курс «Глубокое обучение»

Генерация текстов (NLG)

Александр Дьяконов

18 ноября 2019 года

План

Представление слов

Токенизация на подслова

Посимвольный подход

Гибридный подход

С(у/а)ммаризация текста

seq2seq-подход / + attention

Pointer-Generator Networks

Bottom-up summarization

~~NLG + RL~~

Диалоги

Рассказ историй: Storytelling

Рассказ историй по тексту: Hierarchical Neural Story Generation

Генерация поэзии: Hafez, Deep-speare

Coreference Resolution: SOTA / Clustering-Based

Токенизация на подслова (Subword Tokenization)

«subword» → «sub» + «word»

Решение проблемы OOV слов и,
в принципе, составление более адекватного (по частоте слов) словаря:
1) токенизация на подслова
2) пометка слов не из словаря <UNKNOWN>

Дальше рассматриваем 1й вариант…

https://medium.com/@makcedward/how-subword-helps-on-your-nlp-model-83dd1b836f46
https://mlexplained.com/2019/11/06/a-deep-dive-into-the-wonderful-world-of-preprocessing-in-nlp/

Токенизация на подслова (Subword Tokenization)

Причины:

Неформальные слова

«Yeeees! Goooood!»

Транслитерация

«файнтьюнинг»

Динамический словарь

«айфонизация…»

Сложности с разделеним слов

安理会认可利比亚问题柏林峰会成果

ف ل ق ن ا ه

Lebensversicherungsgesellschaftsangestellter

Представление слов

Токенизация на подслова ← сейчас это

- byte-pair encoding (BPE)
- wordpiece
- unigram language model
- sentencepiece

посимвольный подход (представления слов из анализа символов)

- Посимвольная модель для представления слов: Compositional Character Model
- Посимвольные модели: Character-Aware NLM

гибридный подход (действуем на уровне слов, если надо – на уровне символов)

- Compositional Character Model
- Character-Aware NLM

Byte Pair Encoding (BPE)

Rico Sennrich, Barry Haddow, Alexandra Birch
Neural Machine Translation of Rare Words with Subword Units
https://arxiv.org/abs/1508.07909

идея ~ Huffman encoding
ещё есть работа [Philip Gage, 1994] откуда, собственно, термин BPE

возник в работе по машинному переводу
чтобы обучать на данных с одним словарём,
а работать на более широком разнообразии данных

«Abwasserbehandlungsanlange» (нем.) – станция очистки сточных вод

На этапе препроцессинга данных.

Byte Pair Encoding (BPE)

Обучение BPE

- Слово = последовательность токенов (пока символов) + специальный символ конца • (изначально использовались unicode-символы)
- Словарь = все токены (на нулевой итерации – символы)
- Повторять пока не достигли ограничения на размер словаря
  - Назначаем новым токеном объединение двух существующих токенов, которое встречается чаще других пар в корпусе

Применение BPE (возможны варианты)

идём по всем токенам по убыванию частоты – находим соответствующую последовательность символов в корпусе, заменяем на токен

Byte Pair Encoding (BPE)

AABABCABBAABAC

AA – 2
AB – 4 AB = D
BA – 3
BC – 1
CA – 1
BB – 1
AC – 1

ADDCDBADAC

AD – 2 AD = E
DD – 1
DC – 1
CD – 1
DB – 1
DA – 1
AC – 1

EDCDBEAC

На практике

«l$_{</w>}$ like$_{</w>}$ ke$_{</w>}$» → «l$_{</w>}$», «li», «##ke$_{</w>}$», «ke$_{</w>}$»

- различают токен изолированный и токен внутри слова
- есть специальный символ конца слова

Algorithm 1 Learn BPE operations

```
import re, collections

def get_stats(vocab):
  pairs = collections.defaultdict(int)
  for word, freq in vocab.items():
    symbols = word.split()
    for i in range(len(symbols)-1):
      pairs[symbols[i],symbols[i+1]] += freq
  return pairs

def merge_vocab(pair, v_in):
  v_out = {}
  bigram = re.escape(' '.join(pair))
  p = re.compile(r'(?<!\S)' + bigram + r'(?!\S)')
  for word in v_in:
    w_out = p.sub(''.join(pair), word)
    v_out[w_out] = v_in[word]
  return v_out

vocab = {'l o w </w>' : 5, 'l o w e r </w>' : 2,
         'n e w e s t </w>':6, 'w i d e s t </w>':3}
num_merges = 10
for i in range(num_merges):
  pairs = get_stats(vocab)
  best = max(pairs, key=pairs.get)
  vocab = merge_vocab(best, vocab)
  print(best)
```

r ·	→	r·
l o	→	lo
lo w	→	low
e r·	→	er·

Figure 1: BPE merge operations learned from dictionary {'low', 'lowest', 'newer', 'wider'}.

BPE в GPT2:

- вход – последовательность байтов (а не юникод-символов)
- не сливают символы разного типа (e.g. буквы и знаки пунктуации)

«dog», «dog!», «dog?»

Byte Pair Encoding (BPE)

name	segmentation	shortlist	vocabulary source	vocabulary target	BLEU single	BLEU ens-8	CHRF3 single	CHRF3 ens-8	unigram F_1 (%) all	unigram F_1 (%) rare	unigram F_1 (%) 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. Ens-8: ensemble of 8 models. Best NMT system in bold. Unigram F_1 (with ensembles) is computed for all words ($n = 44085$), rare words (not among top 50 000 in training set; $n = 2900$), and OOVs (not in training set; $n = 1168$).

name	segmentation	shortlist	vocabulary source	vocabulary target	BLEU single	BLEU ens-8	CHRF3 single	CHRF3 ens-8	unigram F_1 (%) all	unigram F_1 (%) rare	unigram F_1 (%) OOV
phrase-based (Haddow et al., 2015)					24.3	-	53.8	-	56.0	31.3	16.5
WUnk	-	-	300 000	500 000	18.8	22.4	46.5	49.9	54.2	25.2	0.0
WDict	-	-	300 000	500 000	19.1	22.8	47.5	51.0	54.8	26.5	6.6
C2-50k	char-bigram	50 000	60 000	60 000	**20.9**	**24.1**	49.0	51.6	55.2	27.8	17.4
BPE-60k	BPE	-	60 000	60 000	20.5	23.6	**49.8**	52.7	55.3	29.7	15.6
BPE-J90k	BPE (joint)	-	90 000	100 000	20.4	**24.1**	49.7	**53.0**	**55.8**	**29.7**	**18.3**

Table 3: English→Russian translation performance (BLEU, CHRF3 and unigram F_1) on newstest2015. Ens-8: ensemble of 8 models. Best NMT system in bold. Unigram F_1 (with ensembles) is computed for all words ($n = 55654$), rare words (not among top 50 000 in training set; $n = 5442$), and OOVs (not in training set; $n = 851$).

из оригинальной статьи (CHRF3 – a character n-gram F_3-score – хорош для оценки перевода [Popović, 2015])

WordPiece

Schuster, Nakajima «Japanese and Korea voice search», 2012
https://static.googleusercontent.com/media/research.google.com/ja//pubs/archive/37842.pdf

– в этой задаче огромный словарь, много произношений у одного символа, мало пробелов

WordPiece – это BPE, но при объединении максимизируем правдоподобие, а не частоту

В BERT был реализован WordPiece, но в RoBERTa показали, что использование BPE особо ничего не меняет

WordPiece (in BERT)

1. Подготавливаем большой корпус
2. Определяем желаемы размер словаря подслов (sudwords)
3. Представляем слово = последовательность букв
4. Строим языковую модель LM
5. Новое слово получаем объединяя 2 существующих, максимизируя правдоподобие
6. Если не достигли ограничения на размер словаря или порога для правдоподобия, повторяем п. 5

WordPiece (in BERT)

If my understanding is correct, this means that aside from just the bigram frequency, the frequency of the original symbols that constitute the bigram are also taken into account. The log likelihood of a sentence in a unigram language model (assuming independence between the words in a sentence) is simply the sum of the log frequencies of its constituent symbols. This means merging two symbols will increase the total log likelihood by the log likelihood of the *merged* symbol and decrease it by the log likelihood of the two *original* symbols. Assuming we merge symbols x and y, the increase in the log likelihood is

$$\log p(x,y) - \log p(x) - \log p(y) = \log \frac{\log(p(x))}{\log(p(x))\log(p(y))}$$

Unigram Language Model

Taku Kudo «Subword Regularization: Improving Neural Network Translation Modelswith Multiple Subword Candidates» // Google, 2018, https://arxiv.org/pdf/1804.10959.pdf

- все слова независимы
- ищем наиболее подходящие сегментации слов
- это вероятностный метод $\Rightarrow$ можно сэмплировать сегментации из распределения

т.к. токенизация зависима от LM (от словаря – как сегментировать текст),
а LM от токенизации (нужны частоты)**...**

Unigram Language Model

The unigram language model makes an assumption that each subword occurs independently, and consequently, the probability of a subword sequence $\mathbf{x} = (x_1, \ldots, x_M)$ is formulated as the product of the subword occurrence probabilities $p(x_i)$[3]:

$$P(\mathbf{x}) = \prod_{i=1}^{M} p(x_i), \qquad (6)$$

$$\forall i \; x_i \in \mathcal{V}, \; \sum_{x \in \mathcal{V}} p(x) = 1,$$

where $\mathcal{V}$ is a pre-determined vocabulary. The most probable segmentation $\mathbf{x}^*$ for the input sentence X is then given by

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{S}(X)} P(\mathbf{x}), \qquad (7)$$

where $\mathcal{S}(X)$ is a set of segmentation candidates built from the input sentence X. $\mathbf{x}^*$ is obtained with the Viterbi algorithm (Viterbi, 1967).

If the vocabulary $\mathcal{V}$ is given, subword occurrence probabilities $p(x_i)$ are estimated via the EM algorithm that maximizes the following marginal likelihood $\mathcal{L}$ assuming that $p(x_i)$ are hidden variables.

$$\mathcal{L} = \sum_{s=1}^{|D|} \log(P(X^{(s)})) = \sum_{s=1}^{|D|} \log\Big(\sum_{\mathbf{x} \in \mathcal{S}(X^{(s)})} P(\mathbf{x}) \Big)$$

Unigram Language Model

1. **Создать большой словарь**
   эвристикой – например, буквы + наиболее частые подслова (Enhanced Suffix Array algorithm [Nong et al., 2009]), можно использовать BPE
2. **Повторять, пока размер словаря не достигнет порога**
   a. **Зафиксировав словарь максимизировать p(x) с помощью EM**
   b. **Для каждого подслова w вычислить $loss_w$ – насколько уменьшится правдоподобие при удалении слова из словаря**
   c. **Оставить 80% слов с максимальным $loss_w$**
      (символы оставлять всегда, чтобы не было OOV)

p(x) см. в картинке

Сэмплирование из сегментаций

можно получать сегментацию и её вероятность…

Subwords (_ means spaces)	Vocabulary id sequence
_Hell/o/_world	13586 137 255
_H/ello/_world	320 7363 255
_He/llo/_world	579 10115 255
_/He/l/l/o/_world	7 18085 356 356 137 255
_H/el/l/o/_/world	320 585 356 137 7 12295

Table 1: Multiple subword sequences encoding the same sentence “Hello World”

Sentencepiece

Taku Kudo, John Richardson «SentencePiece: A simple and language independent subword tokenizerand detokenizer for Neural Text Processing» //
https://www.aclweb.org/anthology/D18-2012.pdf
https://github.com/google/sentencepiece

по сути это библиотека от автора ULM

до сих пор нужна была претокенизация (pretokenization)

1) Не во всех языках слова разделяются пробелами…

2) Есть такая проблема:

«I like it somehow» → #%@$ → «I like it somehow »

теряем информацию о символах (пробелах)

Sentencepiece

- **Пусть наш текст – последовательность unicode characters**
- **Используем BPE или Unigram Language Model**

тогда пробелы включатся в токенизацию

«I like it somehow » → «I», «like», «_it», «__some», «how»

Пример использования библиотеки

```
# загружаем библиотеку
import sentencepiece as spm
# обучение из файла text.txt
spm.SentencePieceTrainer.Train('--input=test/text.txt
                                --model_prefix=m --vocab_size=1000')
# загружаем обученную модель
sp = spm.SentencePieceProcessor()
sp.Load("m.model")
# теперь кодируем текст
sp.EncodeAsIds("I like ke")
```

BPE-Dropout

Ivan Provilkov, Dmitrii Emelianenko, Elena Voita «BPE-Dropout: Simple and Effective Subword Regularization» // https://arxiv.org/pdf/1910.13267.pdf

- **у BPE каждое слово имеет однозначную сегментацию**
- **подслова редких слов не очень интерпретируемы**

BPE-Dropout

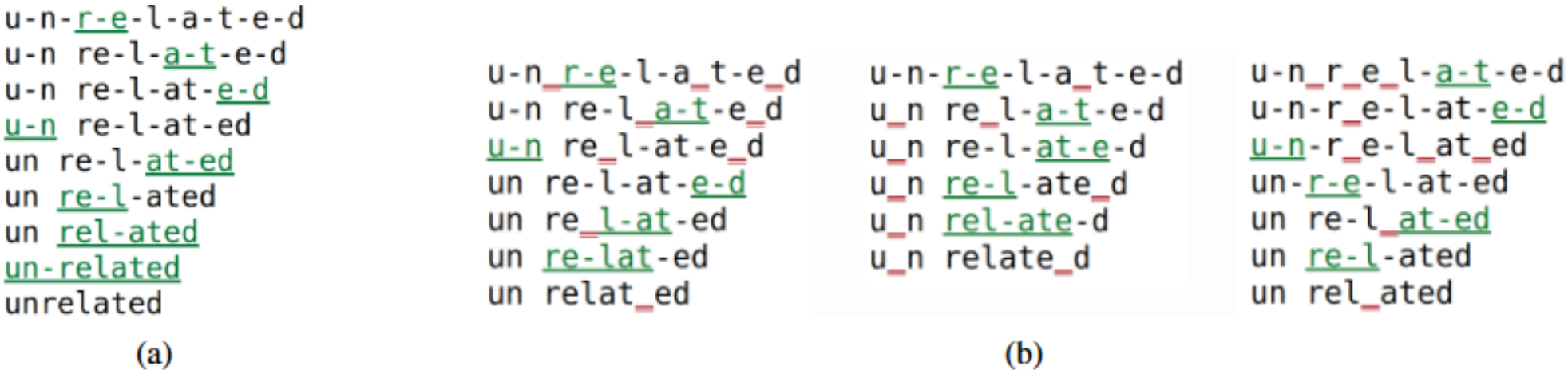


Figure 1: Segmentation process of the word '*unrelated*' using (a) BPE, (b) *BPE-dropout*. Hyphens indicate possible merges (merges which are present in the merge table); merges performed at each iteration are shown in green, dropped – in red.

BPE-Dropout использует словарь и таблицу слияний BPE, но на каждом шаге слияния случайно его пропускает («дропает»)

p=0 $\Rightarrow$ BPE

p=1 $\Rightarrow$ сегментация по буквам

BPE-Dropout

Algorithm 1: BPE-dropout

current_split ← characters from input_word;
do
  merges ← all possible merges of tokens from *current_split*;
  for *merge from merges* **do**
    `/* The only difference from BPE */`
    remove *merge* from *merges* with the probability p;
  end
  if *merges is not empty* **then**
    merge ← select the merge with the highest priority from *merges*;
    apply *merge* to *current_split*;
  end
while *merges is not empty*;
return *current_split*;

	BPE	Kudo (2018)	*BPE-dropout*
IWSLT15			
En-Vi	31.78	32.43	**33.27**
Vi-En	30.83	32.36	**32.99**
En-Zh	21.07	**23.15**	**23.27**
Zh-En	18.29	21.10	**21.45**
IWSLT17			
En-Fr	39.37	39.45	**40.02**
Fr-En	38.18	38.88	**39.39**
En-Ar	13.89	14.43	**15.05**
Ar-En	31.90	32.80	**33.72**
WMT14			
En-De	27.41	**27.82**	**28.01**
De-En	32.69	33.65	**34.19**
ASPEC			
En-Ja	43.69	**44.92**	44.19
Ja-En	30.77	**31.23**	**31.29**

Table 2: BLEU scores. Bold indicates the best score and all scores whose difference from the best is not statistically significant (with p-value of 0.05). (Statistical significance is computed via bootstrapping (Koehn, 2004).)

BPE-Dropout

	BPE	*BPE-dropout*		
		src-only	dst-only	both
250k	26.94	27.98	27.71	**28.40**
500k	29.28	**30.12**	29.40	**29.89**
1m	30.53	**31.09**	30.62	**31.23**
4m	33.38	**33.89**	33.46	**33.85**

Table 3: BLEU scores for models trained with *BPE-dropout* on a single side of a translation pair or on both sides. Models trained on random subsets of WMT14 En-Fr dataset. Bold indicates the best score and all scores whose difference from the best is not statistically significant (with p-value of 0.05).

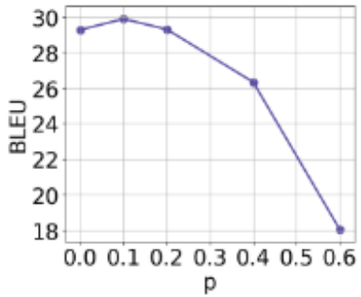


Figure 2: BLEU scores for the models trained with *BPE-dropout* with different values of p. WMT14 En-Fr, 500k sentence pairs.

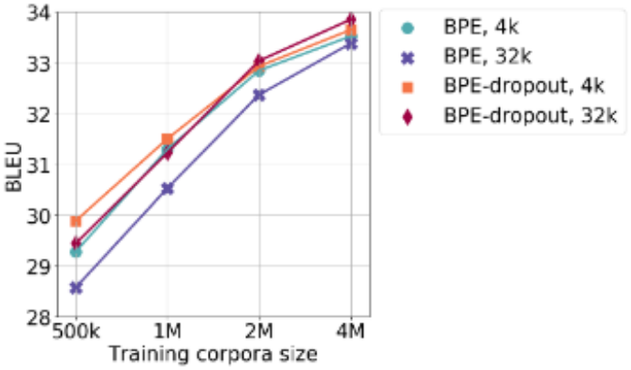


Figure 3: BLEU scores. Models trained on random subsets of WMT14 En-Fr.

withdra		resul		meeting		olec		comptroll	
BPE	BPE-dropout	BPE	BPE-dropout	BPE	BPE-dropout	BPE	BPE-dropout	BPE	BPE-dropout
aimed	withd	undert	result	meetings	meetings	olecular	molec	icial	comptrollership
molecules	withdrawal	checkl	results	meet	meet	molecules	olecular	supervis	comptroller
aromatic	withdraw	maastr	resulting	session	eting	ljubl	molecule	&	troll
specialties	withdrawn	&	resulted	conference	me	zona	molecular	subcomm	controll
publishers	withdrew	unisp	ults	met	etings	choler	molecules	yugosl	contoller
chain	withdrawals	phili	res	workshop	met	oler	aec	trigg	controlled
americ	withdrawing	ζ	resultant	meets	meets	ospheric	oler	sophistic	controllers
chron	dra	preca	ult	sessions	session	olar	tolu	obstac	control
eager	retire	prosecut	ul	convened	et	elic	omet	reag	contro
ighty	reti	tali	outcome	reunion	conference	ochlor	olip	entals	controls

Figure 5: Examples of nearest neighbours in the source embedding space of models trained with BPE and *BPE-dropout* Models trained on WMT14 En-Fr (4m).

качество лучше на вложениях (embeddings)

Итоги

	год	Как строим	Что оптимизируется	Где применяется
BPE	2016 / 1994	Объединяем слова	Частота вхождения	GPT2
WordPiece	2012	Объединяем слова	Правдоподобие	BERT
Unigram Language Model	2018	Вероятностное семейство сегментаций		
Sentencepiece	2018	Больше реализация, чем алгоритм		
BPE-Dropout	2019	BPE + Dropout		

Сравнение BPE-вложения с токенами(word2vec), посимвольным (character) и FastText

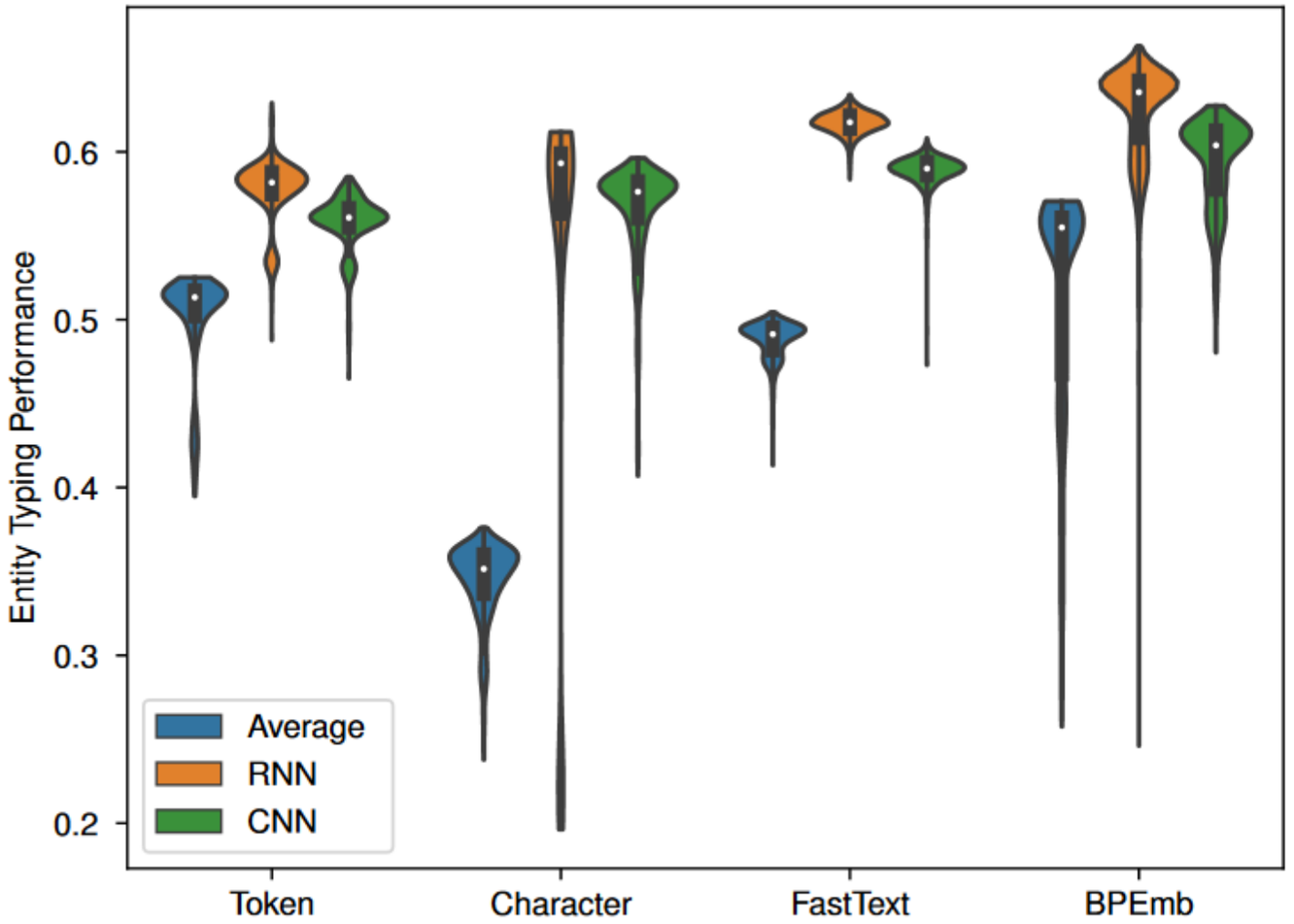


Figure 1: English entity typing performance of subword embeddings across different architectures. This violin plot shows smoothed distributions of the scores obtained during hyper-parameter search. White points represent medians, boxes quartiles. Distributions are cut to reflect highest and lowest scores.

https://arxiv.org/pdf/1710.02187.pdf

Представление слов

~~токенизация на подслова~~

- ~~byte-pair encoding (BPE)~~
- ~~wordpiece~~
- ~~unigram language model~~
- ~~sentencepiece~~

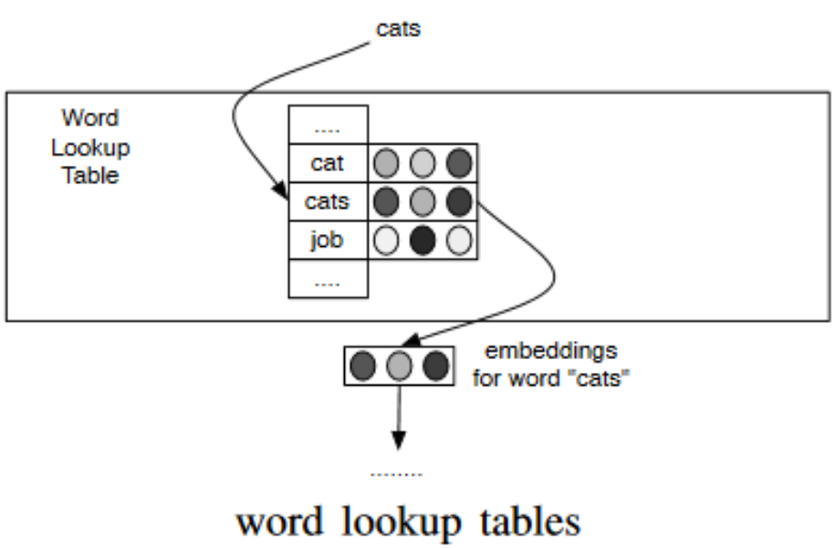
посимвольный подход (представления слов из анализа символов) ← **сейчас это**

- Посимвольная модель для представления слов: Compositional Character Model
- Посимвольные модели: Character-Aware NLM

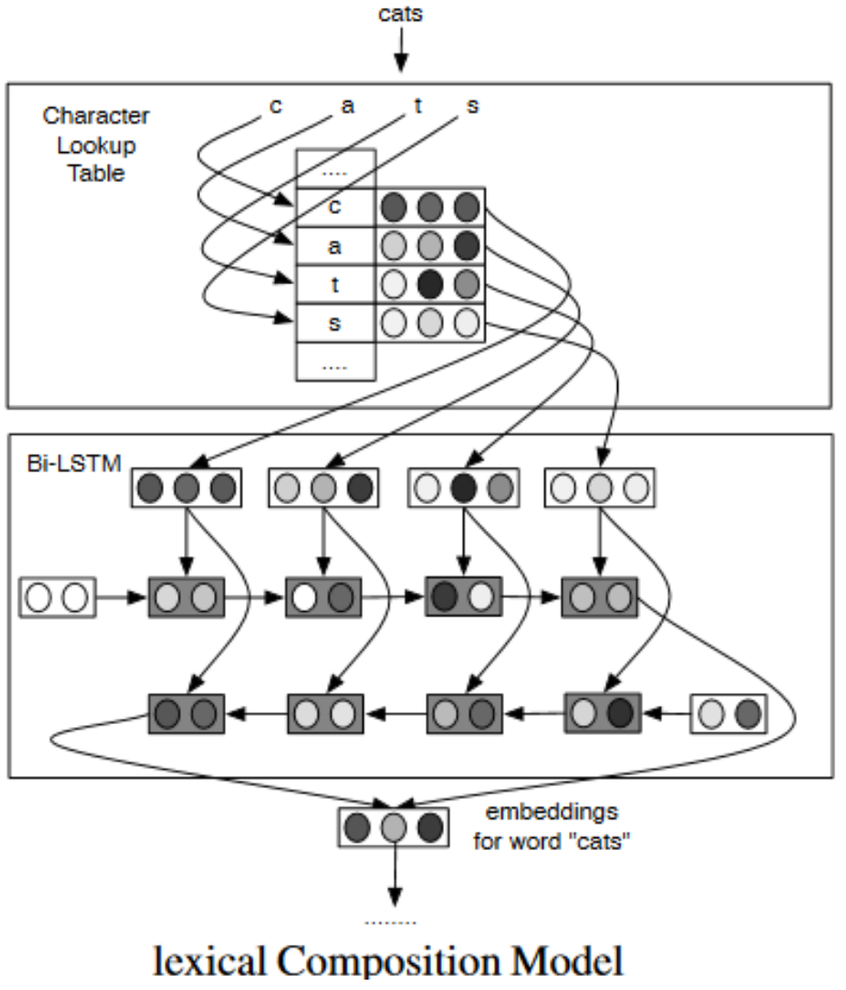
гибридный подход (действуем на уровне слов, если надо – на уровне символов)

- Compositional Character Model
- Character-Aware NLM

Посимвольная модель для представления слов: Compositional Character Model



word lookup tables



lexical Composition Model

Wang Ling et. al. «Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation» // https://www.aclweb.org/anthology/D15-1176.pdf

Посимвольная модель для представления слов: Compositional Character Model

biLSTM

различаем буквы в разных регистрах

каждый символ – ONE · проекционная матрица

(поэтому по-сути табличное представление символа)

Представление слова =

$$e = W_1 h^{\rightarrow}_{\text{last}} + W_2 h^{\leftarrow}_{\text{first}} + b$$

Посимвольная модель для представления слов: Compositional Character Model

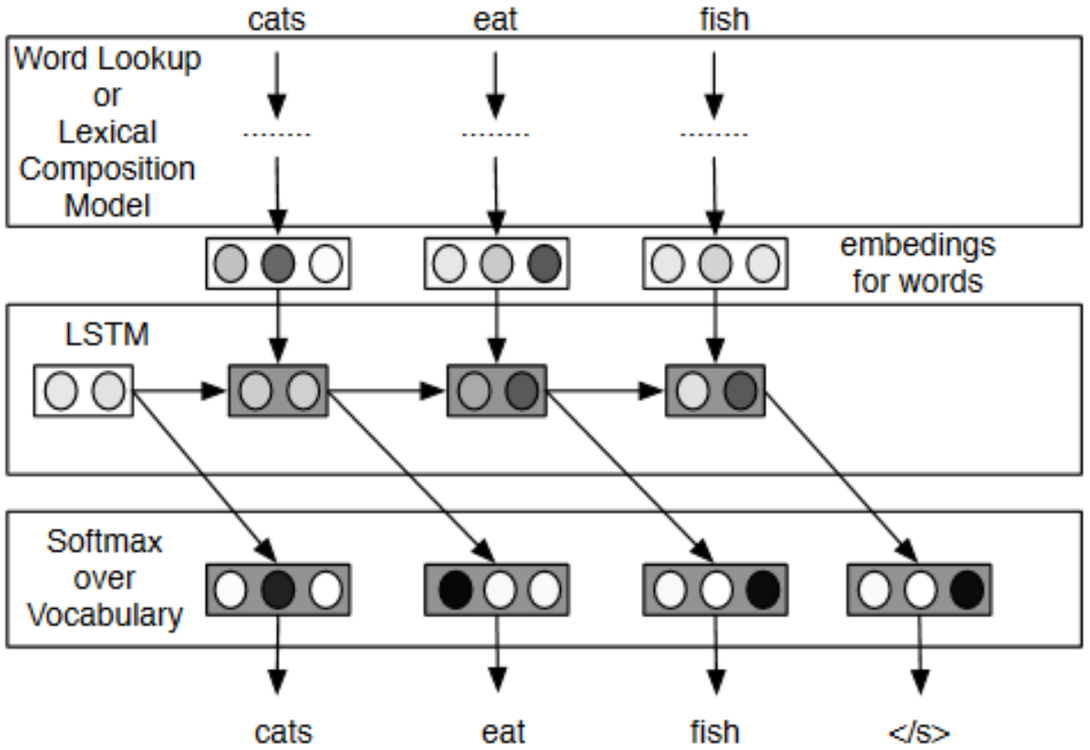


Figure 2: Illustration of our neural network for Language Modeling.

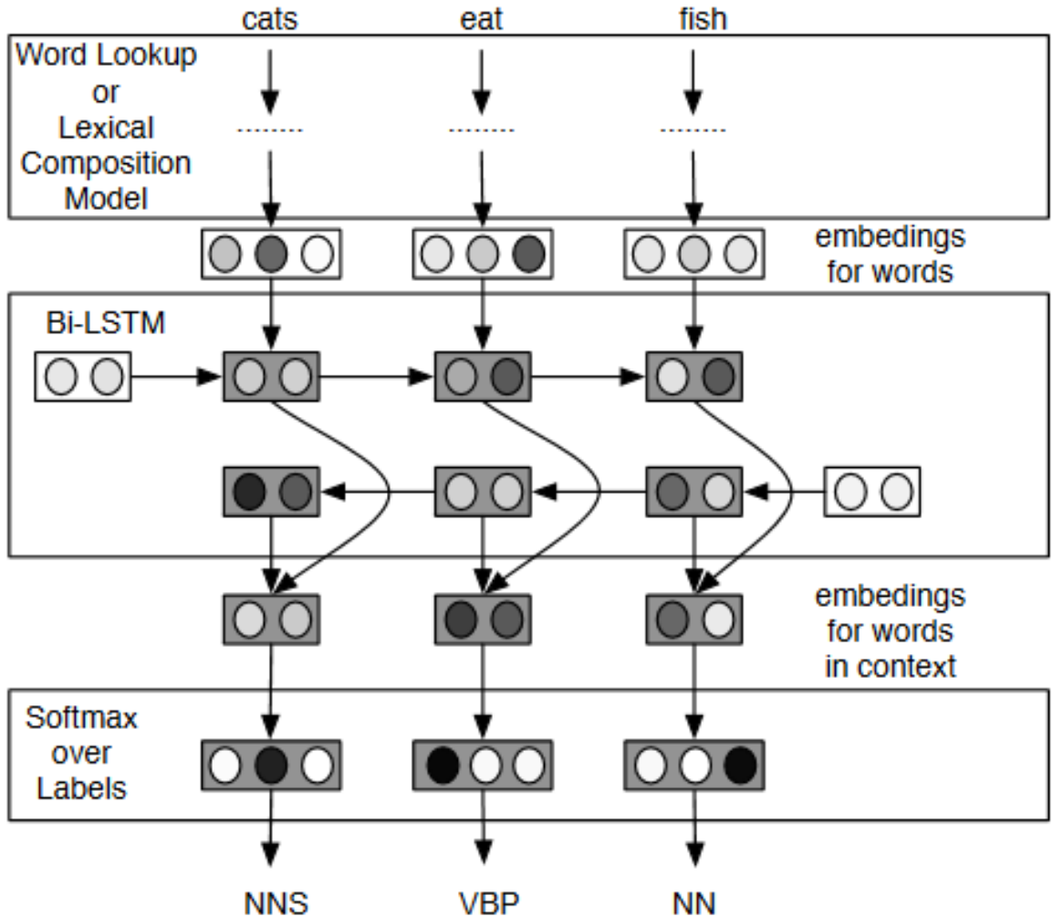


Figure 3: Illustration of our neural network for POS tagging.

Посимвольные модели: Character-Aware NLM

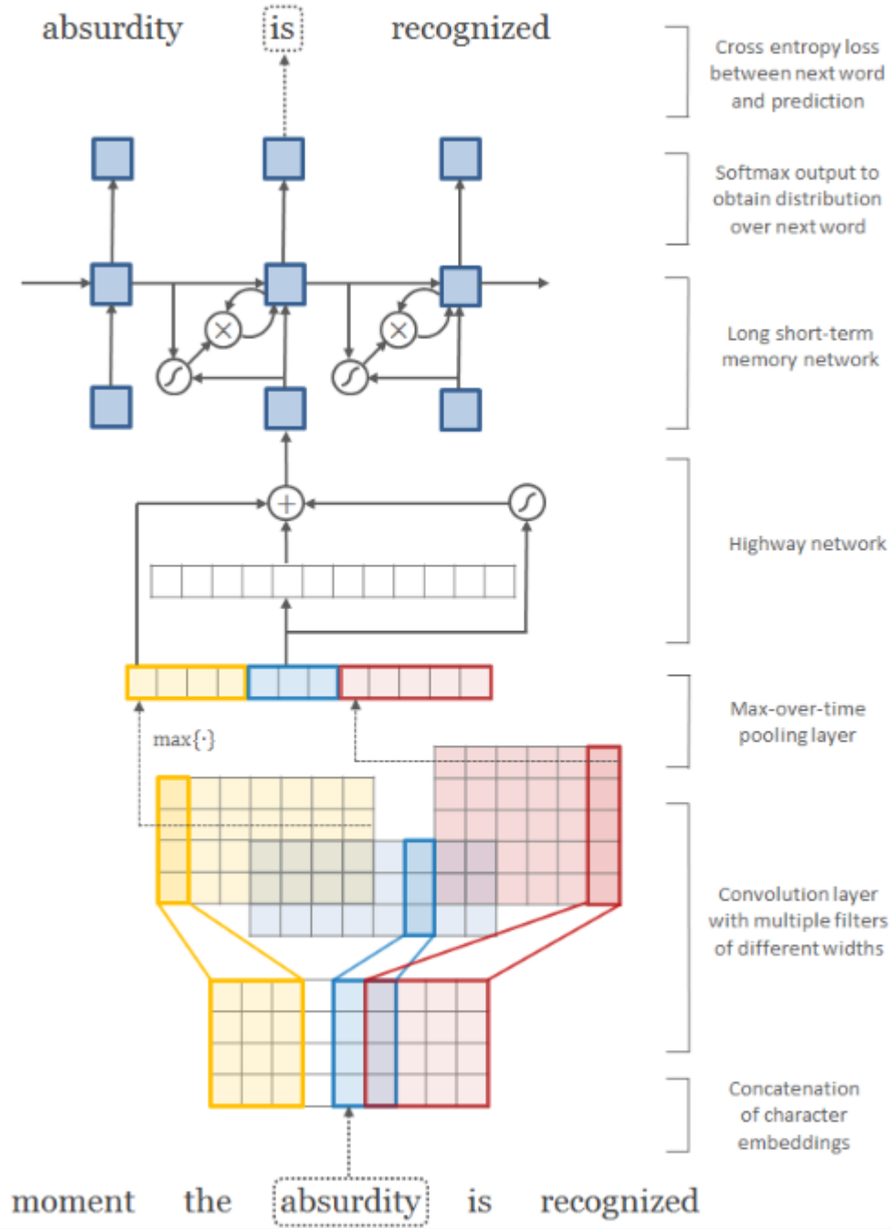


Figure 1: Architecture of our language model applied to an example sentence. Best viewed in color. Here the model takes *absurdity* as the current input and combines it with the history (as represented by the hidden state) to predict the next word, *is*. First layer performs a lookup of character embeddings (of dimension four) and stacks them to form the matrix $\mathbf{C}^k$. Then convolution operations are applied between $\mathbf{C}^k$ and multiple filter matrices. Note that in the above example we have twelve filters—three filters of width two (blue), four filters of width three (yellow), and five filters of width four (red). A max-over-time pooling operation is applied to obtain a fixed-dimensional representation of the word, which is given to the highway network. The highway network's output is used as the input to a multi-layer LSTM. Finally, an affine transformation followed by a softmax is applied over the hidden representation of the LSTM to obtain the distribution over the next word. Cross entropy loss between the (predicted) distribution over next word and the actual next word is minimized. Element-wise addition, multiplication, and sigmoid operators are depicted in circles, and affine transformations (plus nonlinearities where appropriate) are represented by solid arrows.

Посимвольные модели: Character-Aware NLM

слова воспринимаем как цепочки букв, обрабатываем специальной сетью,
а дальше уже всё на уровне слов

свёртки на символьном уровне
max-over-time pooling

highway network из работы [Srivastava et al., 2015]

$$\mathbf{z} = \mathbf{t} \odot g(\mathbf{W}_H \mathbf{y} + \mathbf{b}_H) + (\mathbf{1} - \mathbf{t}) \odot \mathbf{y}$$

~ моделирование n-грамм-закономерностей

иерархический softmax (для большого выходного словаря)
обучение с обрезанным обратным распространением (truncated backprop through time)

Yoon Kim, Yacine Jernite, David Sontag, Alexander M. Rush «Character-Aware Neural Language Models» 2015 // https://arxiv.org/abs/1508.06615

Посимвольные модели: Character-Aware NLM

мораль…

- **не факт, что представления слов нужны на входах в сеть**
- **CNN + Highway Network над символами – выцепляют богатую семантику и структурную информацию**
- **можно строить решение из блоков…**

Другой подход: поиск представления с помощью посимвольной модели

	In Vocabulary					Out-of-Vocabulary		
	while	*his*	*you*	*richard*	*trading*	*computer-aided*	*misinformed*	*looooook*
LSTM-Word	*although*	*your*	*conservatives*	*jonathan*	*advertised*	–	–	–
	letting	*her*	*we*	*robert*	*advertising*	–	–	–
	though	*my*	*guys*	*neil*	*turnover*	–	–	–
	minute	*their*	*i*	*nancy*	*turnover*	–	–	–
LSTM-Char (before highway)	*chile*	*this*	*your*	*hard*	*heading*	*computer-guided*	*informed*	*look*
	whole	*hhs*	*young*	*rich*	*training*	*computerized*	*performed*	*cook*
	meanwhile	*is*	*four*	*richer*	*reading*	*disk-drive*	*transformed*	*looks*
	white	*has*	*youth*	*richter*	*leading*	*computer*	*inform*	*shook*
LSTM-Char (after highway)	*meanwhile*	*hhs*	*we*	*eduard*	*trade*	*computer-guided*	*informed*	*look*
	whole	*this*	*your*	*gerard*	*training*	*computer-driven*	*performed*	*looks*
	though	*their*	*doug*	*edward*	*traded*	*computerized*	*outperformed*	*looked*
	nevertheless	*your*	*i*	*carl*	*trader*	*computer*	*transformed*	*looking*

Table 6: Nearest neighbor words (based on cosine similarity) of word representations from the large word-level and character-level (before and after highway layers) models trained on the PTB. Last three words are OOV words, and therefore they do not have representations in the word-level model.

Другой подход: поиск представления с помощью посимвольной модели

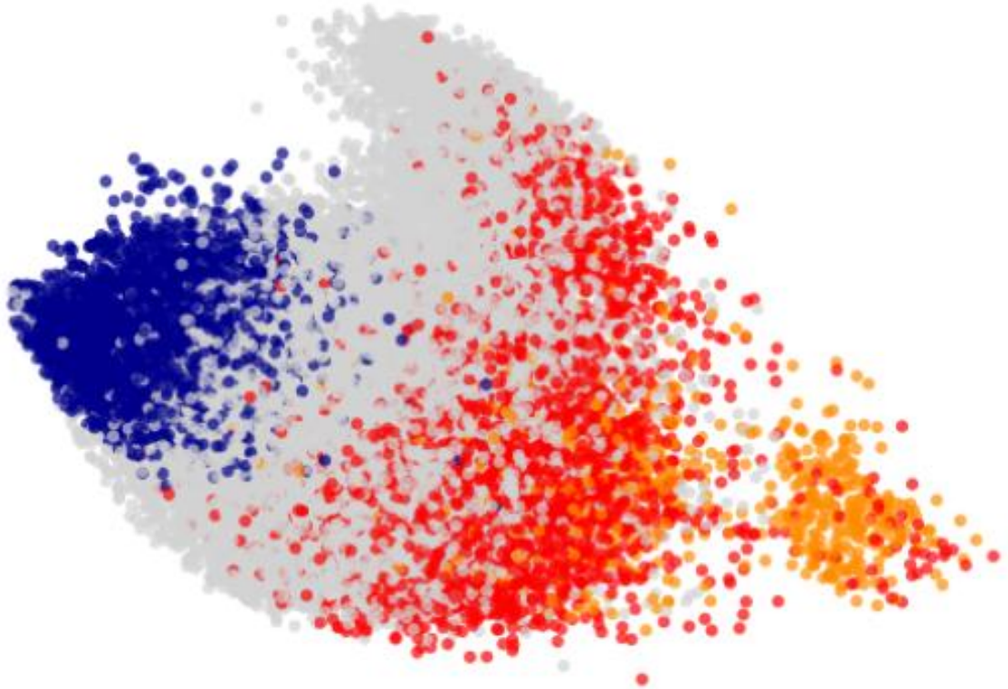


Figure 2: Plot of character n-gram representations via PCA for English. Colors correspond to: prefixes (red), suffixes (blue), hyphenated (orange), and all others (grey). Prefixes refer to character n-grams which start with the start-of-word character. Suffixes likewise refer to character n-grams which end with the end-of-word character.

пропустили n-граммы через сеть…

Представление слов

~~**Слова на кусочки**~~

~~• byte-pair encoding (BPE)~~

~~• wordpiece~~

~~• unigram language model~~

~~• sentencepiece~~

~~**посимвольный подход** (представления слов из анализа символов)~~

~~• Посимвольная модель для представления слов: Compositional Character Model~~

~~• Посимвольные модели: Character-Aware NLM~~

сейчас это ↓

гибридный подход (действуем на уровне слов, если надо – на уровне символов)

• Compositional Character Model

• Character-Aware NLM

Гибридный подход: Hybrid NMT

работаем на уровне слов (4 слойная LSTM)

если надо спускаемся на уровень букв

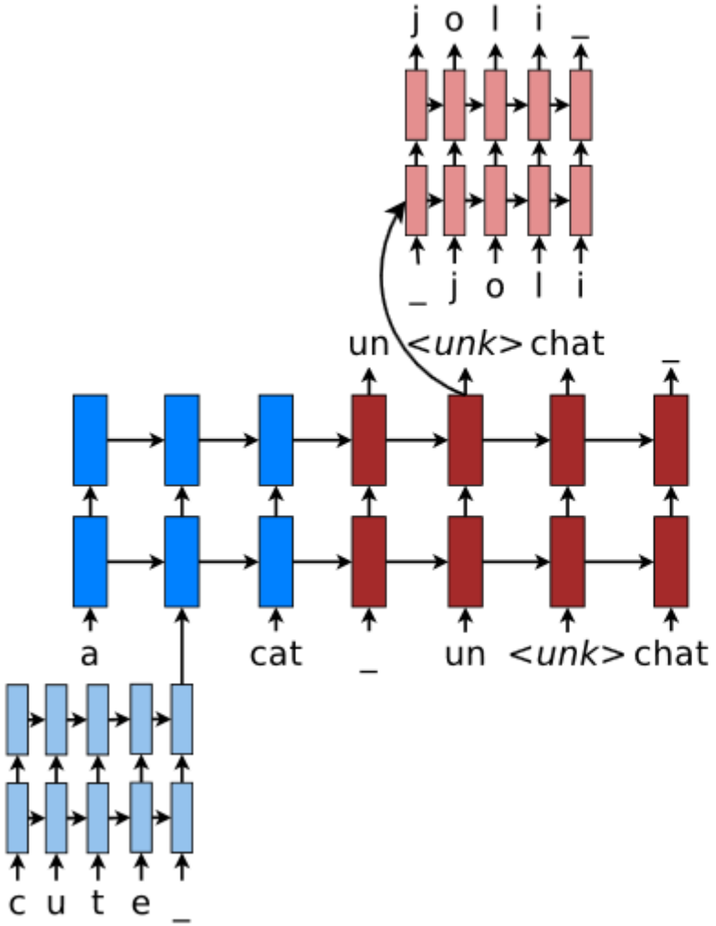


Figure 1: **Hybrid NMT** – example of a word-character model for translating "a cute cat" into "un joli chat". Hybrid NMT translates at the word level. For rare tokens, the character-level components build source representations and recover target <*unk*>. "_" marks sequence boundaries.

Thang Luong and Chris Manning. Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. ACL 2016. // https://arxiv.org/abs/1604.00788

Гибридный подход: Hybrid NMT

метод луча (beam search)
на уровне слов и букв

SOTA по BLEU на момент публикации

Гибридный подход: Hybrid NMT

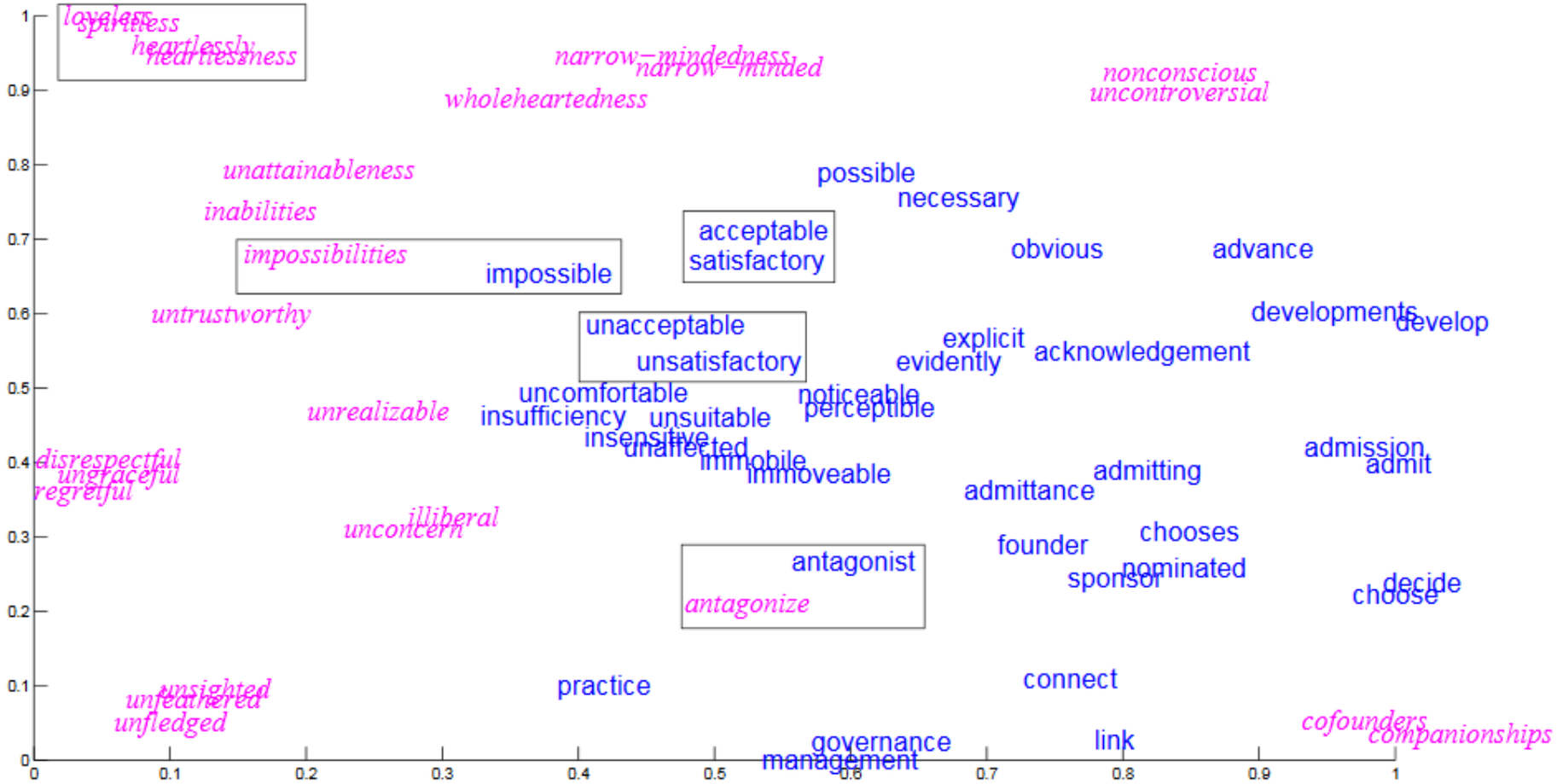


Figure 4: **Barnes-Hut-SNE visualization of source word representations** – shown are sample words from the *Rare Word* dataset. We differentiate two types of embeddings: frequent words in which encoder embeddings are looked up directly and *rare* words where we build representations from characters. Boxes highlight examples that we will discuss in the text. We use the hybrid model *(l)* in this visualization.

Другие подходы…

«Chars for word embeddings»

не будем рассказывать…

Kris Cao and Marek Rei «A Joint Model for Word Embedding and Word Morphology» //
https://arxiv.org/pdf/1606.02601.pdf

Следующая тема

~~Представления слов~~

~~Токенизация на подслова~~

~~Посимвольный подход~~

~~Гибридный подход~~

С(у/а)ммаризация текстов ← сейчас это

Рассказ историй: Storytelling

Рассказ историй по тексту: Hierarchical Neural Story Generation

Генерация поэзии: Hafez, Deep-speare

Coreference Resolution: SOTA / Clustering-Based

С(а/у)ммаризация текстов

текст или набор текстов → краткое содержание

Extractive summarization

Выбрать части данного текста, чтобы составить саммари

легче

Abstractive summarization

Сгенерировать новый текст

более ИИ

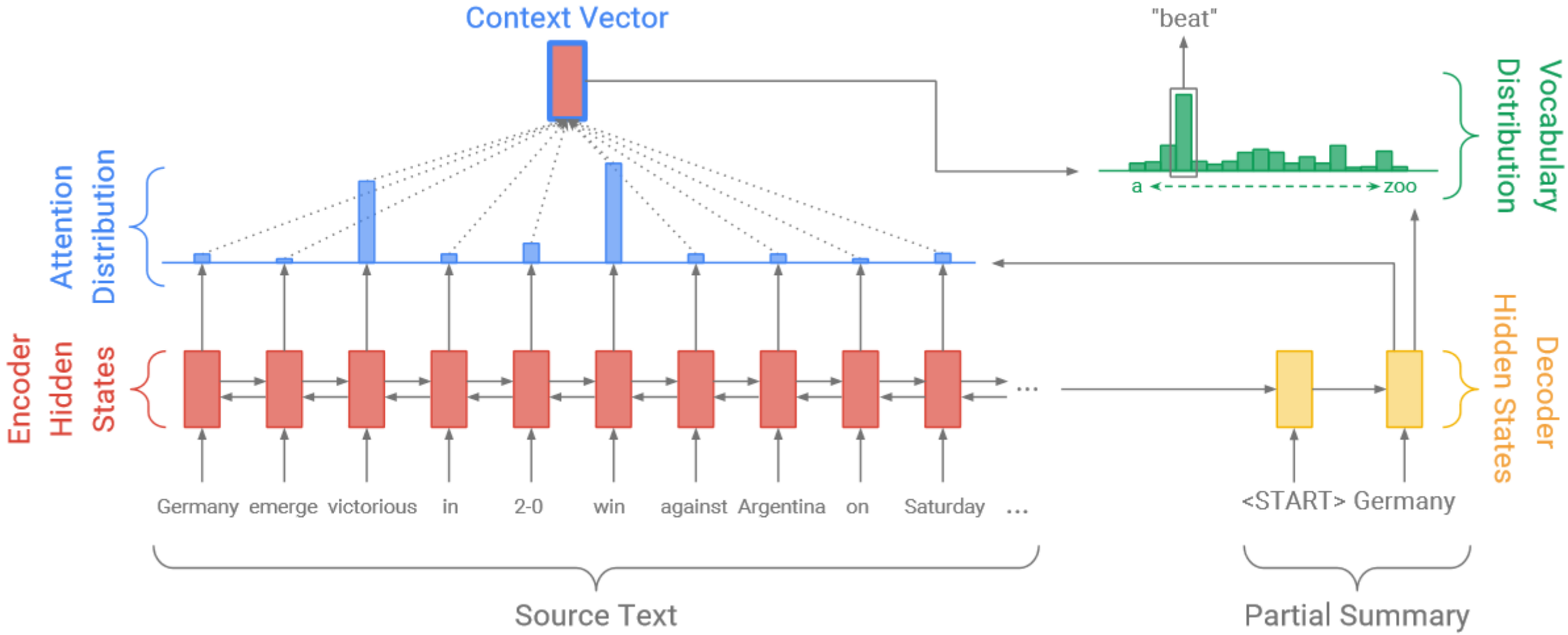
«simplification»
– упрощение текста

Богатый источник информации и данных

https://github.com/mathsyouth/awesome-text-summarization

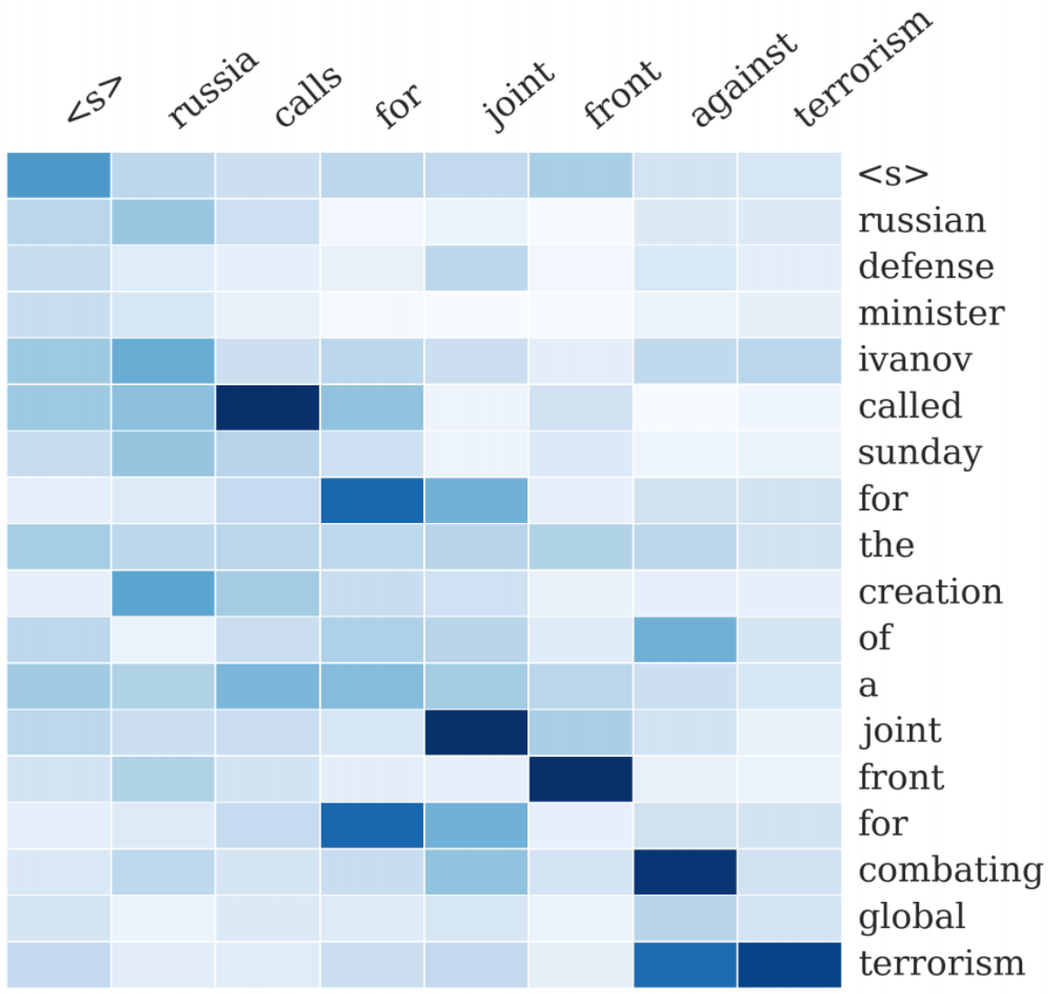
С(a/y)ммаризация текстов: seq2seq-подход / + attention

суммаризация – задача «перевода»



из работы https://arxiv.org/pdf/1704.04368.pdf

С(а/у)ммаризация текстов: seq2seq-подход / + attention



Rush et al «A Neural Attention Model for Abstractive Sentence Summarization» 2015 // https://arxiv.org/pdf/1509.00685.pdf

С(а/у)ммаризация текстов

много улучшений:

иерархическое внимание (Hierarchical / multi-level attention)

More global / high-level content selection

RL для минимизации метрики ROUGE

A Survey on Neural Network-Based Summarization Methods, Dong, 2018 https://arxiv.org/pdf/1804.04589.pdf

нужен механизм копирования – для точности к деталям

Language as a Latent Variable: Discrete Generative Models for Sentence Compression, Miao et al, 2016 https://arxiv.org/pdf/1609.07317.pdf

Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond, Nallapati et al, 2016 https://arxiv.org/pdf/1602.06023.pdf

Incorporating Copying Mechanism in Sequence-to-Sequence Learning, Gu et al, 2016 https://arxiv.org/pdf/1603.06393.pdf

С(а/у)ммаризация текстов: Pointer-Generator Networks

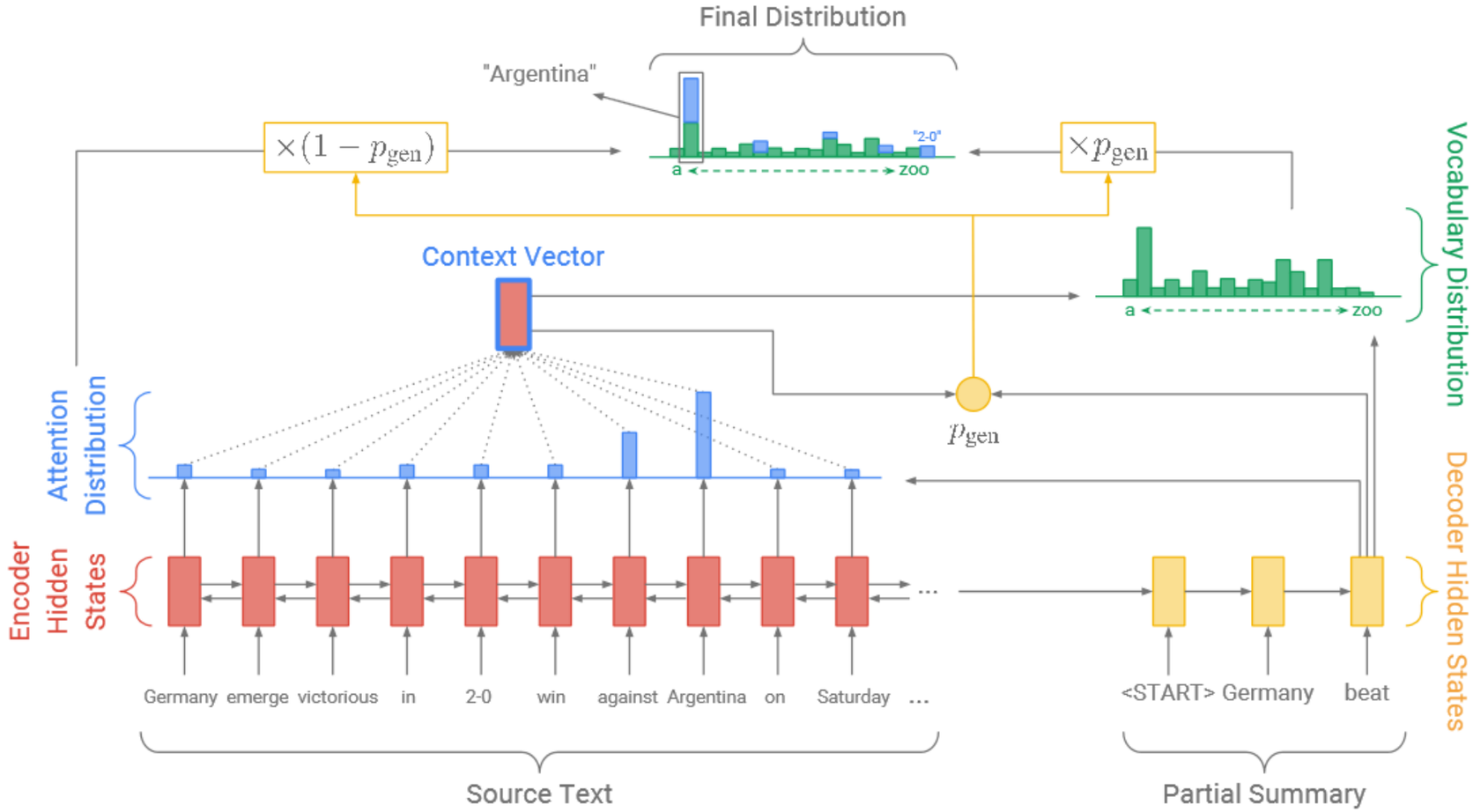


Figure 3: Pointer-generator model. For each decoder timestep a generation probability $p_{\text{gen}} \in [0, 1]$ is calculated, which weights the probability of *generating* words from the vocabulary, versus *copying* words from the source text. The vocabulary distribution and the attention distribution are weighted and summed to obtain the final distribution, from which we make our prediction. Note that out-of-vocabulary article words such as *2-0* are included in the final distribution. Best viewed in color.

С(а/у)ммаризация текстов: Pointer-Generator Networks

1) hybrid pointer-generator network – для копирования слов с помощью «указания» для аккуратного копирования информации

гибридность означает, что можно копировать и генерировать слова…
на каждом шаге t оценивать вероятность генерирования слова, а не копирования

$$p = \sigma(w_1^{\mathrm{T}} h_t^* + w_2^{\mathrm{T}} s_t + w_3^{\mathrm{T}} x_t + b)$$

s – состояние декодера

x – вход декодера

h – контекстный вектор (сумма состояний кодировщика):

$$h_t^* = \sum_i a_{[i]}^t h_i$$

$$P(w) = pP_{\text{vocab}}(w) + (1-p) \sum_{i:w_i=w} a_{[i]}^t$$

левая часть – генерация из словаря, правая – копирование

С(а/у)ммаризация текстов: Pointer-Generator Networks

2) Использование «покрытия» (coverage из [Tu et al., 2016]) того, что попало в суммаризацию, чтобы исключить повторения

покрытие – сумма attention-векторов

$$c^t = \sum_{t'<t} a^{t'}$$

и оно учитывается при вычислении внимания:

$$e_i^t = w^{\mathrm{T}} \tanh(W_1 h_i + W_2 s_i + w_0 c_{[i]}^t + b)$$

и ещё «coverage loss»

$$\mathrm{covloss}_t = \sum_i \min(a_{[i]}^t, c_{[i]}^t)$$

See et al, 2017, Get To The Point: Summarization with Pointer-Generator Networks, https://arxiv.org/pdf/1704.04368.pdf

Original Text (truncated): lagos, nigeria (cnn) a day after winning nigeria's presidency, *muhammadu buhari* told cnn's christiane amanpour that **he plans to aggressively fight corruption that has long plagued nigeria** and go after the root of the nation's unrest. *buhari* said he'll "rapidly give attention" to curbing violence in the northeast part of nigeria, where the terrorist group boko haram operates. by cooperating with neighboring nations chad, cameroon and niger, **he said his administration is confident it will be able to thwart criminals** and others contributing to nigeria's instability. for the first time in nigeria's history, the opposition defeated the ruling party in democratic elections. *buhari* defeated incumbent goodluck jonathan by about 2 million votes, according to nigeria's independent national electoral commission. **the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.**

Baseline Seq2Seq + Attention: **UNK UNK** says his administration is confident it will be able to **destabilize nigeria's economy**. **UNK** says his administration is confident it will be able to thwart criminals and other **nigerians**. **he says the country has long nigeria and nigeria's economy.**

Pointer-Gen: *muhammadu buhari* says he plans to aggressively fight corruption **in the northeast part of nigeria**. he says he'll "rapidly give attention" to curbing violence **in the northeast part of nigeria**. he says his administration is confident it will be able to thwart criminals.

Pointer-Gen + Coverage: *muhammadu buhari* says he plans to aggressively fight corruption that has long plagued nigeria. he says his administration is confident it will be able to thwart criminals. the win comes after a long history of military rule, coups and botched attempts at democracy in africa's most populous nation.

Figure 1: Comparison of output of 3 abstractive summarization models on a news article. The baseline model makes **factual errors**, a **nonsensical sentence** and struggles with OOV words *muhammadu buhari*. The pointer-generator model is accurate but **repeats itself**. Coverage eliminates repetition. The final summary is composed from **several fragments**.

C(a/y)ммаризация текстов: Pointer-Generator Networks

	ROUGE			METEOR	
	1	2	L	exact match	+ stem/syn/para
abstractive model (Nallapati et al., 2016)*	35.46	13.30	32.65	-	-
seq-to-seq + attn baseline (150k vocab)	30.49	11.17	28.08	11.65	12.86
seq-to-seq + attn baseline (50k vocab)	31.33	11.81	28.83	12.03	13.20
pointer-generator	36.44	15.66	33.42	15.35	16.65
pointer-generator + coverage	**39.53**	**17.28**	**36.38**	17.32	18.72
lead-3 baseline (ours)	40.34	17.70	36.57	20.48	22.21
lead-3 baseline (Nallapati et al., 2017)*	39.2	15.7	35.5	-	-
extractive model (Nallapati et al., 2017)*	39.6	16.2	35.3	-	-

Table 1: ROUGE F_1 and METEOR scores on the test set. Models and baselines in the top half are abstractive, while those in the bottom half are extractive. Those marked with * were trained and evaluated on the anonymized dataset, and so are not strictly comparable to our results on the original text. All our ROUGE scores have a 95% confidence interval of at most ± 0.25 as reported by the official ROUGE script. The METEOR improvement from the 50k baseline to the pointer-generator model, and from the pointer-generator to the pointer-generator+coverage model, were both found to be statistically significant using an approximate randomization test with $p < 0.01$.

Суммаризация: Проблемы с копированием

– копируют слишком много

– плохи в общем выборе контента

делают несколько стадий:

выбор контента

генерация текста

можно – суммаризация «снизу-вверх»

С(а/у)ммаризация текстов: Bottom-up summarization

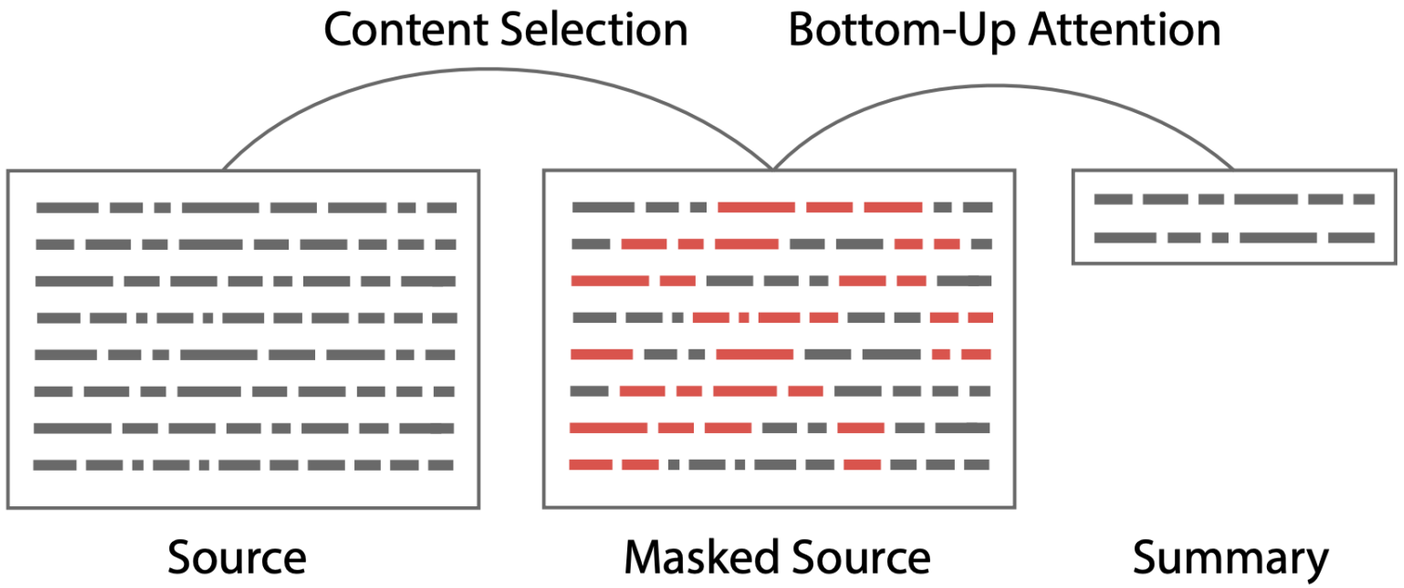


Figure 2: Overview of the selection and generation processes described throughout Section 4.

Content selection stage

– пометка слов тэгами «include» / «don't»

Bottom-up attention stage

Pointer-Generator

не посещаем слова с тегами «don't»

просто и эффективно!

Gehrmann et al «Bottom-Up Abstractive Summarization», 2018

https://arxiv.org/pdf/1808.10792v1.pdf

С(а/у)ммаризация текстов: Bottom-up summarization

Method	R-1	R-2	R-L
Pointer-Generator (See et al., 2017)	36.44	15.66	33.42
Pointer-Generator + Coverage (See et al., 2017)	39.53	17.28	36.38
ML + Intra-Attention (Paulus et al., 2017)	38.30	14.81	35.49
ML + RL (Paulus et al., 2017)	39.87	15.82	36.90
Saliency + Entailment reward (Pasunuru and Bansal, 2018)	40.43	18.00	37.10
Key information guide network (Li et al., 2018a)	38.95	17.12	35.68
Inconsistency loss (Hsu et al., 2018)	40.68	17.97	37.13
Sentence Rewriting (Chen and Bansal, 2018)	40.88	17.80	**38.54**
Pointer-Generator (our implementation)	36.25	16.17	33.41
Pointer-Generator + Coverage Penalty	39.12	17.35	36.12
Pointer-Generator + Mask Only	37.70	15.63	35.49
Pointer-Generator + Multi-Task	37.67	15.59	35.47
Pointer-Generator + DiffMask	38.45	16.88	35.81
Bottom-Up Summarization	**41.22**	**18.68**	38.34

Table 1: Results of abstractive summarizers on the CNN-DM dataset.[2] The first section shows encoder-decoder abstractive baselines trained with cross-entropy. The second section describes reinforcement-learning based approaches. The third section presents our baselines and the attention masking methods described in this work.

С(а/у)ммаризация текстов: + RL нужно знание RL – пропускаем

оптимизация ROUDE-L с помощью RL – функция не дифференцируемая

но получается не очень «человечный» текст

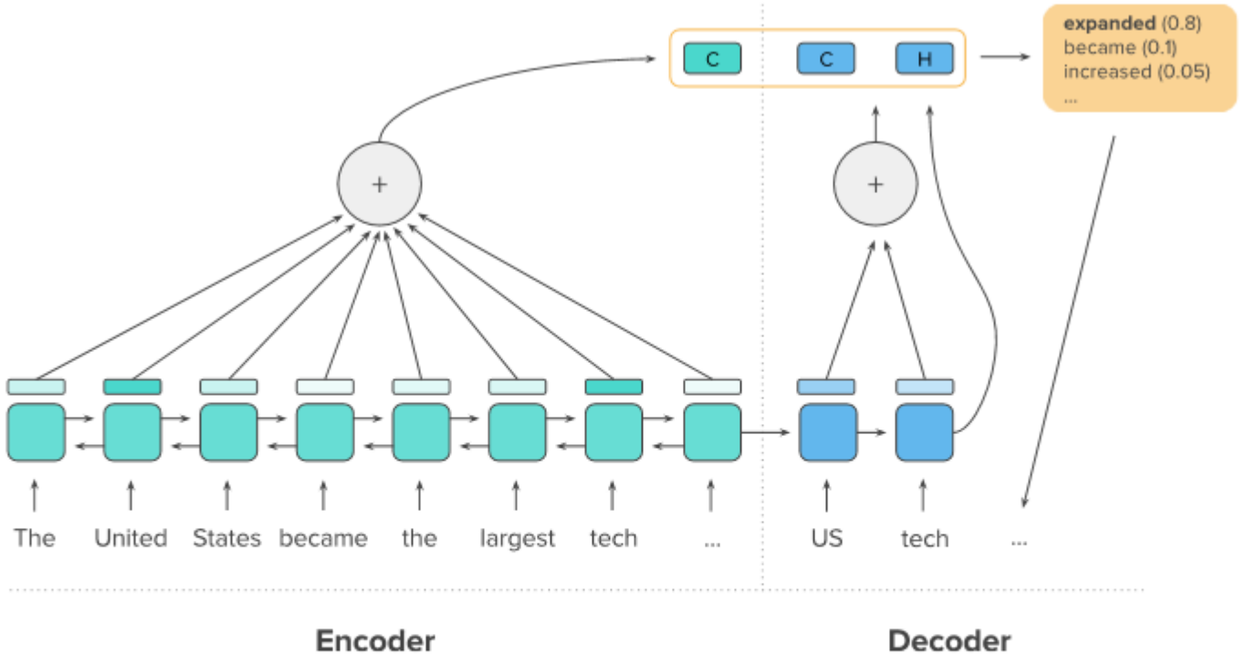


Figure 1: Illustration of the encoder and decoder attention functions combined. The two context vectors (marked "C") are computed from attending over the encoder hidden states and decoder hidden states. Using these two contexts and the current decoder hidden state ("H"), a new word is generated and added to the output sequence.

Paulus et al «A Deep Reinforced Model for Abstractive Summarization», 2017 //

https://arxiv.org/pdf/1705.04304.pdf

simplification: DRESS (Deep REinforcement Sentence Simplification)

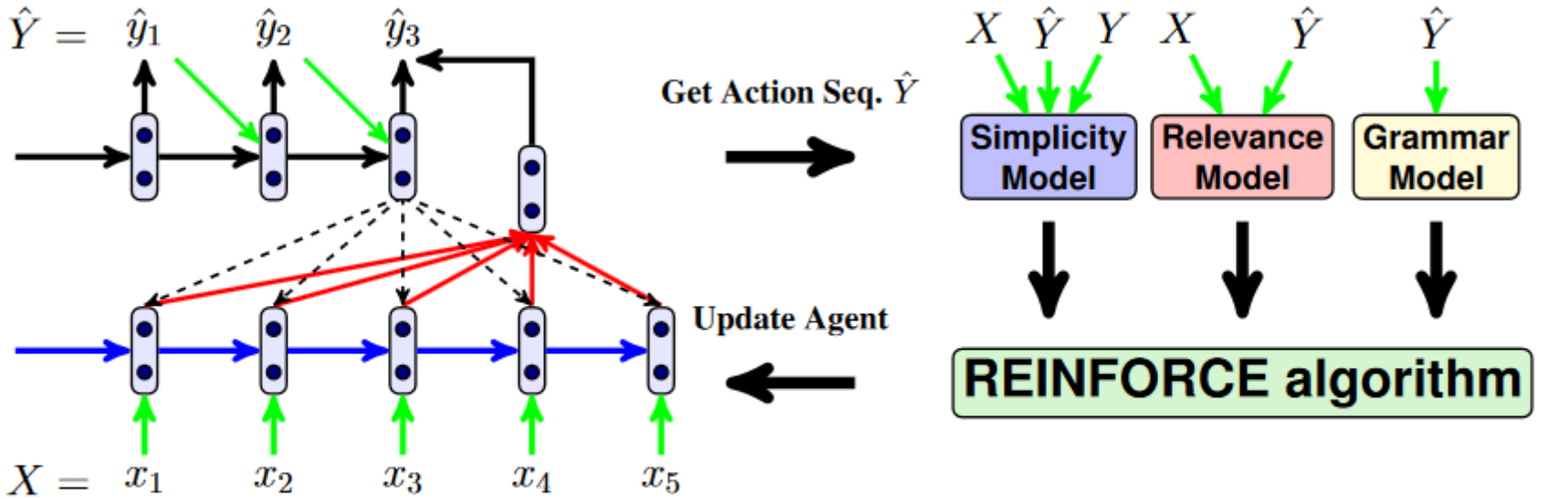


Figure 1: Deep reinforcement learning simplification model. X is the complex sentence, Y the reference (simple) sentence and $\hat{Y}$ the action sequence (simplification) produced by the encoder-decoder model.

Действия агента – упрощение предложения

Награда – по метрике похожесть на целевое упрощение + простота + соответствие языковой модели

Xingxing Zhangand, Mirella Lapata «Sentence Simplification with Deep Reinforcement Learning» https://www.aclweb.org/anthology/D17-1062.pdf

Complex	There's just one major hitch: the primary purpose of education is to develop citizens with a wide variety of skills.
Reference	The purpose of education is to develop a wide range of skills.
PBMT-R	It's just one major hitch: the purpose of education is to **make people** with a wide variety of skills.
Hybrid	one hitch the purpose is to develop citizens.
EncDecA	The **key** of education is to develop **people** with a wide variety of skills.
DRESS	There's just one major hitch: the **main goal** of education is to develop **people** with **lots of** skills.
DRESS-LS	There's just one major hitch: the **main goal** of education is to develop citizens with **lots of** skills.
Complex	"They were so burdened by the past they couldn't think about the future," said Barnet, 62, who was president of Columbia Records, the No.1 record label in the United States, before joining Capitol.
Reference	Capitol was stuck in the past. It could not think about the future, Barnett said.
PBMT-R	"They were so **affected** by the past they couldn't think about the future," said Barnett, 62, was president of Columbia Records, before joining Capitol **building**.
Hybrid	"They were so burdened by the past they couldn't think about the future," said Barnett, 62, who was Columbia Records, president of the No.1 record label in the united states, before joining Capitol.
EncDecA	"They were so burdened by the past they couldn't think about the future," said Barnett, who was president of Columbia Records, the No.1 record labels in the United States.
DRESS	"They were so **sicker** by the past they couldn't think about the future," said Barnett, who was president of Columbia Records.
DRESS-LS	"They were so burdened by the past they couldn't think about the future," said Barnett, who was president of Columbia Records.

Table 3: System output for two sentences (Newsela development set). Substitutions are shown in bold.

Extractive summarization: SummaRuNNer

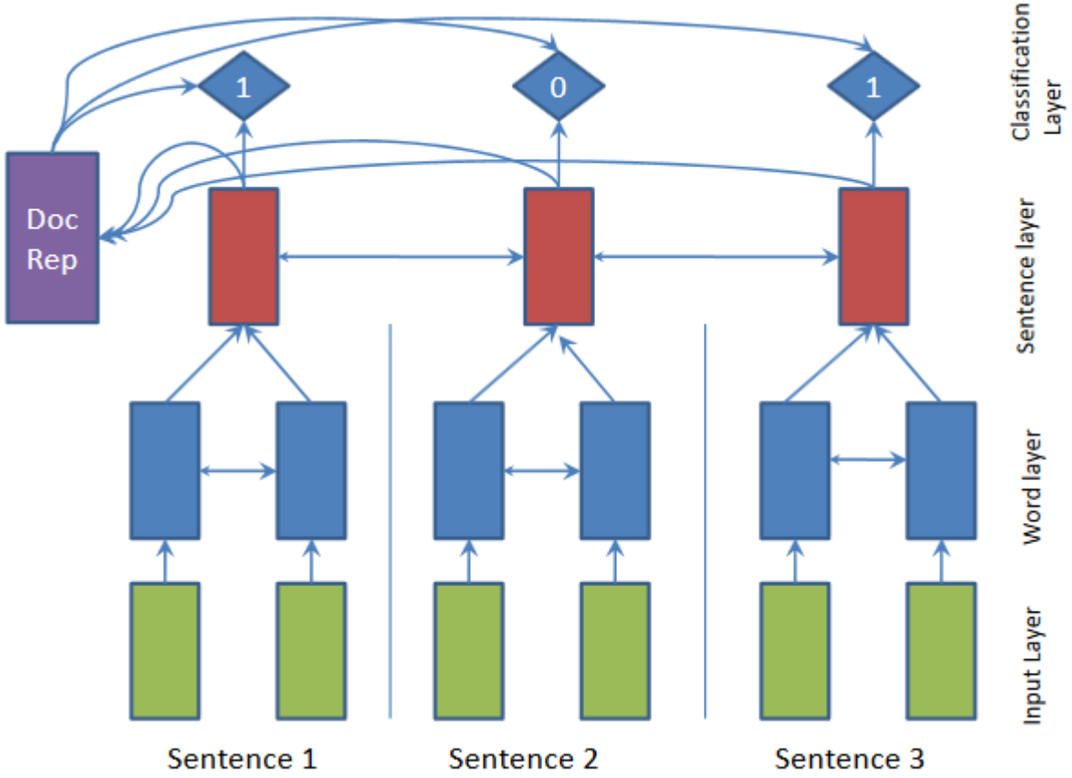


Figure 1: SummaRuNNer: A two-layer RNN based sequence classifier: the bottom layer operates at word level within each sentence, while the top layer runs over sentences. Double-pointed arrows indicate a bi-directional RNN. The top layer with 1's and 0's is the sigmoid activation based classification layer that decides whether or not each sentence belongs to the summary. The decision at each sentence depends on the content richness of the sentence, its salience with respect to the document, its novelty with respect to the accumulated summary representation and other positional features.

один из первых RNN-подходов

Ramesh Nallapati et. al. «SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents» https://arxiv.org/pdf/1611.04230.pdf

SummaRuNNer: интерпретация

Gold Summary: Redpath has ended his eight-year association with Sale Sharks. Redpath spent five years as a player and three as a coach at sale. He has thanked the owners, coaches and players for their support.	Salience	Content	Novelty	Position	Prob.
Bryan Redpath has left his coaching role at Sale Sharks with immediate effect.	0.1	0.1	0.9	0.1	0.3
The 43 - year - old Scot ends an eight-year association with the Aviva Premiership side, having spent five years with them as a player and three as a coach.	0.9	0.6	0.9	0.9	0.7
Redpath returned to Sale in June 2012 as director of rugby after starting a coaching career at Gloucester and progressing to the top job at Kingsholm .	0.8	0.5	0.5	0.9	0.6
Redpath spent five years with Sale Sharks as a player and a further three as a coach but with Sale Sharks struggling four months into Redpath's tenure, he was removed from the director of rugby role at the Salford-based side and has since been operating as head coach .	0.8	0.9	0.7	0.8	**0.9**
'I would like to thank the owners, coaches, players and staff for all their help and support since I returned to the club in 2012.	0.4	0.1	0.1	0.7	0.2
Also to the supporters who have been great with me both as a player and as a coach,' Redpath said.	0.6	0.0	0.2	0.3	0.2

Figure 2: Visualization of SummaRuNNer output on a representative document. Each row is a sentence in the document, while the shading-color intensity is proportional to its probability of being in the summary, as estimated by the RNN-based sequence classifier. In the columns are the normalized scores from each of the abstract features in Eqn. (6) as well as the final prediction probability (last column). Sentence 2 is estimated to be the most salient, while the longest one, sentence 4, is considered the most content-rich, and not surprisingly, the first sentence the most novel. The third sentence gets the best position based score.

Abstractive Summarization: TCONVS2S

«Extreme summarization»

Сгенерировать по тексту одно предложение – ответ на вопрос «о чём текст»

Идея архитектуры (свёрточная) взята из machine translation (Gehring et al., 2017a,b) будем проходить story generation (Fan et al.,2018)

Shashi Narayan, Shay B Cohen, and Mirella Lapata «Don't give me the details, just thesummary! Topic-aware convolutional neural net-works for extreme summarization» arXiv preprintarXiv:1808.08745.

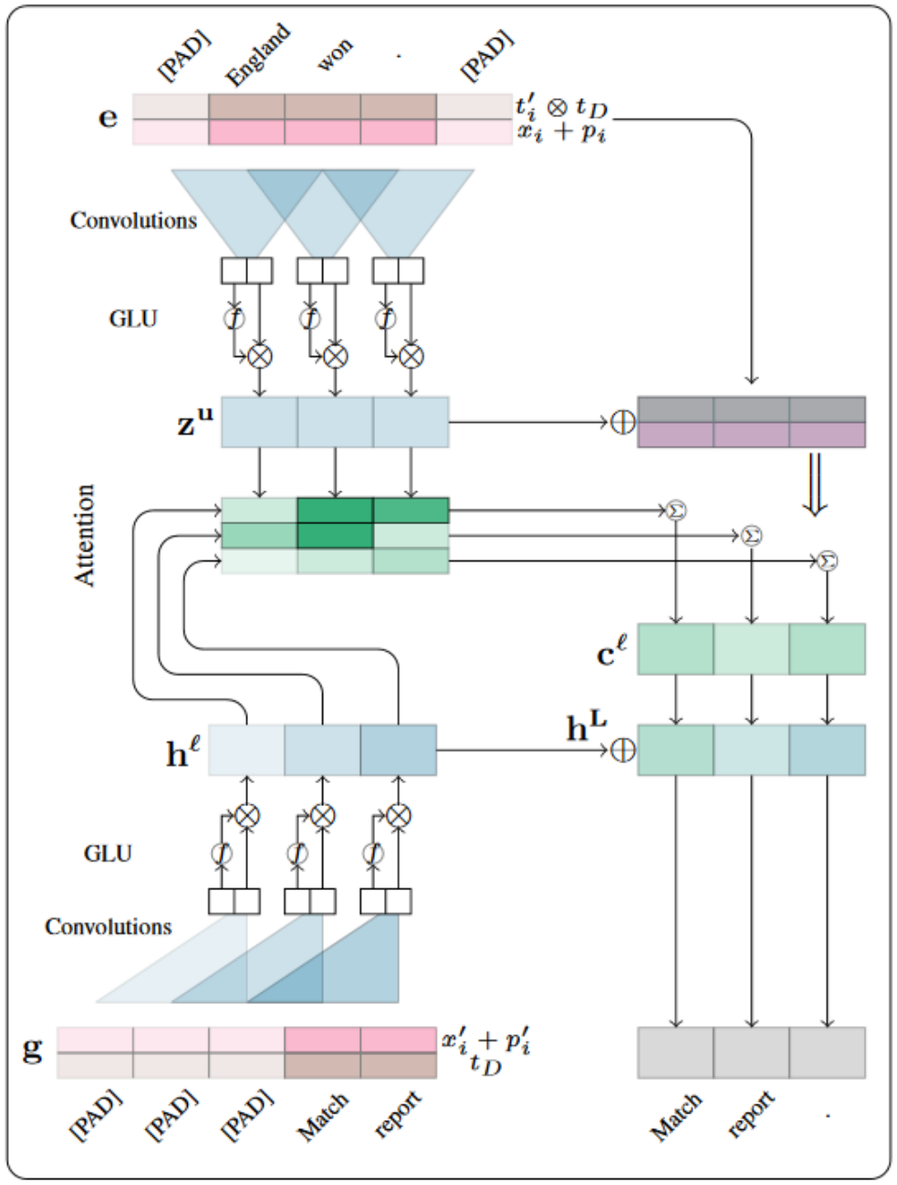


Figure 2: Topic-conditioned convolutional model for extreme summarization.

SUMMARY: *A man and a child have been killed after a light aircraft made an emergency landing on a beach in Portugal.*

DOCUMENT: Authorities said the incident took place on Sao Joao beach in Caparica, south-west of Lisbon.
The National Maritime Authority said a middle-aged man and a young girl died after they were unable to avoid the plane.
[*6 sentences with 139 words are abbreviated from here.*]
Other reports said the victims had been sunbathing when the plane made its emergency landing.
[*Another 4 sentences with 67 words are abbreviated from here.*]
Video footage from the scene carried by local broadcasters showed a small recreational plane parked on the sand, apparently intact and surrounded by beachgoers and emergency workers.
[*Last 2 sentences with 19 words are abbreviated.*]

Figure 1: An abridged example from our extreme summarization dataset showing the document and its one-line summary. Document content present in the summary is color-coded.

Суммаризация с BERT: BertSum

общий подход для extractive / abstractive суммаризации

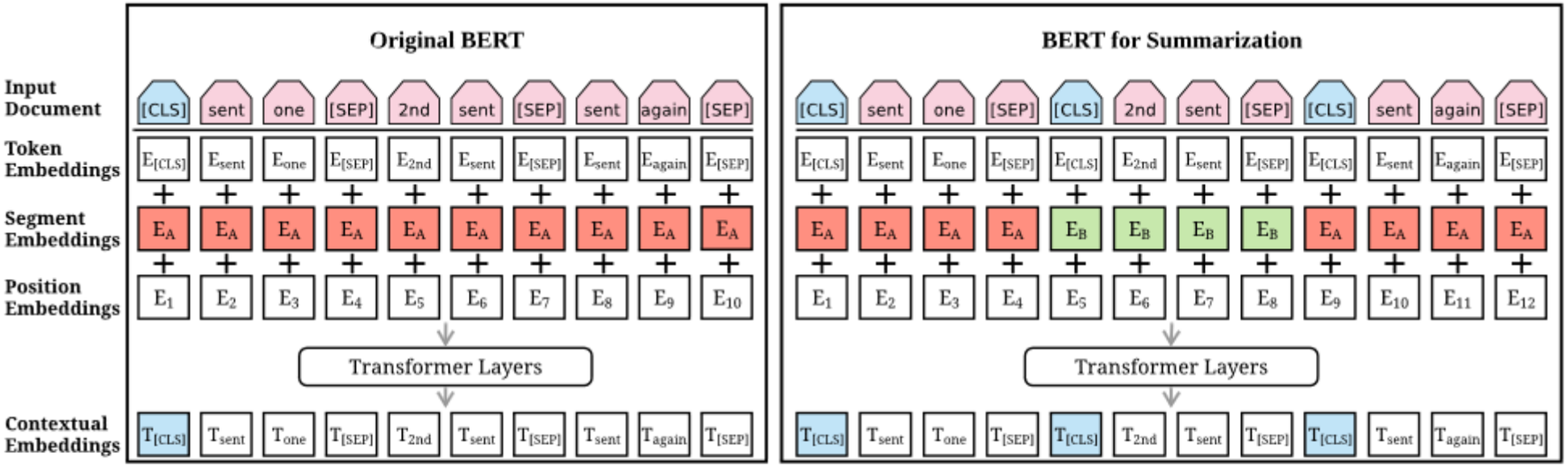


Figure 1: Architecture of the original BERT model (left) and BERTSUM (right). The sequence on top is the input document, followed by the summation of three kinds of embeddings for each token. The summed vectors are used as input embeddings to several bidirectional Transformer layers, generating contextual vectors for each token. BERTSUM extends BERT by inserting multiple [CLS] symbols to learn sentence representations and using interval segmentation embeddings (illustrated in red and green color) to distinguish multiple sentences.

на вход текст, поэтому [CLS] разделяет предложения и потом собирает информацию о след. предложении, специальная кодировка чётности предложения

Yang Liuand, Mirella Lapata «Text Summarization with Pretrained Encoders» // https://arxiv.org/pdf/1908.08345.pdf

Суммаризация с BERT: BertSum

Extractive Summarization

целевые пометки – оставляем или нет предложение

Abstractive Summarization

кодировщик-декодировщик

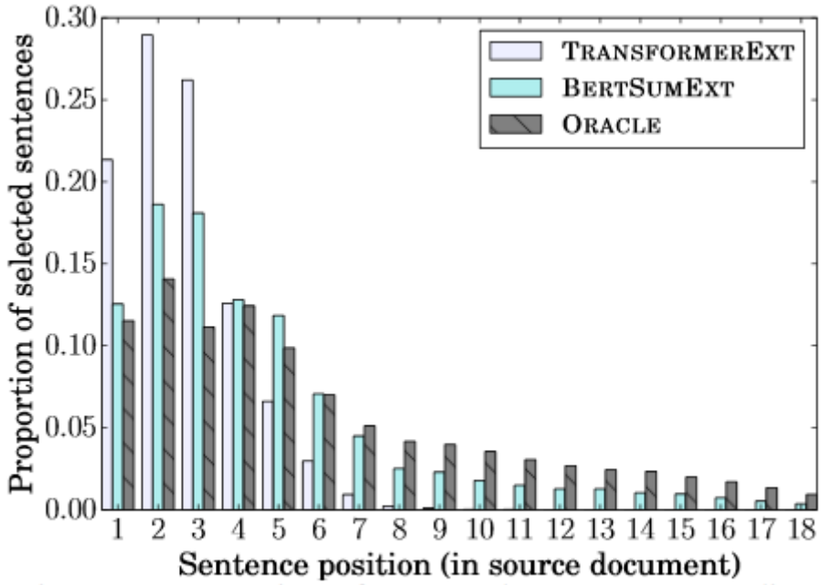


Figure 2: Proportion of extracted sentences according to their position in the original document.

Model	R1	R2	RL
Oracle	52.59	31.24	48.87
Lead-3	40.42	17.62	36.67
Extractive			
SummaRuNNer (Nallapati et al., 2017)	39.60	16.20	35.30
Refresh (Narayan et al., 2018b)	40.00	18.20	36.60
Latent (Zhang et al., 2018)	41.05	18.77	37.54
NeuSum (Zhou et al., 2018)	41.59	19.01	37.98
Sumo (Liu et al., 2019)	41.00	18.40	37.20
TransformerExt	40.90	18.02	37.17
Abstractive			
PTGen (See et al., 2017)	36.44	15.66	33.42
PTGen+Cov (See et al., 2017)	39.53	17.28	36.38
DRM (Paulus et al., 2018)	39.87	15.82	36.90
BottomUp (Gehrmann et al., 2018)	41.22	18.68	38.34
DCA (Celikyilmaz et al., 2018)	41.69	19.47	37.92
TransformerAbs	40.21	17.76	37.09
Bert-based			
BertSumExt	43.25	20.24	39.63
BertSumExt w/o interval embeddings	43.20	20.22	39.59
BertSumExt (large)	43.85	20.34	39.90
BertSumAbs	41.72	19.39	38.76
BertSumExtAbs	42.13	19.60	39.18

Table 2: ROUGE F1 results on **CNN/DailyMail** test set (R1 and R2 are shorthands for unigram and bigram overlap; RL is the longest common subsequence). Results for comparison systems are taken from the authors' respective papers or obtained on our data by running publicly released software.

Диалоги

Целевые (Task-oriented dialogue)	**Ассистенты (Assistive)** Помощь в действиях (покупка, бронирование и т.п.) **Кооперация (Co-operative)** Есть совместная задача, например, диалог **Состязательный (Adversarial)** Противостояние в диалоге
Социальные (Social dialogue)	**Chit-chat** диалоги для развлечения **Therapy / mental wellbeing** терапия

Диалоги

seq2seq

- A Neural Conversational Model, Vinyals et al, 2015
https://arxiv.org/pdf/1506.05869.pdf
- Neural Responding Machine for Short-Text Conversation, Shang et al, 2015
https://www.aclweb.org/anthology/P15-1152

– повторения

– нерелевантные ответы

– общие ответы («я не знаю»)

– потеря контекста

Диалоги: борьба с нерелевантностью

Maximum Mutual Information (MMI) between input S and response T

A Diversity-Promoting Objective Function for Neural Conversation Models, Li et al, 2016

https://arxiv.org/pdf/1510.03055.pdf

повышать вероятности редких слов

другие стратегии сэмплирования (не только beam search)

Why are Sequence-to-Sequence Models So Dull?, Jiang et al, 2018

https://staff.fnwi.uva.nl/m.derijke/wp-content/papercite-data/pdf/jiang-why-2018.pdf

подстройка под собеседника

например, векторное представление собеседника

A Persona-Based Neural Conversation Model, Li et al 2016, https://arxiv.org/pdf/1603.06155.pdf

Personalizing Dialogue Agents: I have a dog, do you have pets too?, Zhang et al, 2018

https://arxiv.org/pdf/1801.07243.pdf

Диалоги: борьба с повторами

запрещать повторы

простое решение

можно вставить в функцию ошибки (не всегда дифференцируема $\Rightarrow$ RL)

можно контролировать механизм внимания

не смотрел на одни и те же слова много раз

Рассказ историй: Storytelling

по контексту (изображение, запрос, начало рассказа) сгенерировать текст

Generated story about image
Model: Taylor Swift Lyrics

"I don't see the expression on my face, you know, that's what I want to do, I guess, if you're a cat, I bear it away."

https://medium.com/@samim/generating-stories-about-images-d163ba41e4ed

Рассказ историй: Storytelling

Generated story about image
Model: Taylor Swift Lyrics

"You are in a crowd of people, I thought, I know what it 's like to be honest with you. I thought , Oh, God , if you come up with a fire in it, I 'm gonna lose you."

нет парной разметки для обучения
надо использовать общее пространство представлений
(common sentence-encoding space)

Рассказ историй: Storytelling

Соединяем этапы:

используем метод Skip-thought vectors для представления предложений

используя COCO (изображения с заголовками) учим отображение картинки → тексты

учим RNN генерировать текст по Skip-thought vector (датасет Taylor Swift lyrics)

Рассказ историй по тексту: Hierarchical Neural Story Generation

Prompt: The Mage, the Warrior, and the Priest

Story: A light breeze swept the ground, and carried with it still the distant scents of dust and time-worn stone. The Warrior led the way, heaving her mass of armour and muscle over the uneven terrain. She soon crested the last of the low embankments, which still bore the unmistakable fingerprints of haste and fear. She lifted herself up onto the top the rise, and looked out at the scene before her. [...]

Figure 1: Example prompt and beginning of a story from our dataset. We train a hierarchical model that first generates a prompt, and then conditions on the prompt when generating a story.

«очень нетривиальное решение»

Fan et al «Hierarchical Neural Story Generation», 2018 https://arxiv.org/pdf/1805.04833.pdf

Рассказ историй по тексту: Hierarchical Neural Story Generation

seq2seq – но качество не очень – игнорирую контекст, больше моделируют текст

1) Model fusion mechanism из [Sriram et al., 2017]

учим seq2seq (общая LM),

потом вторую (учитывает запрос), которая имеет доступ к скрытым состояниям первой

~ как бустинг

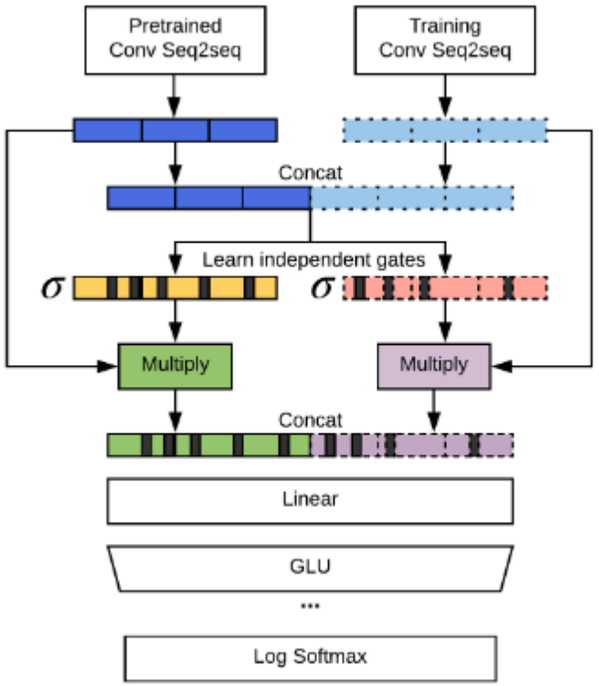


Figure 4: Diagram of our fusion model, which learns a second seq2seq model to improve a pre-trained model. The separate hidden states are combined after gating through concatenation.

Рассказ историй по тексту: Hierarchical Neural Story Generation

2) Gated multi-head multi-scale self-attention

см. дальше Q,K,V не линейные проекции, а сложнее –
Gated Linear Unit activations [Dauphin et al., 2017]

downsample – первая головка видит всё, вторая – чётные позиции,
третья – кратные 3м и т.д.

Рассказ историй по тексту: Hierarchical Neural Story Generation

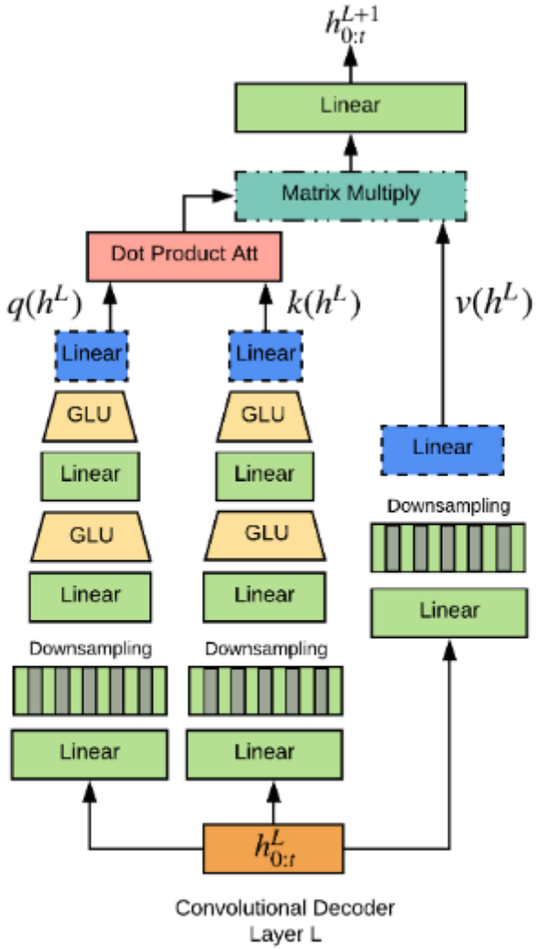


Figure 2: Self-Attention Mechanism of a single head, with GLU gating and downsampling. Multiple heads are concatenated, with each head using a separate downsampling function.

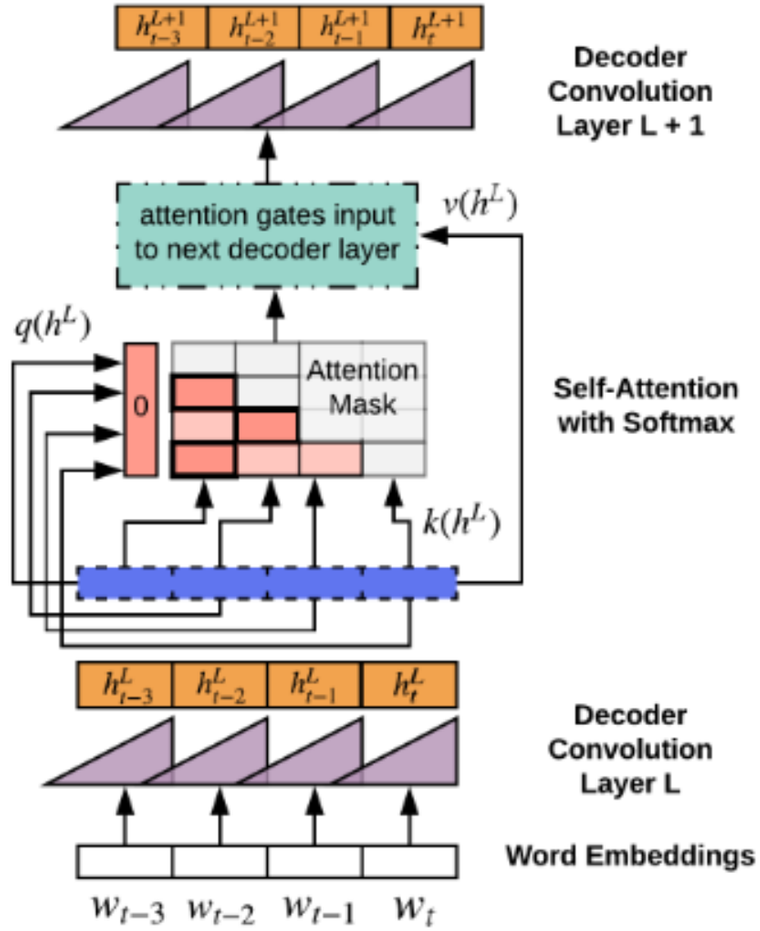


Figure 3: Multihead self-attention mechanism. The decoder layer depicted attends with itself to gate the input of the subsequent decoder layer.

Рассказ историй по тексту: Hierarchical Neural Story Generation

Model	# Parameters (mil)	Valid Perplexity	Test Perplexity
GCNN LM	123.4	54.50	54.79
GCNN + self-attention LM	126.4	51.84	51.18
LSTM seq2seq	110.3	46.83	46.79
Conv seq2seq	113.0	45.27	45.54
Conv seq2seq + self-attention	134.7	37.37	37.94
Ensemble: Conv seq2seq + self-attention	270.3	36.63	36.93
Fusion: Conv seq2seq + self-attention	255.4	**36.08**	**36.56**

Table 3: Perplexity on WRITINGPROMPTS. We dramatically improve over standard seq2seq models.

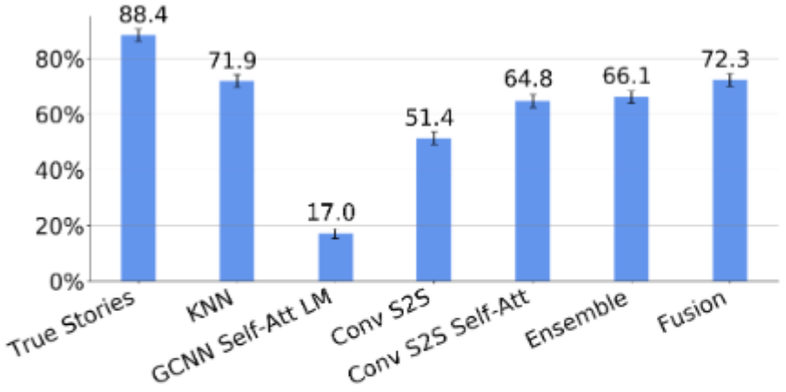


Figure 5: Human accuracy at pairing stories with the prompts used to generate them. People find that our fusion model significantly improves the link between the prompt and generated stories.

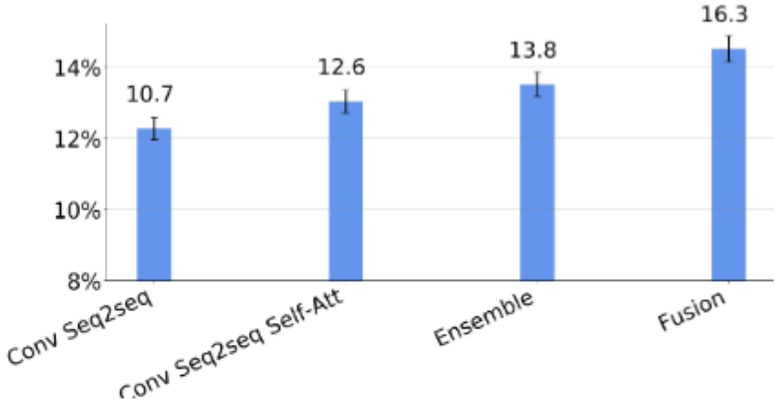


Figure 6: Accuracy of prompt ranking. The fusion model most accurately pairs prompt and stories.

Рассказ историй по тексту: глобальная проблема

LM ~ последовательность слов, а история ~ последовательность событий

- события и их структура
- логика событий (наследство – конфликт – преступление)
- персонажи (их мотивация, описание, предыстория)
- атмосфера (реальный мир)
- Принципы правильного рассказа (ex: ружьё в театре)

https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17046/15769
https://homes.cs.washington.edu/~yejin/

Генерация поэзии

Finite State Acceptor (FSA) – генерация всех возможных последовательностей, которые удовлетворяют ритму

one line of iambic pentameter →

```
010     1  0    10      101
Attending on his golden pilgramage
```

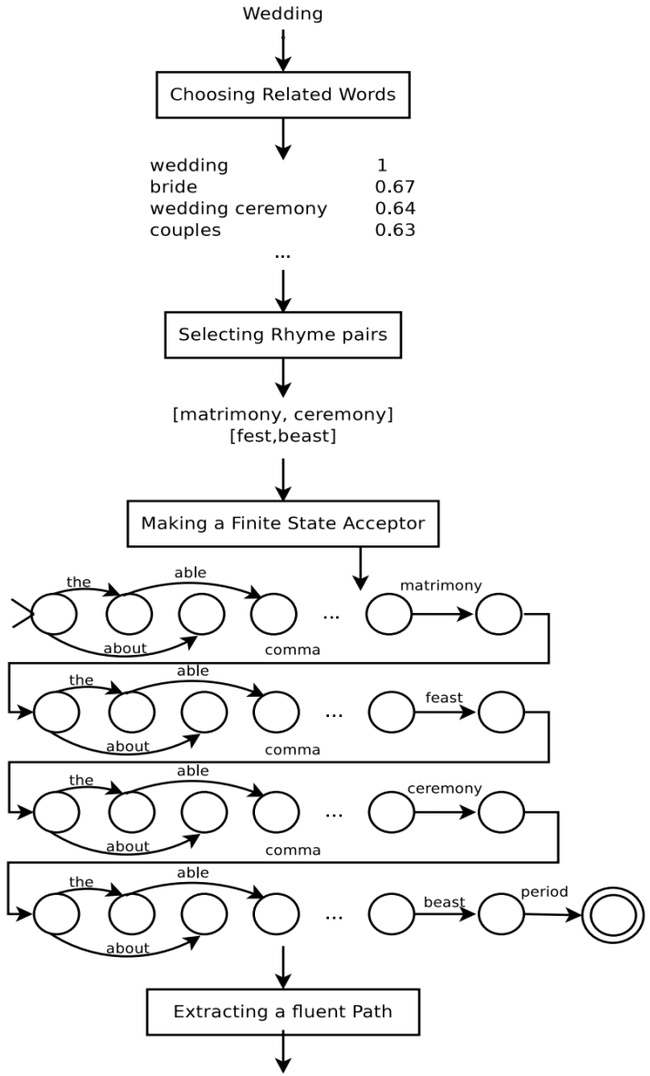
слога + бинарная пометка ударности

word	stress pattern	strict rhyme class	slant rhyme class (coarse version)
needing	10	IY1 D IH0 NG	IY1 * IH0 NG
ordinary	1010	EH1 R IY0	EH1 * IY0
obligate	101	EY1 T	last syllable stressed, no slant rhyme

Table 1: Sample word analyses.

Generating Topical Poetry, Ghazvininejad et al, 2016 // http://www.aclweb.org/anthology/D16-1126
Hafez: an Interactive Poetry Generation System, Ghazvininejad et al, 2017 // http://www.aclweb.org/anthology/P17-4008

Генерация поэзии: Hafez



Начитаем с «названия» (topic word)

Получаем множество слов, подходящих по теме
Из них – слова-окончания

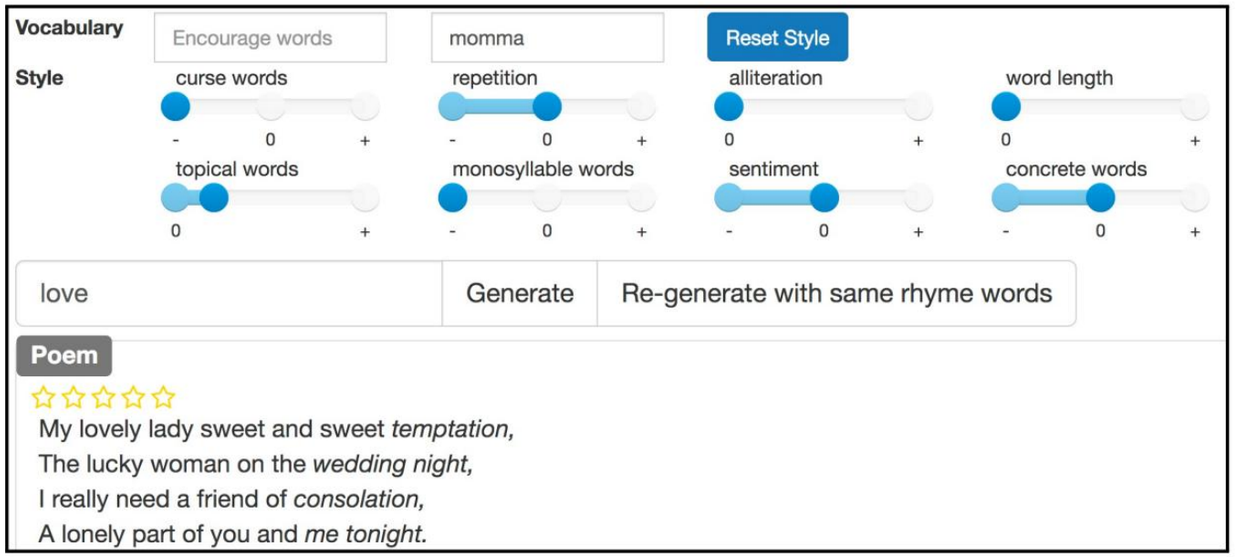
- **User-supplied input topic:** *colonel*
- **Output:** *colonel* (1.00), *lieutenant_colonel* (0.77), *brigadier_general* (0.73), *commander* (0.67) ... *army* (0.55) ...

Идентифицируем ритмичность – чтобы были правильные окончания в строках

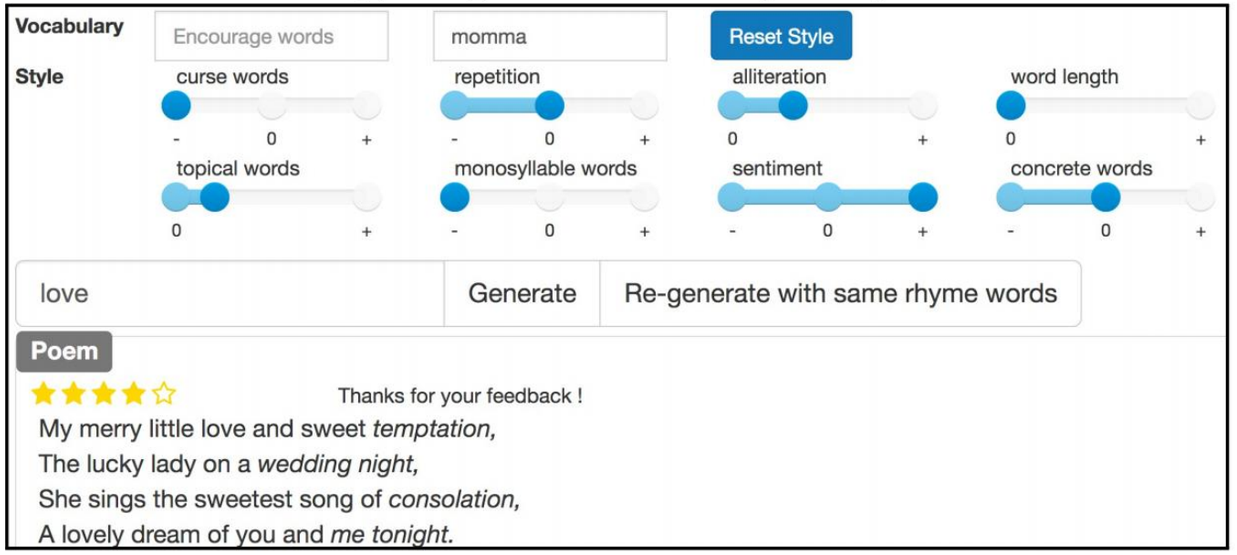
Finite-state acceptor (FSA) – для ритма

RNN-LM(←) + FSA

Генерация поэзии: Hafez



(a) Poem generated with default style settings



(b) Poem generated with user adjusted style settings

Генерация поэзии: Hafez – примеры

Love at First Sight

An early morning on a rainy night,
Relax and make the other people happy,
Or maybe get a little out of sight,
And wander down the streets of Cincinnati.

Girlfriend

Another party started getting heavy.
And never had a little bit of Bobby,
Or something going by the name of Eddie,
And got a finger on the trigger sloppy

Noodles

The people wanna drink spaghetti alla,
And maybe eat a lot of other crackers,
Or sit around and talk about the salsa,
A little bit of nothing really matters.

Civil War

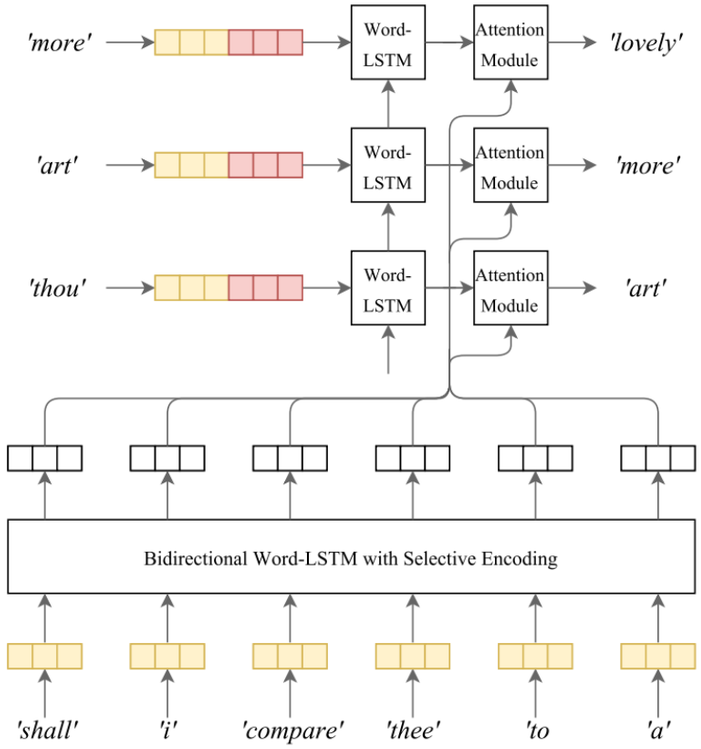
Creating new entire revolution,
An endless nation on eternal war,
United as a peaceful resolution,
Or not exist together any more.

Генерация поэзии: Deep-speare

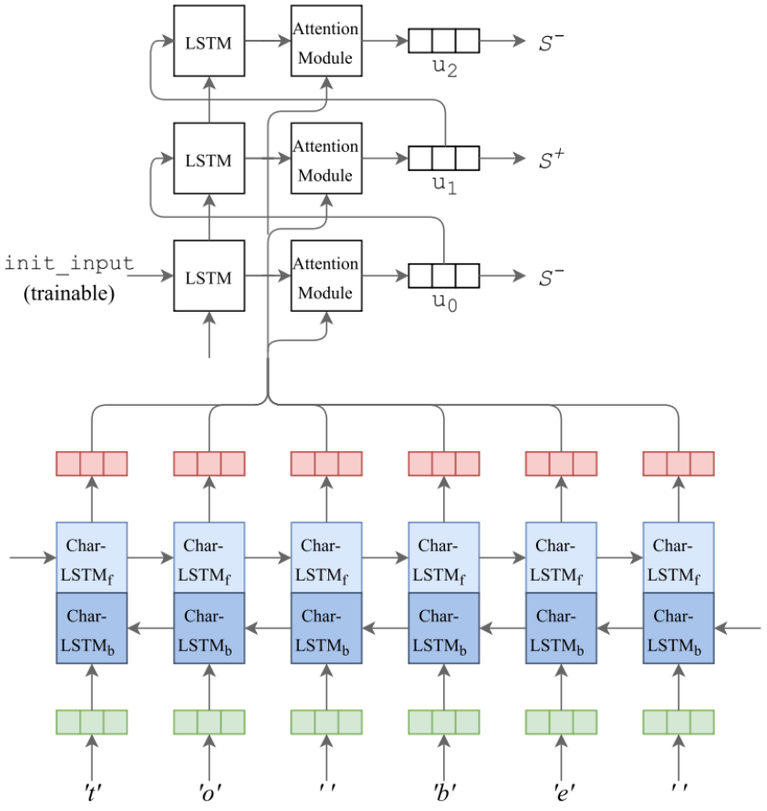
более end-to-end решение

3 компоненты – учим вместе как «multi-task problem»:
LM
pentameter model
rhyme model

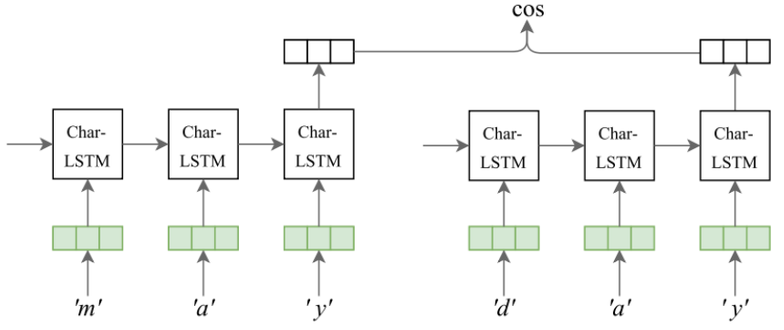
Lau et al «Deep-speare: A joint neural model of poetic language, meter and rhyme», 2018
http://aclweb.org/anthology/P18-1181



(a) Language model



(b) Pentameter model



(c) Rhyme model

Deep-speare – примеры

CMU Rhyming Pairs		CMU Non-Rhyming Pairs	
Word Pair	Cos	Word Pair	Cos
(*endeavour*, *never*)	0.028	(*blood*, *stood*)	1.000
(*nowhere*, *compare*)	0.098	(*mood*, *stood*)	1.000
(*supply*, *sigh*)	0.164	(*overgrown*, *frown*)	1.000
(*sky*, *high*)	0.164	(*understood*, *food*)	1.000
(*me*, *maybe*)	0.165	(*brood*, *wood*)	1.000
(*cursed*, *burst*)	0.172	(*rove*, *love*)	0.999
(*weigh*, *way*)	0.200	(*sire*, *ire*)	0.999
(*royally*, *we*)	0.217	(*moves*, *shoves*)	0.998
(*use*, *juice*)	0.402	(*afraid*, *said*)	0.998
(*dim*, *limb*)	0.497	(*queen*, *been*)	0.996

Table 3: Rhyming errors produced by the model. Examples on the left (right) side are rhyming (non-rhyming) word pairs — determined using the CMU dictionary — that have low (high) cosine similarity. "Cos" denote the system predicted cosine similarity for the word pair.

```
python sonnet_gen.py -m trained_model/ -d 1

Temperature = 0.6 - 0.8
  01   [0.43]  with joyous gambols gay and still array
  02   [0.44]  no longer when he twas, while in his day
  03   [0.00]  at first to pass in all delightful ways
  04   [0.40]  around him, charming and of all his days


python sonnet_gen.py -m trained_model/ -d 2

Temperature = 0.6 - 0.8
  01   [0.44]  shall i behold him in his cloudy state
  02   [0.00]  for just but tempteth me to stop and pray
  03   [0.00]  a cry: if it will drag me, find no way
  04   [0.40]  from pardon to him, who will stand and wait
```

https://github.com/jhlau/deepspeare

Проблемы метрики качества для саммаризации / диалогов / описания

– нет хорошей метрики качества

точно плохи повторы

хороши вопросы ⇒ вовлечённость пользователя

хороша клиенто/контексто-ориентированность

«Человечность» отличается от «качество диалога»

Why We Need New Evaluation Metrics for NLG, Novikova et al, 2017 https://arxiv.org/pdf/1707.06875.pdf
What makes a good conversation? How controllable attributes affect human judgments,
See et al, 2019 https://arxiv.org/pdf/1902.08654.pdf

Тренды в NLG (генерации текстов)

использование латентных переменных (ex: стиль текста)

Не однонаправленная генерация (слева-направо)

параллельная, последовательное улучшение (iterative refinement)

другие функции ошибки

больше на анализ предложений, а не отдельных слов

Итог

Проще задачи с ограничениями, open-ended-задачи сложнее

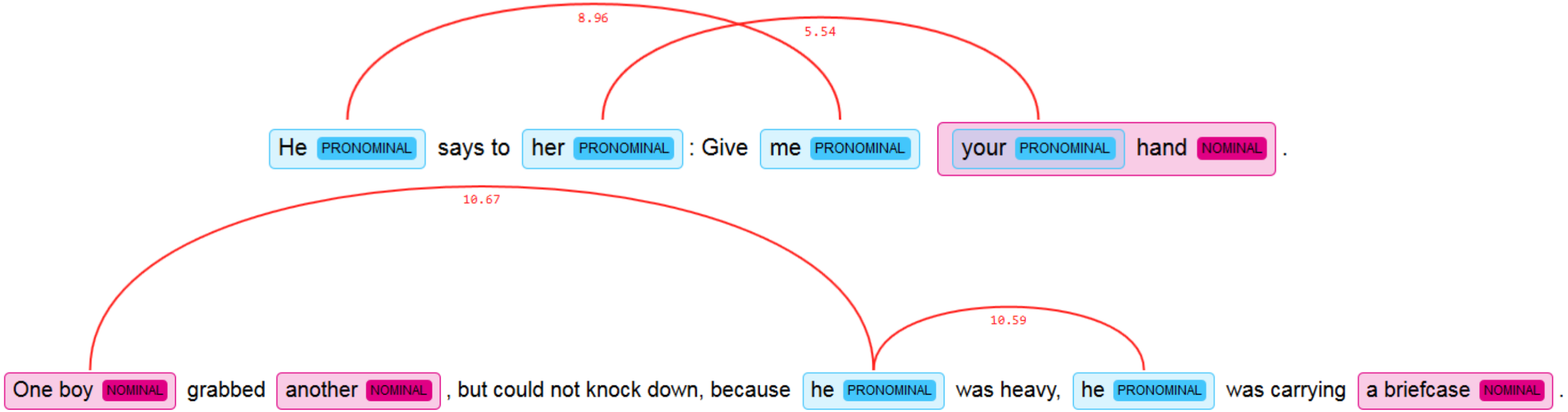
улучшение LM почти всегда приводит к улучшению качества генерируемого текста

нужно смотреть на много (неадекватных) метрик

Coreference Resolution

найти все сущности в тексте и кластеризовать их

понять, где речь об одной сущности



задача уровня теста Тьюринга

http://corenlp.run/

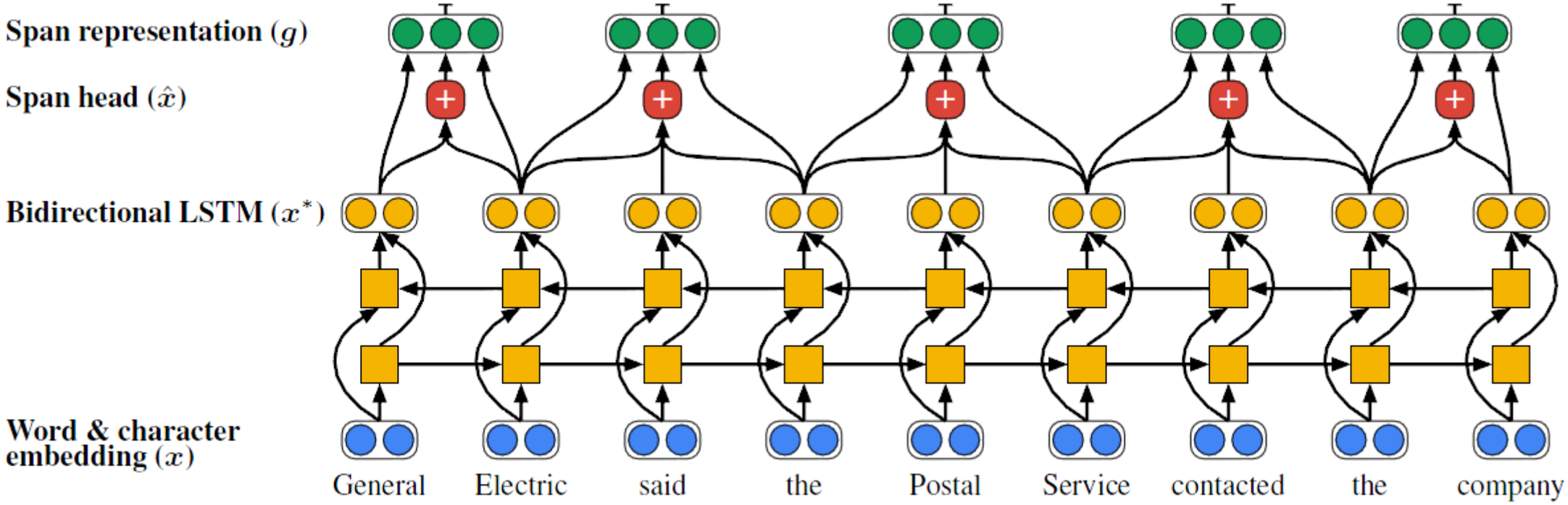
https://huggingface.co/coref/

Coreference Resolution: SOTA

рассмотреть все отрезки текста (span of text) до какой-то длины

LSTM + attention

оценить каждую пару отрезков: одинаковая ли сущность

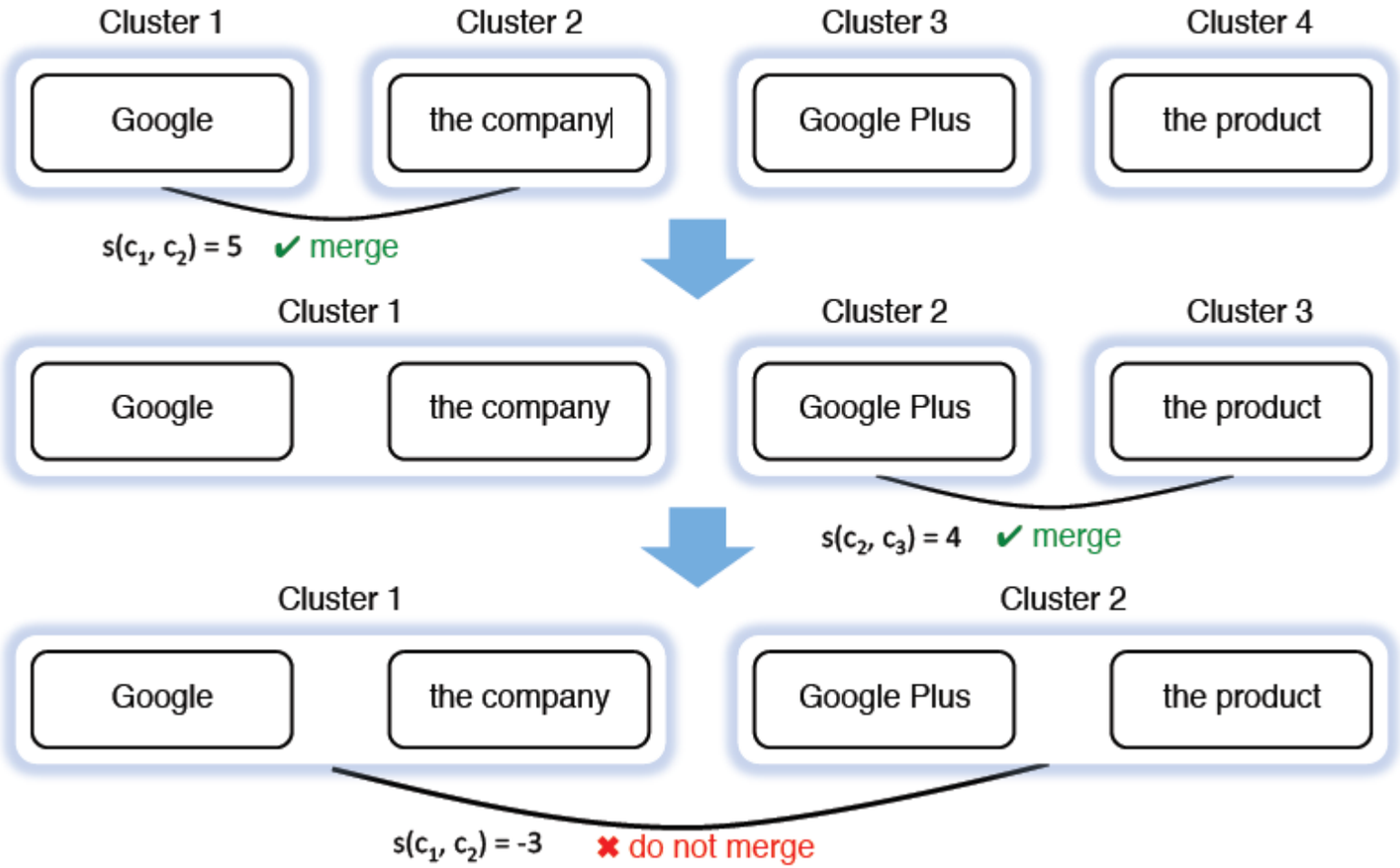


Kenton Lee et al. from UW, EMNLP 2017

Coreference Resolution: Clustering-Based

агломеративная кластеризация (собственно, по постановке задачи)

Google recently … **the company** announced **Google Plus** … **the product** features …



Clark & Manning, 2016 http://web.stanford.edu/class/cs224n/

Итоги

слова можно / нужно расщеплять на подслова
есть проблема OOV-слов!
можно обрабатывать слова посимвольно

суммаризация – механизм копировоания
+ механизм контроля, что копировать

Истории – иерархические модели

…