translation system: Bridging the gap between human and machine translation» arXiv preprintarXiv:1609.08144

Машинный перевод с помощью НС

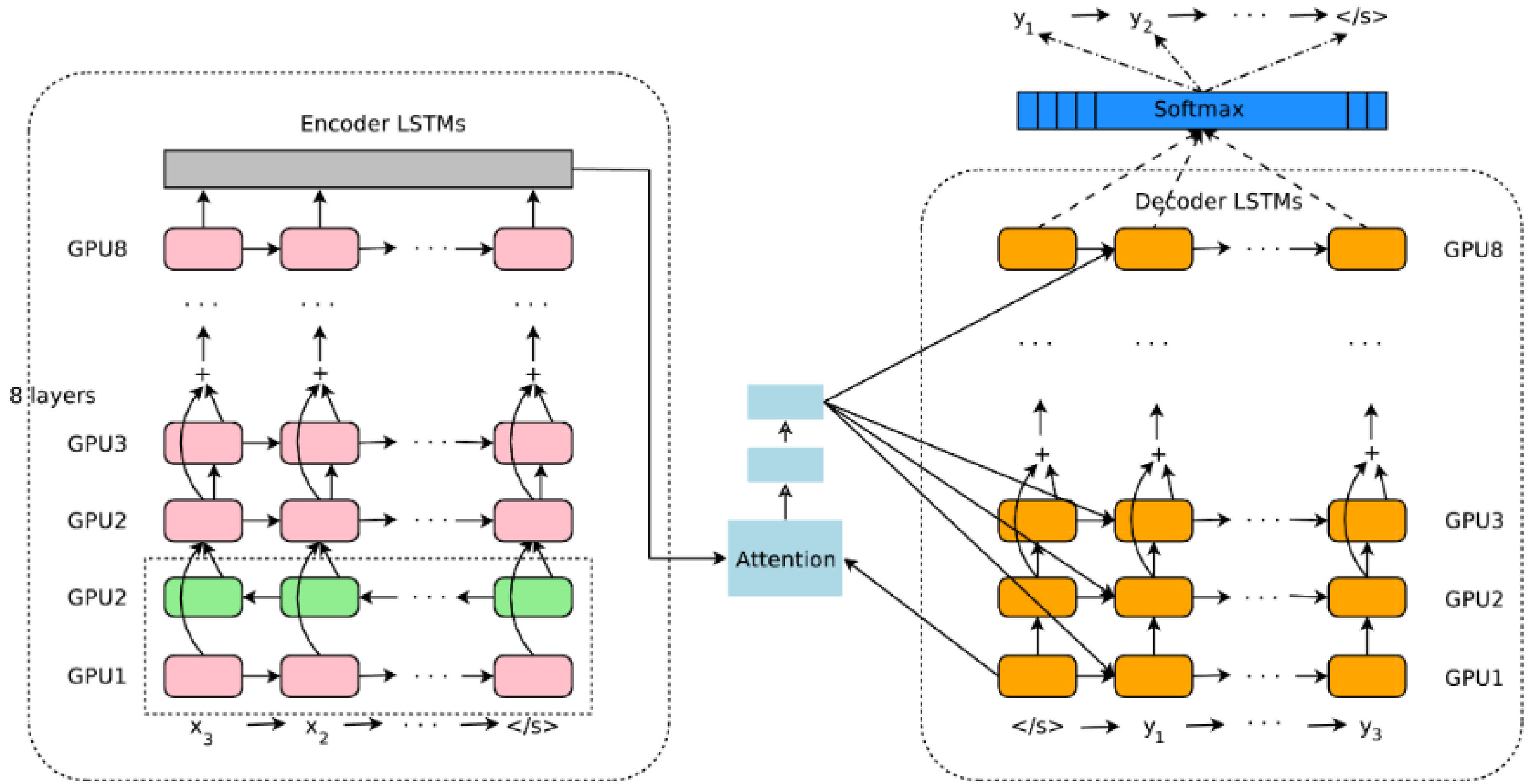


Figure 1: The model architecture of GNMT, Google's Neural Machine Translation system. On the left is the encoder network, on the right is the decoder network, in the middle is the attention module. The bottom encoder layer is bi-directional: the pink nodes gather information from left to right while the green nodes gather information from right to left. The other layers of the encoder are uni-directional. Residual connections start from the layer third from the bottom in the encoder and decoder. The model is partitioned into multiple GPUs to speed up training. In our setup, we have 8 encoder LSTM layers (1 bi-directional layer and 7 uni-directional layers), and 8 decoder layers. With this setting, one model replica is partitioned 8-ways and is placed on 8 different GPUs typically belonging to one host machine. During training, the bottom bi-directional encoder layers compute in parallel first. Once both finish, the uni-directional encoder layers can start computing, each on a separate GPU. To retain as much parallelism as possible during running the decoder layers, we use the bottom decoder layer output only for obtaining recurrent attention context, which is sent directly to all the remaining decoder layers. The softmax layer is also partitioned and placed on multiple GPUs. Depending on the output vocabulary size we either have them run on the same GPUs as the encoder and decoder networks, or have them run on a separate set of dedicated GPUs.

GNMT (Google’s Neural Machine Translation System)

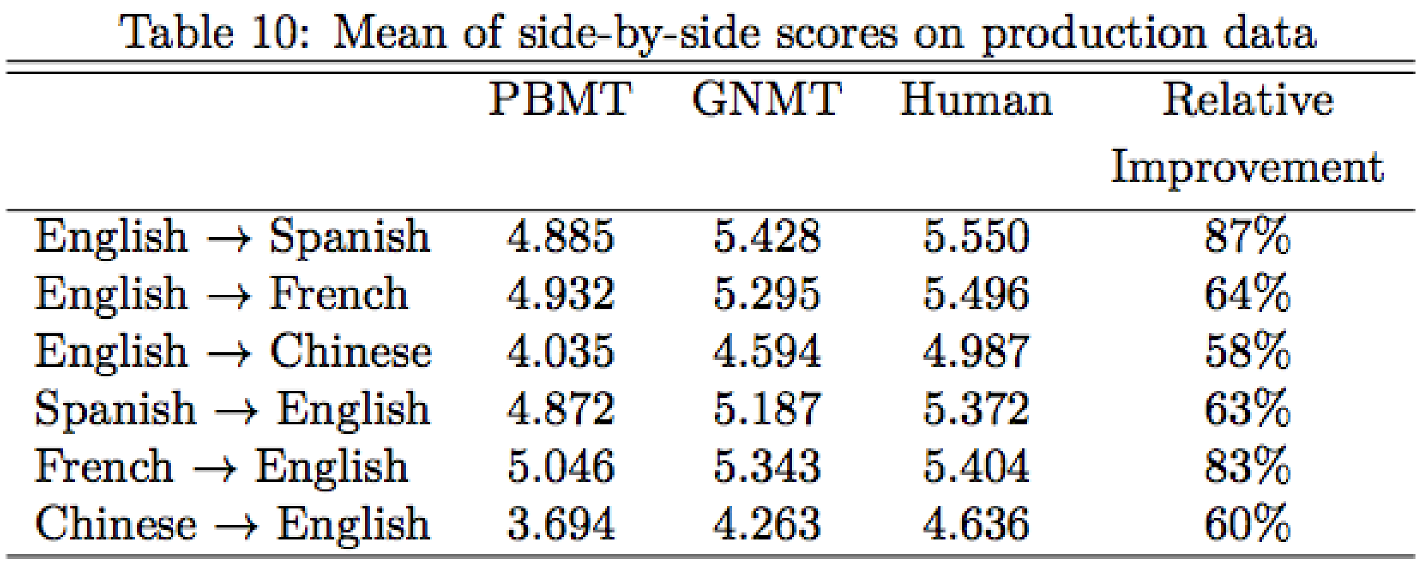


Table 10: Mean of side-by-side scores on production data

	PBMT	GNMT	Human	Relative Improvement
English → Spanish	4.885	5.428	5.550	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.035	4.594	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%

PBMT: Translation by phrase-based statistical translation system used by Google,
GNMT: Translation by our GNMT system
Human: Translation by humans fluent in both languages

RNMT+

endecoder: 6 biLSTM (а была 1bi + 7uni)

decoder: 7 uni

Residual connections: прибавка к слоям >2

Multi-head additive attention

synchronous training strategy

per-gate layernormalization

Chen «The best of both worlds: Combining recent ad-vances in neural machine translation» arXiv preprint arXiv:1804.09849.

доработали идею GNMT

RNMT+

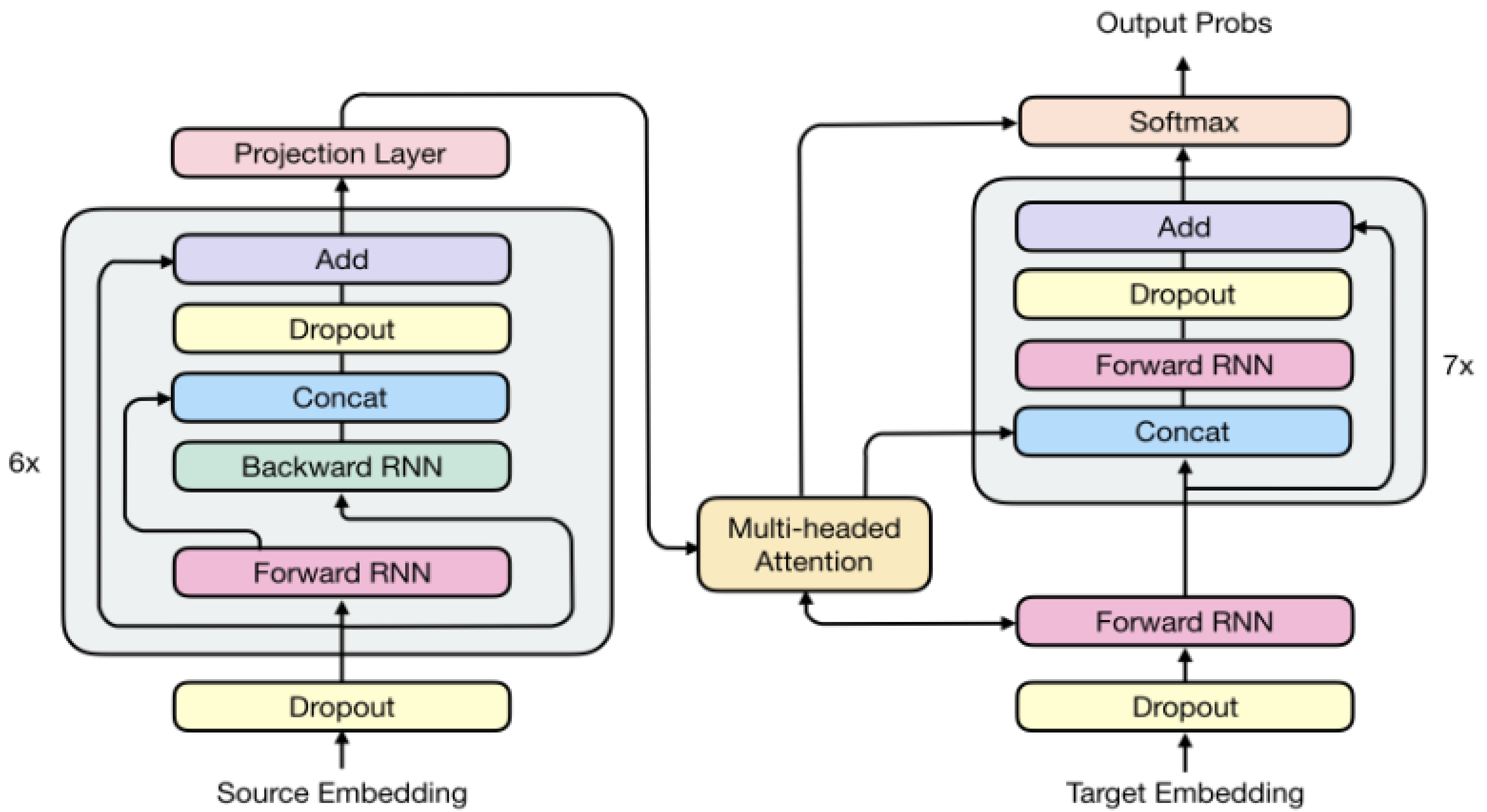


Figure 1: Model architecture of RNMT+. On the left side, the encoder network has 6 bidirectional LSTM layers. At the end of each bidirectional layer, the outputs of the forward layer and the backward layer are concatenated. On the right side, the decoder network has 8 unidirectional LSTM layers, with the first layer used for obtaining the attention context vector through multi-head additive attention. The attention context vector is then fed directly into the rest of the decoder layers as well as the softmax layer.

The Transformer model (Vaswani et al., 2017)

Уже подробно разбирали

normalize → transform → dropout → residual-add

https://github.com/tensorflow/tensor2tensor

Проблема редких слов

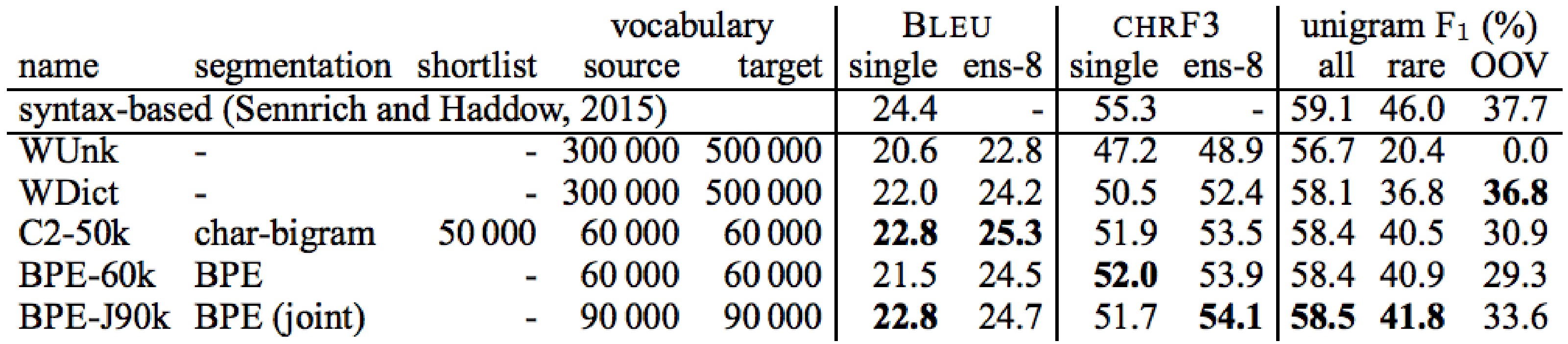
увеличение словаря

подслова (+символы, BPE)

English: Claustrophobia

German: Klaustrophobie

Russian: Клаустрофобия



			vocabulary		BLEU		CHRF3		unigram F_1 (%)		
name	segmentation	shortlist	source	target	single	ens-8	single	ens-8	all	rare	OOV
syntax-based (Sennrich and Haddow, 2015)					24.4	-	55.3	-	59.1	46.0	37.7
WUnk	-	-	300 000	500 000	20.6	22.8	47.2	48.9	56.7	20.4	0.0
WDict	-	-	300 000	500 000	22.0	24.2	50.5	52.4	58.1	36.8	**36.8**
C2-50k	char-bigram	50 000	60 000	60 000	**22.8**	**25.3**	51.9	53.5	58.4	40.5	30.9
BPE-60k	BPE	-	60 000	60 000	21.5	24.5	**52.0**	53.9	58.4	40.9	29.3
BPE-J90k	BPE (joint)	-	90 000	90 000	**22.8**	24.7	51.7	**54.1**	**58.5**	**41.8**	33.6

Table 2: English→German translation performance (BLEU, CHRF3 and unigram F_1) on newstest2015.

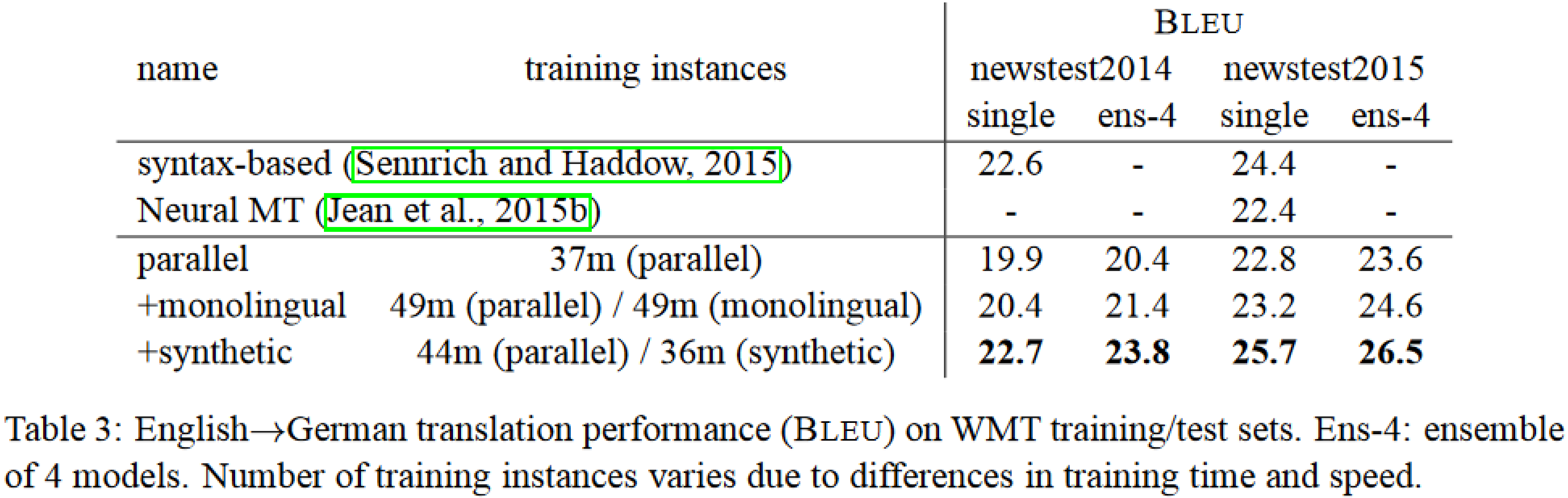
Luong, Minh-Thang, et al. "Addressing the rare word problem in neural machine translation." arXiv preprint arXiv:1410.8206 (2014).

Языковая модель в NMT

Архитектура encoder-decoder NMT имеет достаточную ёмкость, чтобы учить то же, что и языковая модель

Можно использовать данные одного языка, чтобы (под)обучать её

- в паре с пустым предложением
- в паре с синтетическим предложением (back-translation)

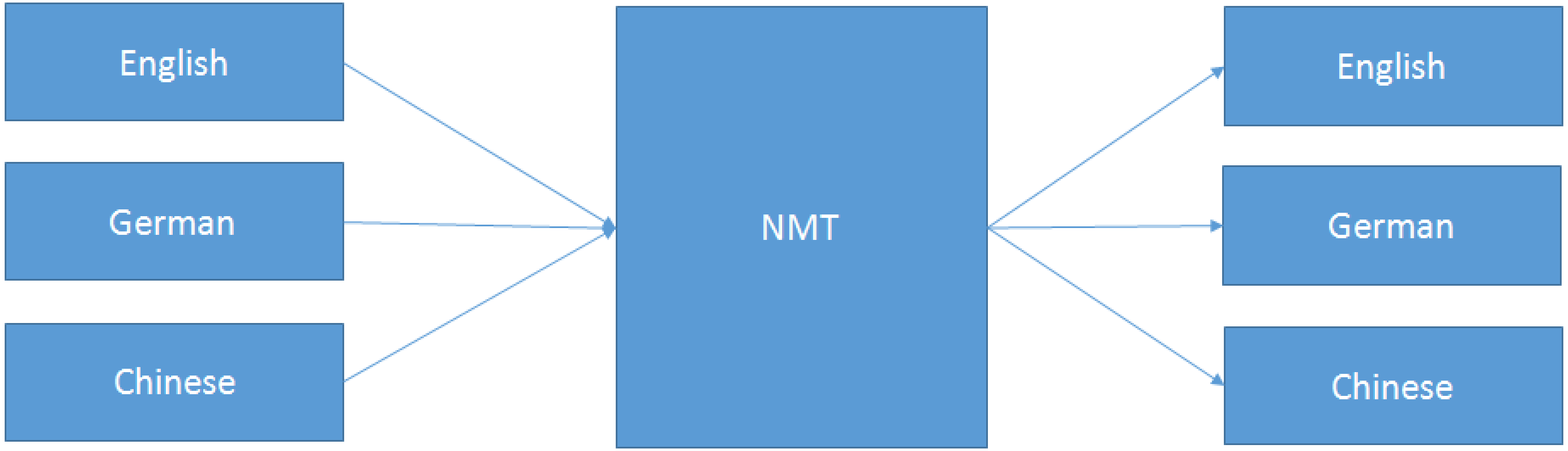


name	training instances	BLEU			
		newstest2014		newstest2015	
		single	ens-4	single	ens-4
syntax-based (Sennrich and Haddow, 2015)		22.6	-	24.4	-
Neural MT (Jean et al., 2015b)		-	-	22.4	-
parallel	37m (parallel)	19.9	20.4	22.8	23.6
+monolingual	49m (parallel) / 49m (monolingual)	20.4	21.4	23.2	24.6
+synthetic	44m (parallel) / 36m (synthetic)	**22.7**	**23.8**	**25.7**	**26.5**

Table 3: English→German translation performance (BLEU) on WMT training/test sets. Ens-4: ensemble of 4 models. Number of training instances varies due to differences in training time and speed.

Sennrich, Rico, Barry Haddow, and Alexandra Birch «Improving neural machine translation models with monolingual data» https://arxiv.org/pdf/1511.06709.pdf

Использование нескольких языков



один кодировщик и один декодировщик на один язык

разделяемый механизм внимания

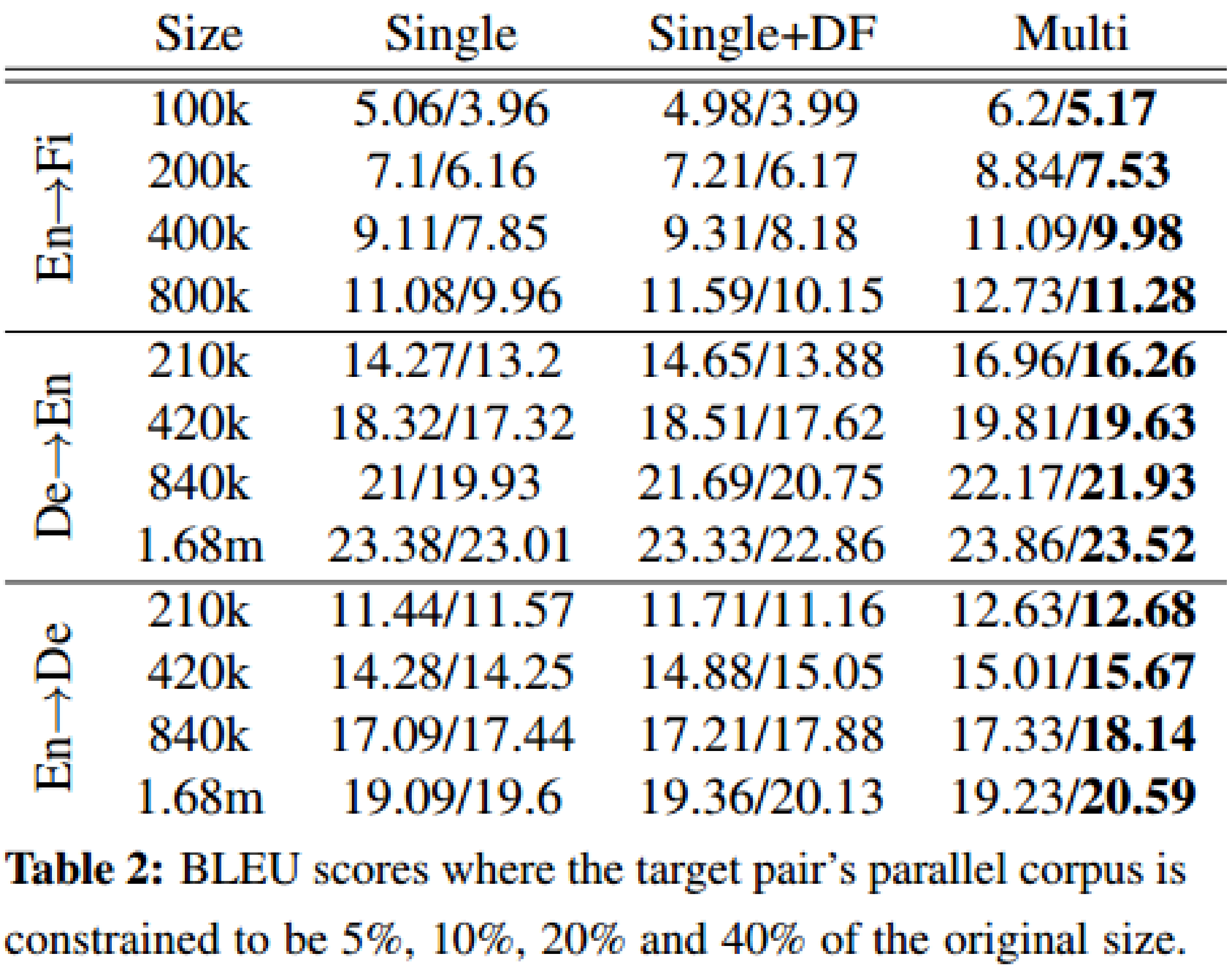
Firat, Orhan, Kyunghyun Cho, and Yoshua Bengio. "Multi-way, multilingual neural machine translation with a shared attention mechanism." arXiv preprint arXiv:1601.01073 (2016).

Много пар языков – разделить кодировщики и декодировщики

Johnson et al. "Google's multilingual NMT system…" ACL 2017

Aharoni et al. "Massively multilingual NMT" ACL 2019

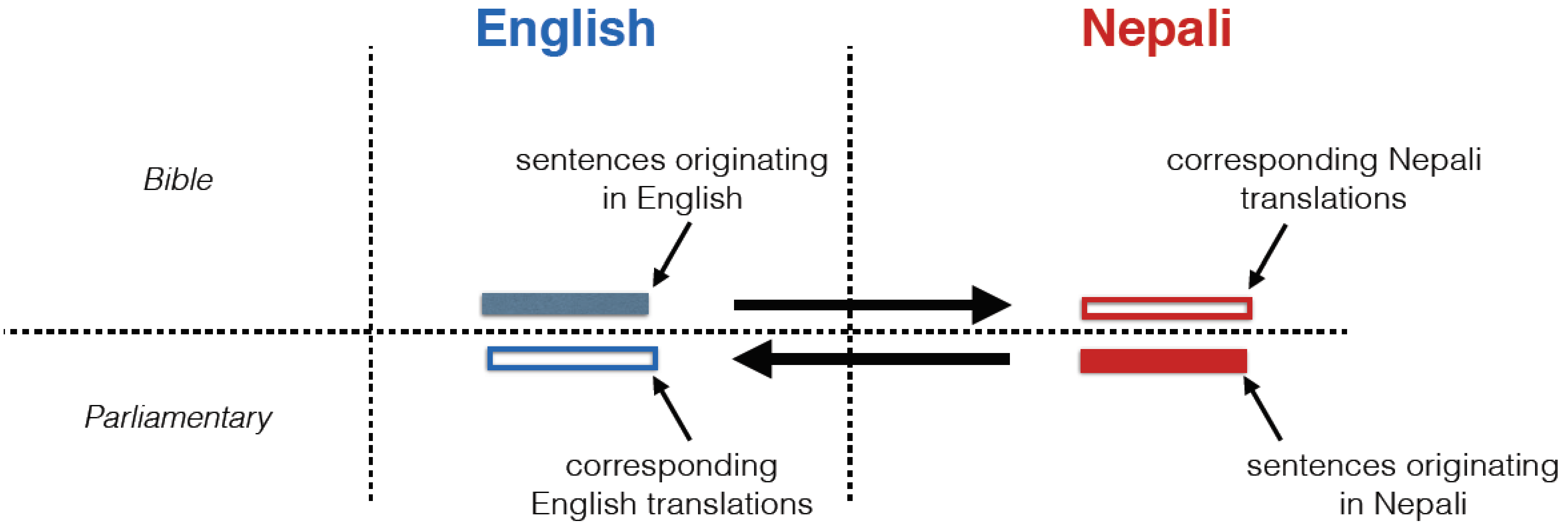
Использование нескольких языков



	Size	Single	Single+DF	Multi
En→Fi	100k	5.06/3.96	4.98/3.99	6.2/**5.17**
	200k	7.1/6.16	7.21/6.17	8.84/**7.53**
	400k	9.11/7.85	9.31/8.18	11.09/**9.98**
	800k	11.08/9.96	11.59/10.15	12.73/**11.28**
De→En	210k	14.27/13.2	14.65/13.88	16.96/**16.26**
	420k	18.32/17.32	18.51/17.62	19.81/**19.63**
	840k	21/19.93	21.69/20.75	22.17/**21.93**
	1.68m	23.38/23.01	23.33/22.86	23.86/**23.52**
En→De	210k	11.44/11.57	11.71/11.16	12.63/**12.68**
	420k	14.28/14.25	14.88/15.05	15.01/**15.67**
	840k	17.09/17.44	17.21/17.88	17.33/**18.14**
	1.68m	19.09/19.6	19.36/20.13	19.23/**20.59**

Table 2: BLEU scores where the target pair's parallel corpus is constrained to be 5%, 10%, 20% and 40% of the original size.

Редкие пары языков

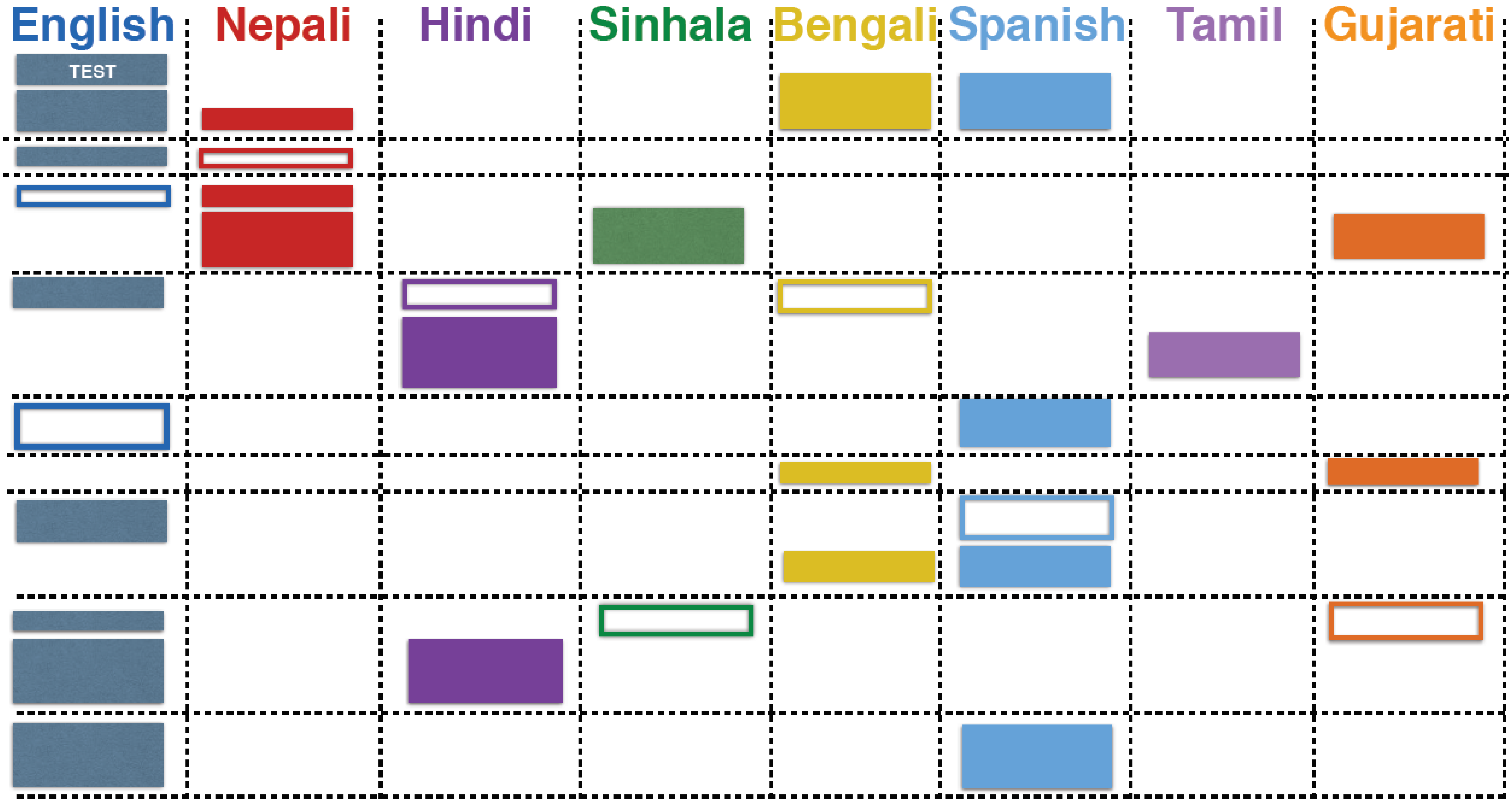

English
Nepali
Bible
Parliamentary
sentences originating in English
corresponding Nepali translations
corresponding English translations
sentences originating in Nepali

человеческий перевод – незакрашенный прямоугольник

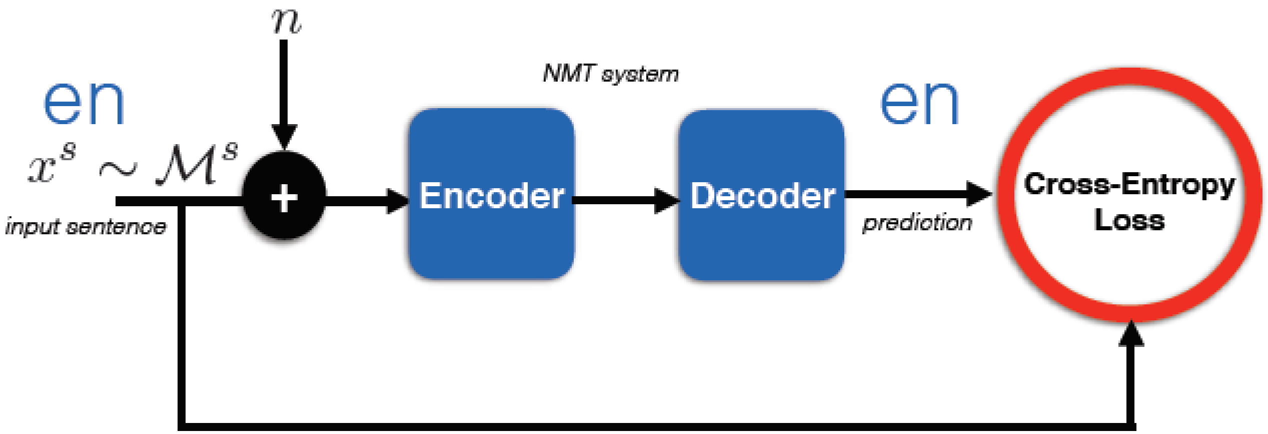
есть пары языков, в которых мало данных для обучения

http://web.stanford.edu/class/cs224n/

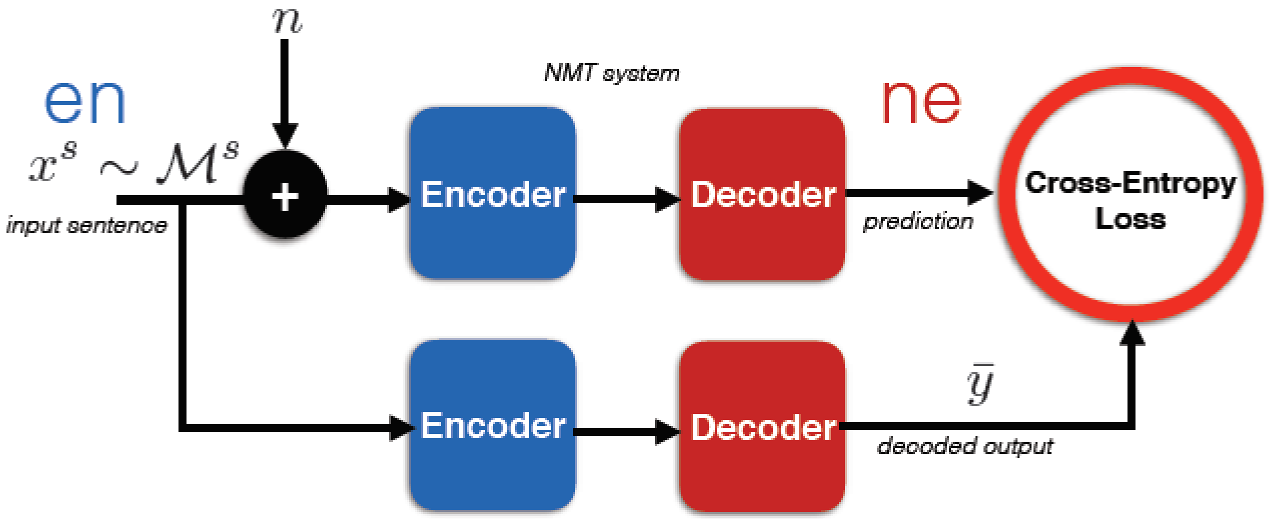
Редкие пары языков



Редкие пары языков



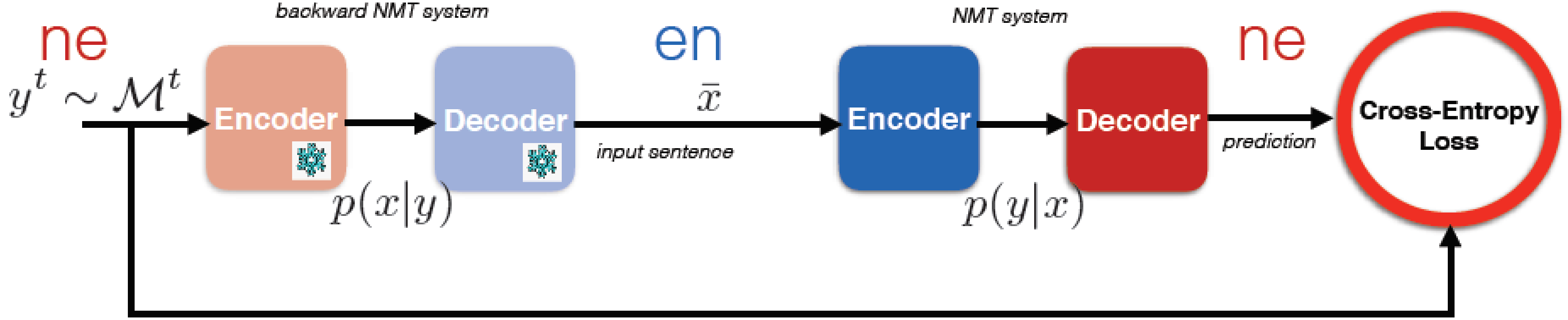
учим автокодировщик, устраняющий шум (+ синонимы и т.п.)



Liu et al. “Multilingual denoising pretraining for NMT” arXiv:2001.08210 2020

He et al. “Revisiting self-training for neural sequence generation” ICLR 2020

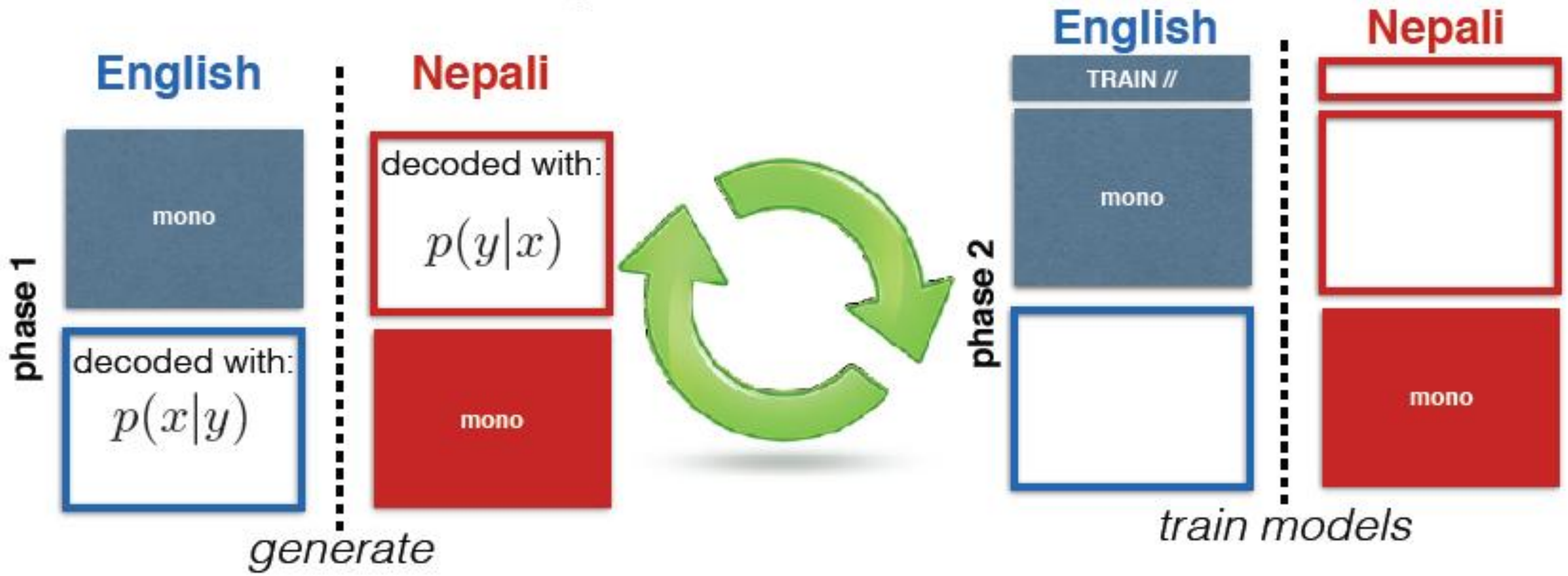
Редкие пары языков



много данных одного на одном из языке – «циклический» кодировщик-декодировщик

Sennrich et al. “Improving NMT models with monolingual data” ACL 2016

Artetxe et al. “An effective approach to unsupervised MT” ACL 2019



Shen et al. “The source-target domain mismatch problem in MT” arXiv:1909.13151 2019

Chen et al. “FBAI WAT’19 Myanmar-English translation task submission” WAT@EMNLP 2019

Проблемы переводчика

неучёт контекста (в том числе забывание начала текста)

пока ещё плохо переводят длинные предложения

различие в домене обучения и теста

гендерные и прочие дискриминации

«он начальник», «она швея»

OOV

редкие пары языков

Выводы

NMT может выполняться end2end-сетью
(не надо оптимизировать отдельные компоненты)

Один подход годится для всех пар языков…

MT – локомотивная область в DL
большинство методов, архитектур появляется здесь

Обзор

Felix Stahlberg "Neural Machine Translation: A Review" // https://arxiv.org/pdf/1912.02047.pdf

Shuoheng Yang, Yuxin Wang, Xiaowen Chu "A Survey of Deep Learning Techniques for NeuralMachine Translation" // https://arxiv.org/pdf/2002.07526.pdf