курс «Глубокое обучение»

Векторные представления слов и текстов

Александр Дьяконов

30 марта 2020 года

План

классические способы представления слов

OHE, counts, LSA, кластеризация, LDA

DL-классика

word2vec, fasttext, Glove

учёт контекста

CoVe, ELMo, FLAIR

представление текстов

Doc2Vec / paragraph2vec, The skip-thoughts model,
Autoencoder pretraining, StarSpace, DAN
Universal Sentence Encoder

DSSM

Способы кодирования / представления слов

- **OHE**

слишком большая размерность, нет хорошей близости

- **counts (сумма OHE соседей)**

более нетривиальная оценка близости с помощью cos

- **вложение (embeddings)**

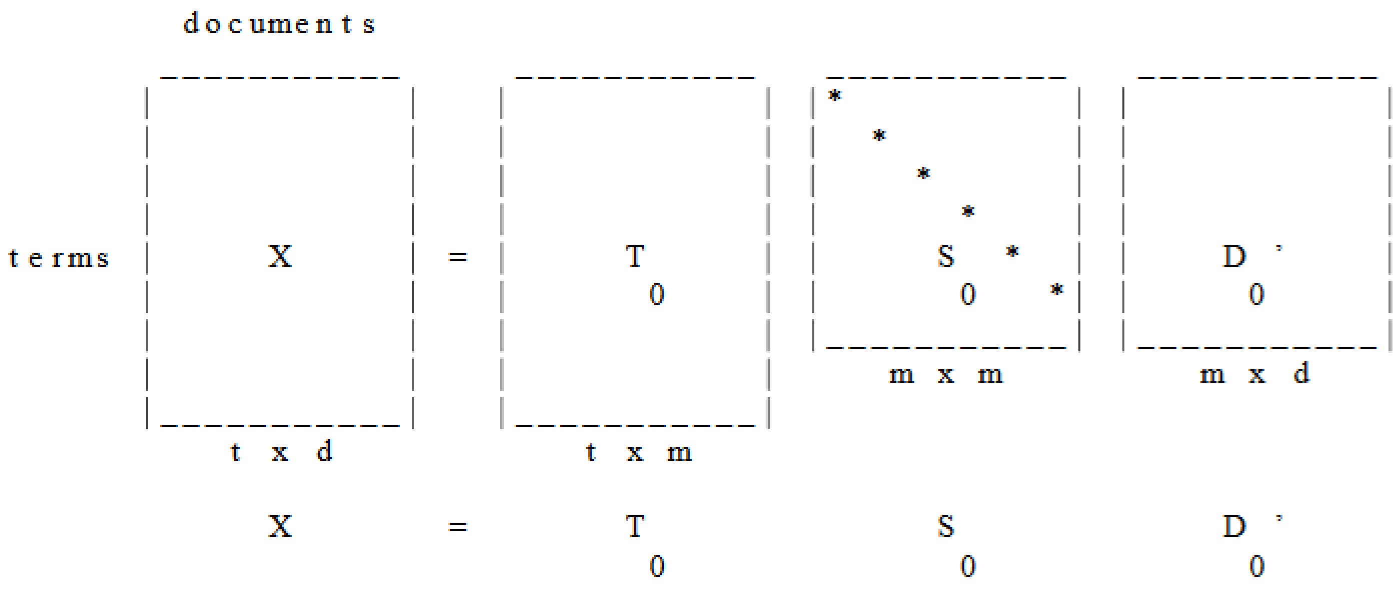
умный алгоритм задания кодировки

«word embeddings»

Представления слов в вещественном многомерном пространстве
$\Rightarrow$ можно использовать в матмоделях

Предобученные **Обученные для конкретной задачи**

Классические способы представления слов: LSA



S Deerwester «Indexing by latent semantic analysis», 1990

http://lsa.colorado.edu/papers/JASIS.lsi.90.pdf

Классические способы представления слов: кластеризация слов

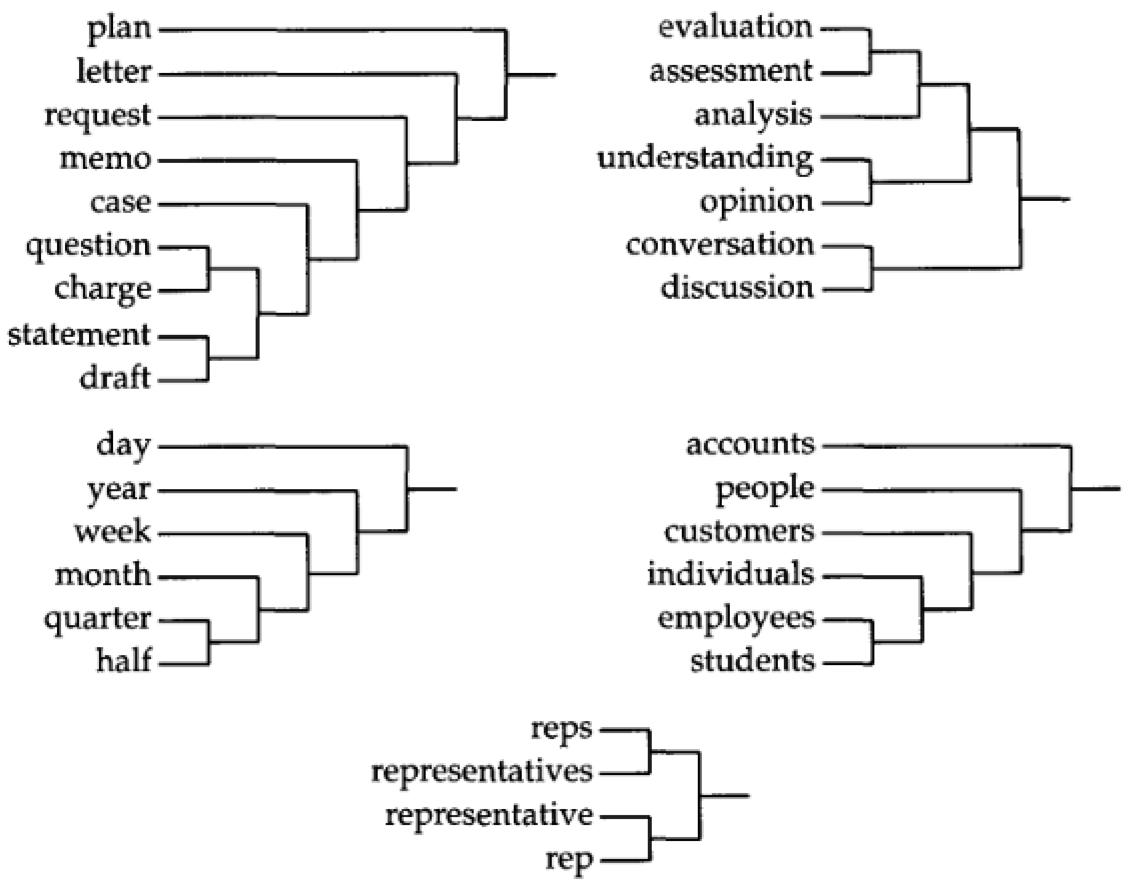


Figure 2
Sample subtrees from a 1,000-word mutual information tree.

Peter F. Brown et. al. «Class-Based n-gram Models of Natural Language»
https://www.aclweb.org/anthology/J92-4003.pdf

Классические способы представления слов: LDA

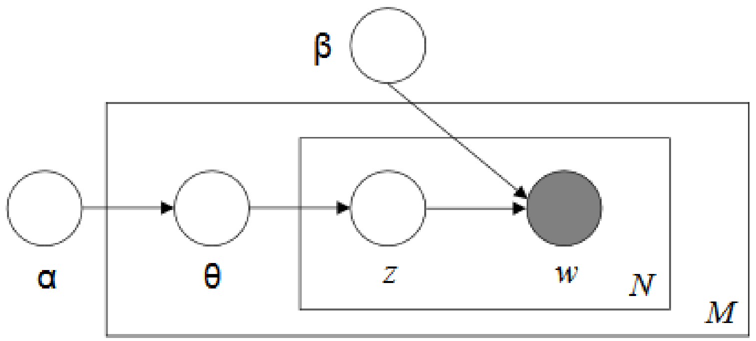


Figure 1: Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.

“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

D.M. Blei «Latent Dirichlet Allocation» // Journal of Machine Learning, 2003

http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Для чего использовались: n-граммная языковая модель

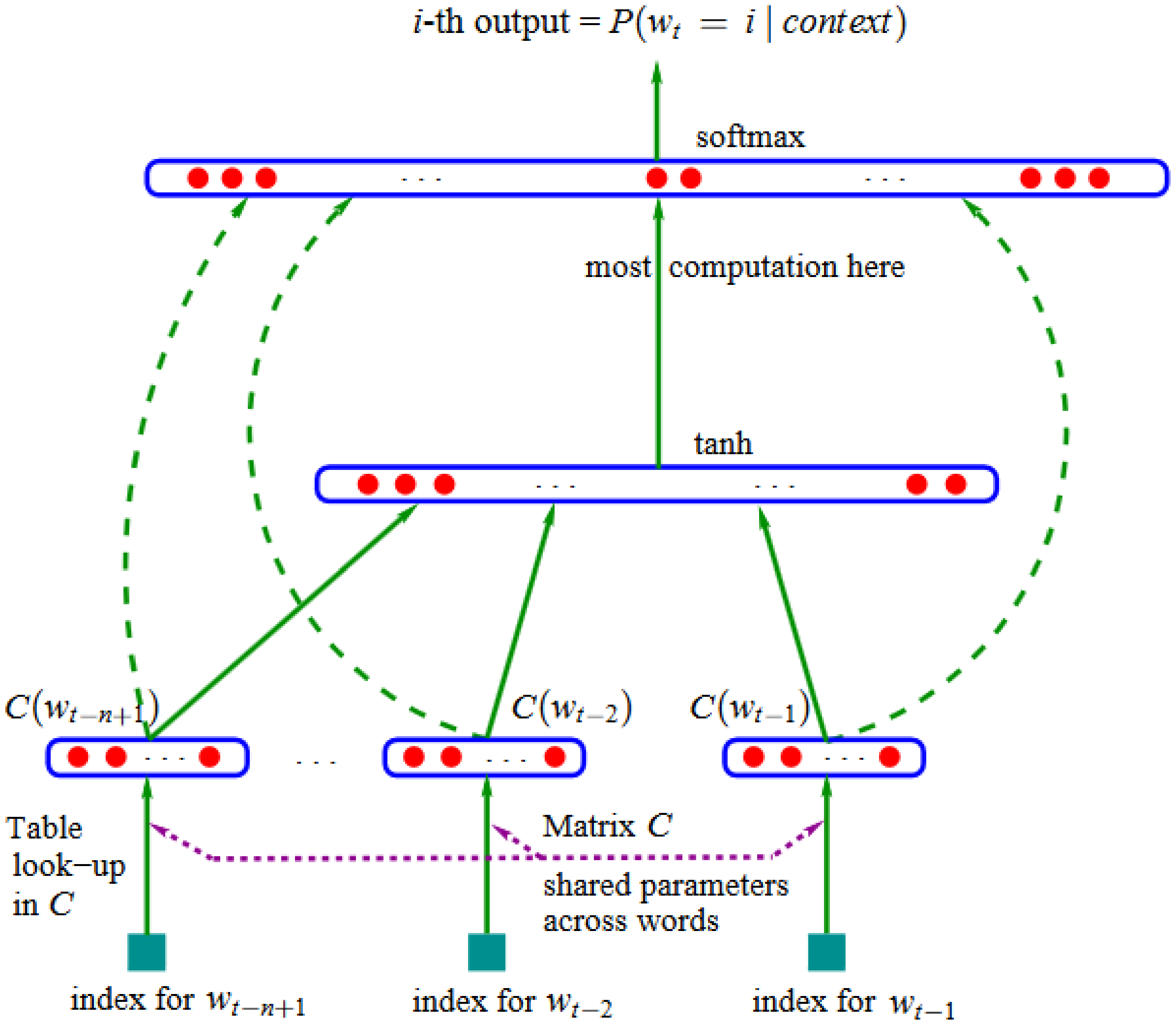
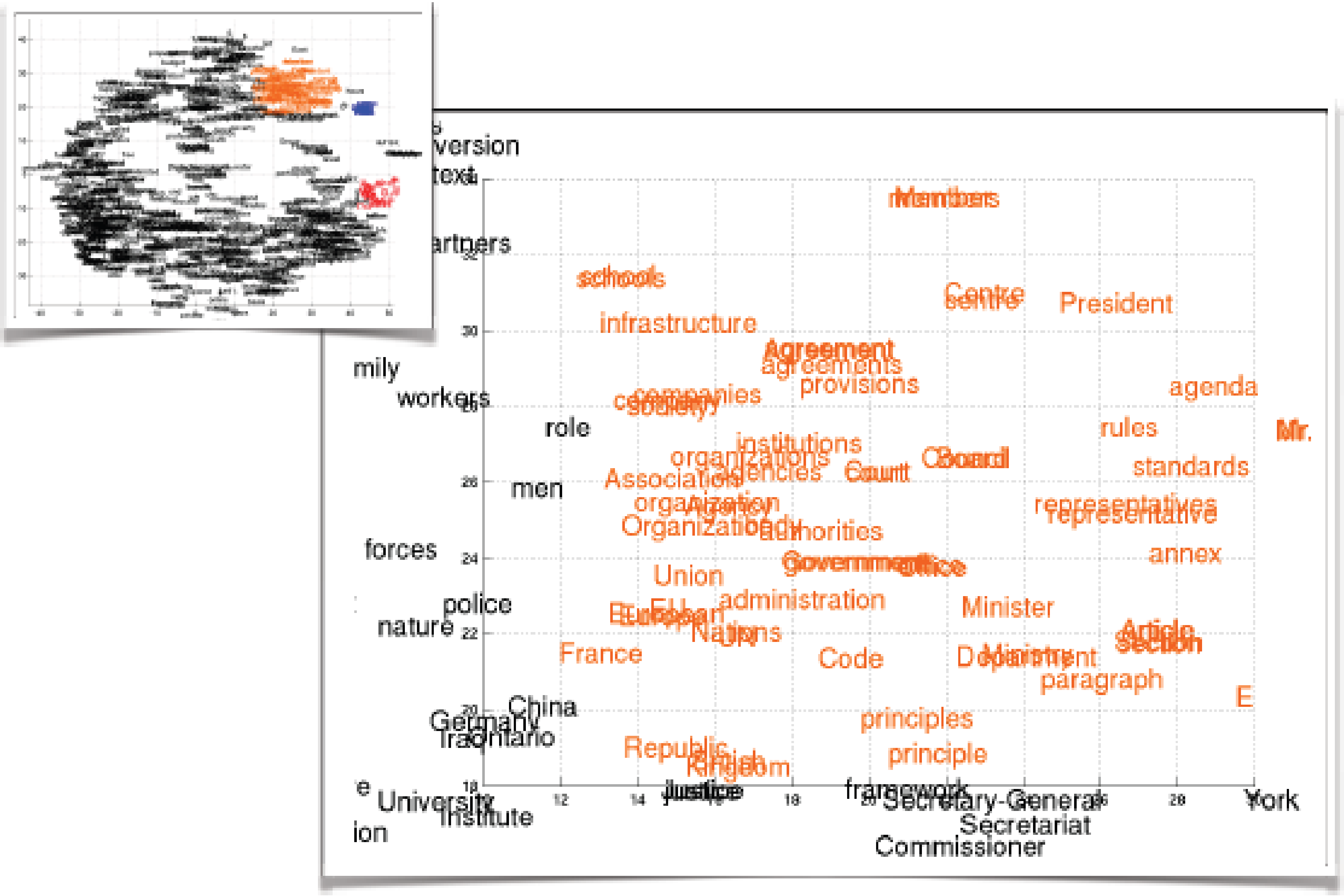


Figure 1: Neural architecture: $f(i, w_{t-1}, \cdots, w_{t-n+1}) = g(i, C(w_{t-1}), \cdots, C(w_{t-n+1}))$ where g is the neural network and $C(i)$ is the i-th word feature vector.

Yoshua Bengio «Neural Probabilistic Language Model» Journal of Machine Learning Research

http://www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf

Вложение слов в непрерывное пространство (embedding)



DL-классика

Несколько популярных способов
context-free – не учитывающих контекст
(точнее, ограниченно учитывающих)

- **word2vec** = предсказания слово ↔ контекст
- **fasttext** = word2vec + ngrams
- **Glove** = разложение матрицы совместной встречаемости

word2vec – дистрибутивная семантика

Трюк: настраиваем модель, но не для использования в задаче, которой учим (нас интересуют формируемые внутренние представления)
Аналогично было в автокодировщиках;)

Термины «distributional semantics»

Смысл слова определяется контекстом

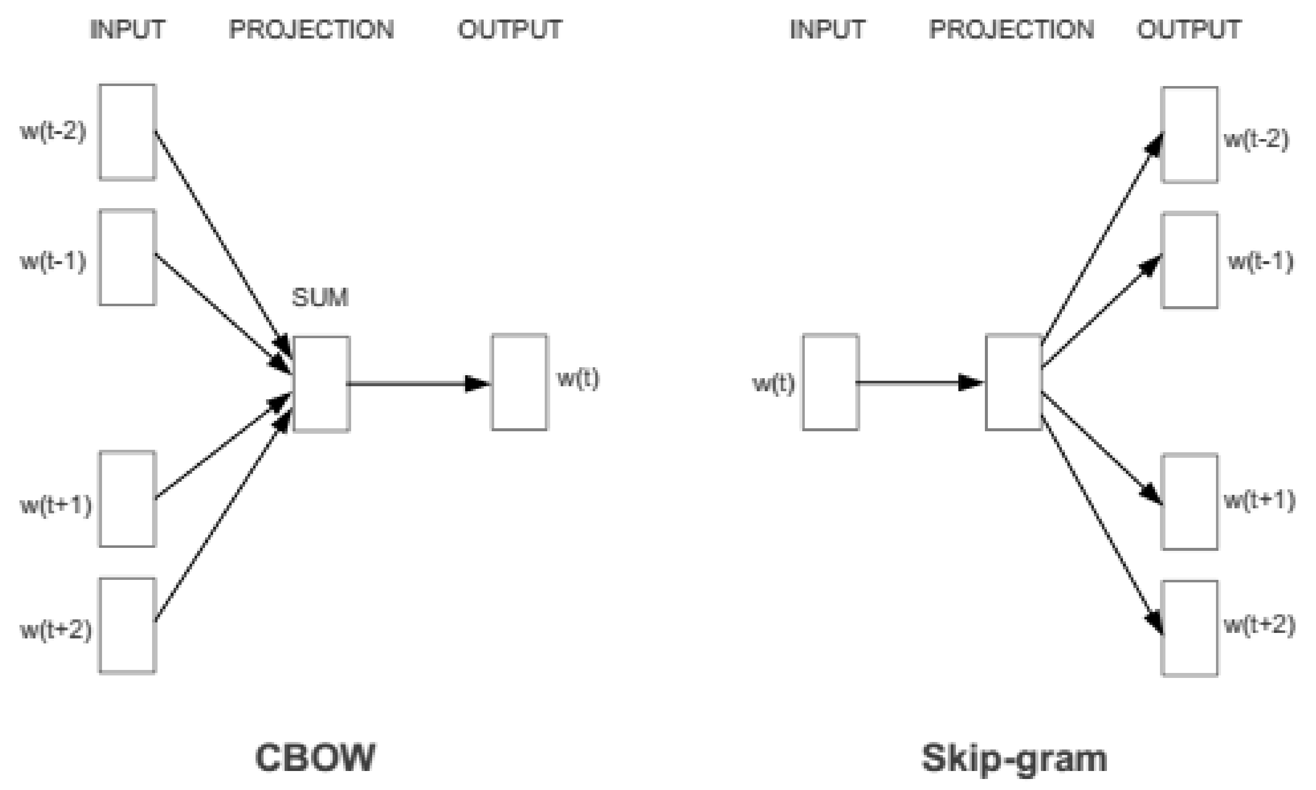
Полосатая маленькая ***** мурлычит и пьёт молоко

Весна

Ручьи
Тает
Цветёт
Зеленеет
Прилетают

[Mikolov et al. 2013]

word2vec: два подхода к реализации



CBOW = Continuous Bag of Words (быстрее, окно ~ 5, большие копуса)
skipgram model (лучше, окно ~ 10, небольшие корпуса)

word2vec: два метода обучения

позже

- **Hierarchical Softmax**
- **Negative Sampling**

word2vec: CBOW

Предсказываем слово по контексту

используется реже, чем следующая реализация

$$P(x_t \mid \text{context}(x_t)) = \text{softmax}\left(V\left(W \sum_{x_i \in \text{context}(x_t)} OHE(x_i)\right)\right)$$

выделено то, что будем считать кодировкой

контекст – слово (слова), которое недалеко располагается (в окрестности)

http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

word2vec: CBOW

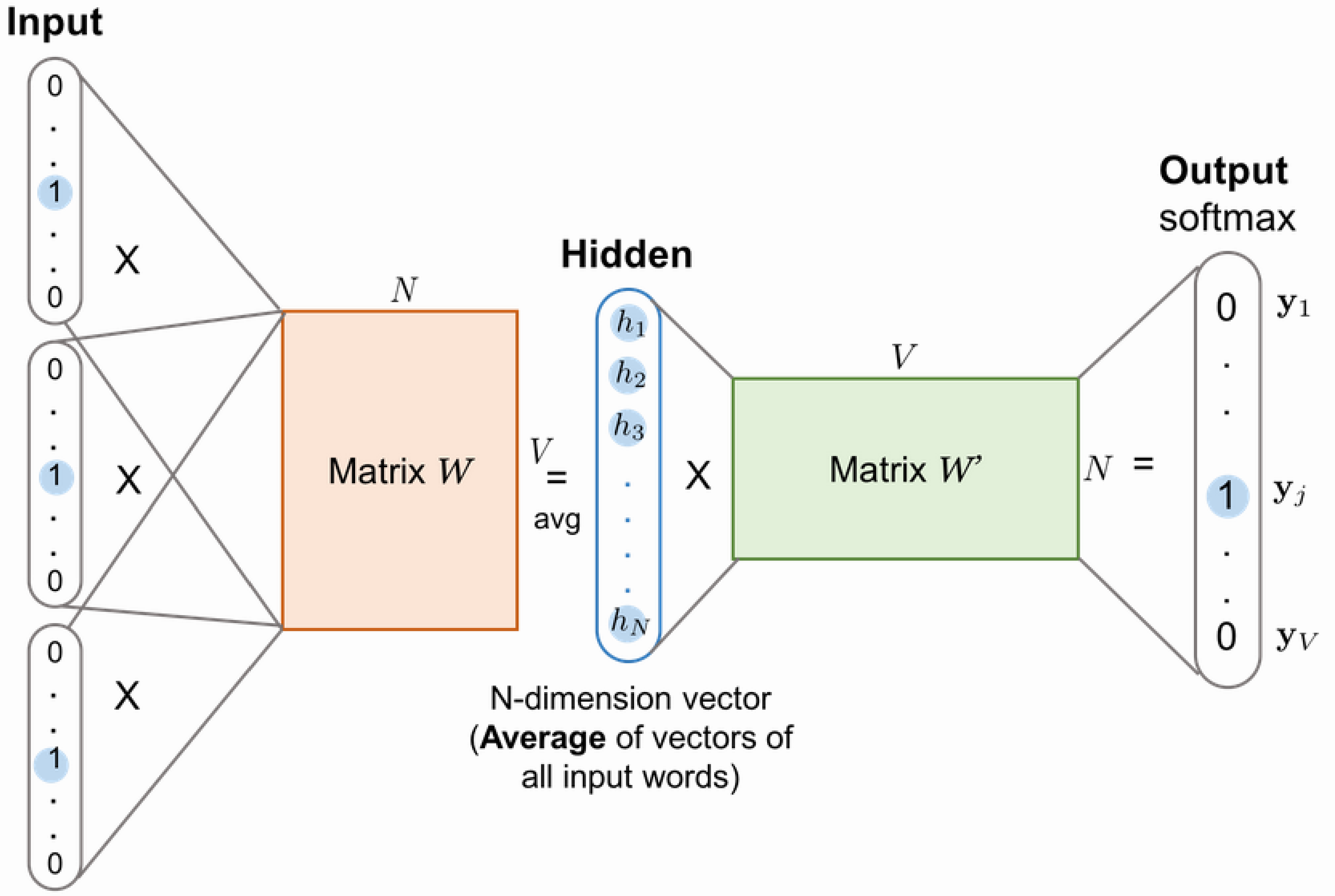


Fig. 2. The CBOW model. Word vectors of multiple context words are averaged to get a fixed-length vector as in the hidden layer. Other symbols have the same meanings as in Fig 1.

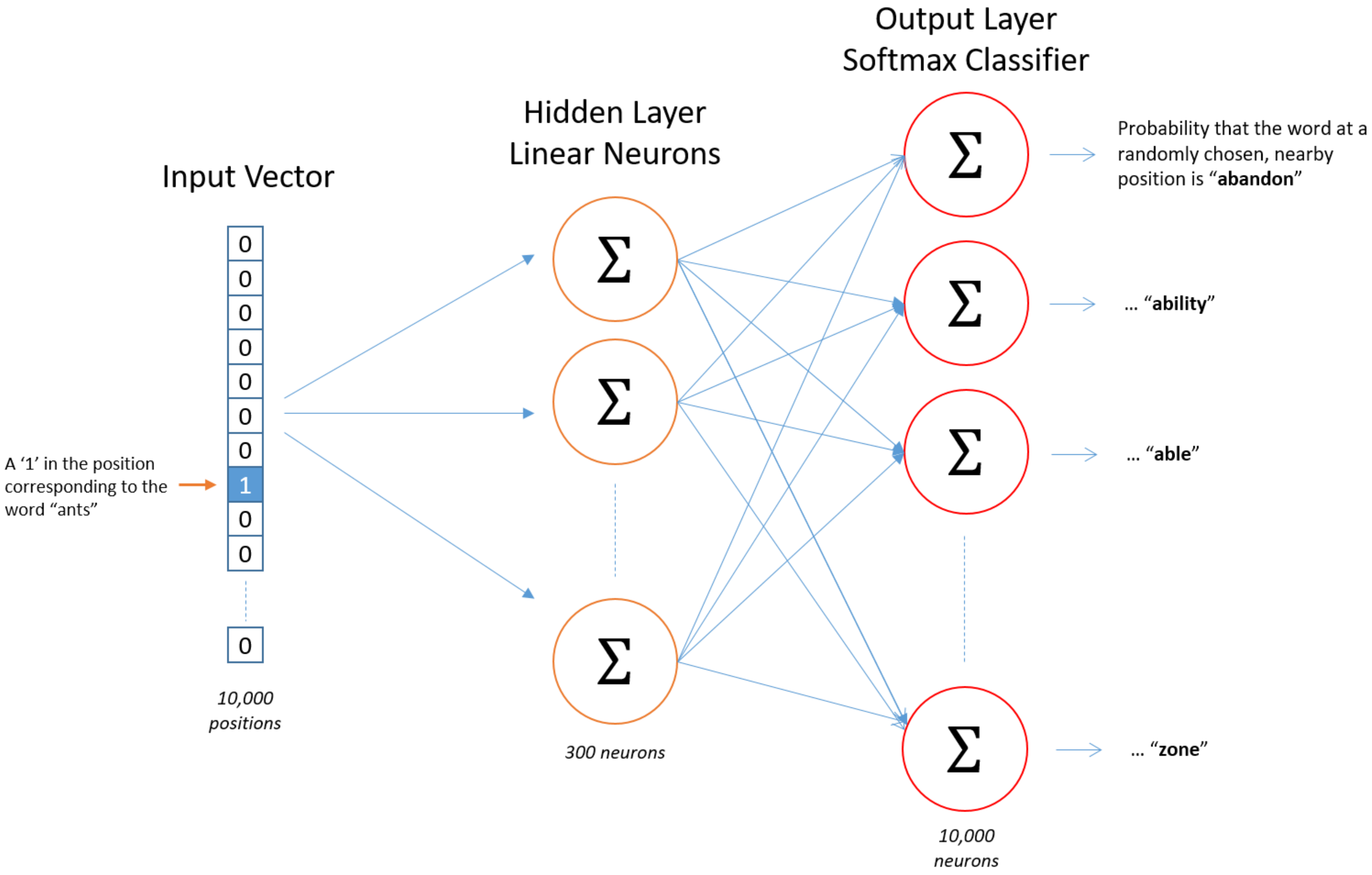
https://lilianweng.github.io/lil-log/2017/10/15/learning-word-embedding.html

word2vec: skip-gram

Предсказываем контекст по слову

Source Text	Training Samples
The quick brown fox jumps over the lazy dog. ⟹	(the, quick) (the, brown)
The **quick** brown fox jumps over the lazy dog. ⟹	(quick, the) (quick, brown) (quick, fox)
The quick **brown** fox jumps over the lazy dog. ⟹	(brown, the) (brown, quick) (brown, fox) (brown, jumps)
The quick brown **fox** jumps over the lazy dog. ⟹	(fox, quick) (fox, brown) (fox, jumps) (fox, over)

word2vec: skip-gram



вход: **ОНЕ-кодировка слова** **выход:** **распределение вероятностей**

Средний слой – для нашего кодирования

word2vec: skip-gram

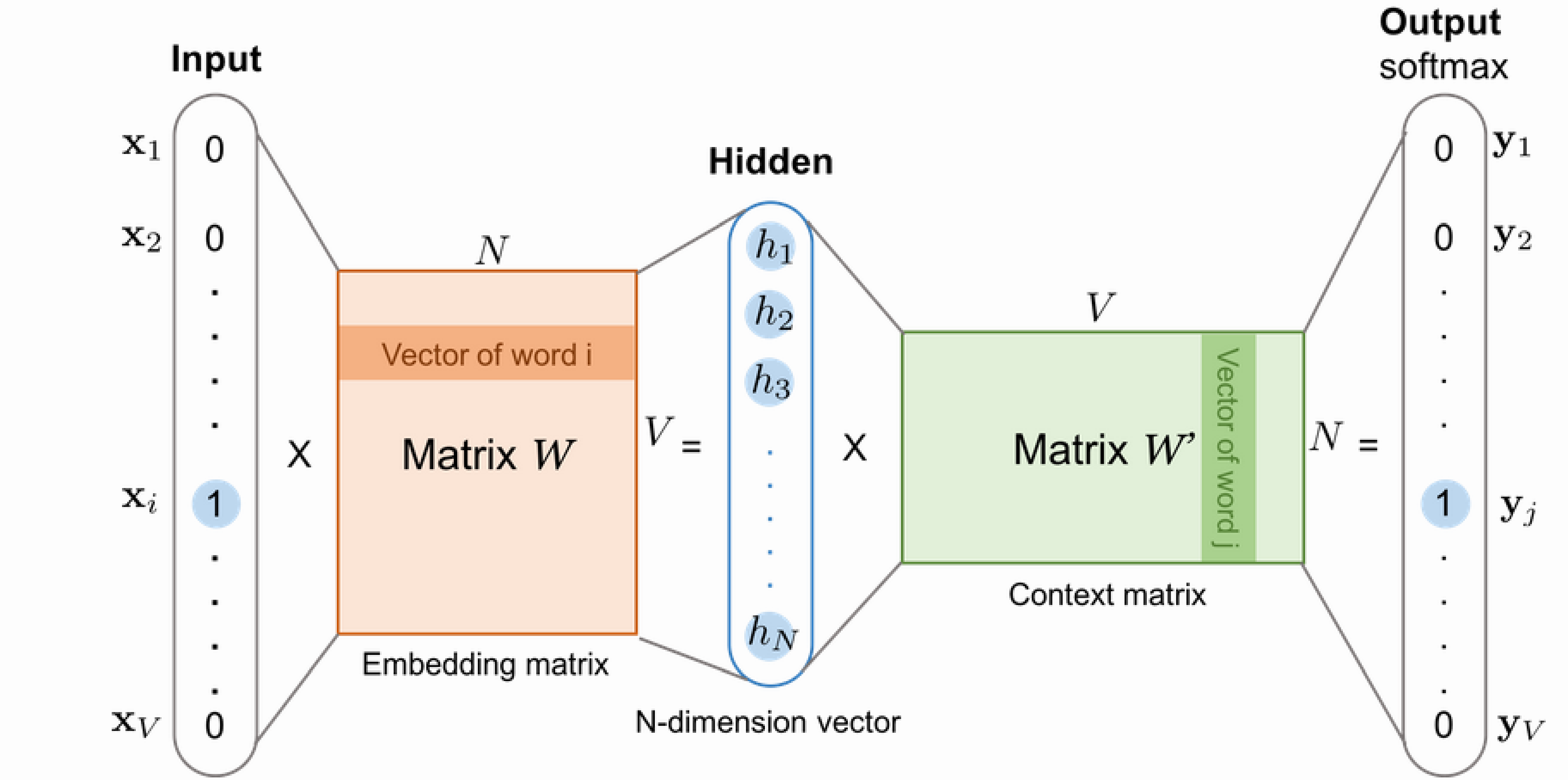


Fig. 1. The skip-gram model. Both the input vector $\mathbf{x}$ *and the output* $\mathbf{y}$ *are one-hot encoded word representations. The hidden layer is the word embedding of size* N.

https://lilianweng.github.io/lil-log/2017/10/15/learning-word-embedding.html

word2vec

Огромная НС

Первый слой – #слов × размерность предствления

Как обучать????

«Distributed Representations of Words and Phrases and their Compositionality» [Mikolov T. 2013 https://arxiv.org/pdf/1310.4546.pdf]

/ код слова = строка первой матрицы + столбец второй

Следующие слайды по
http://mccormickml.com/2017/01/11/word2vec-tutorial-part-2-negative-sampling/

Есть отличия между реализацией и статьёй!

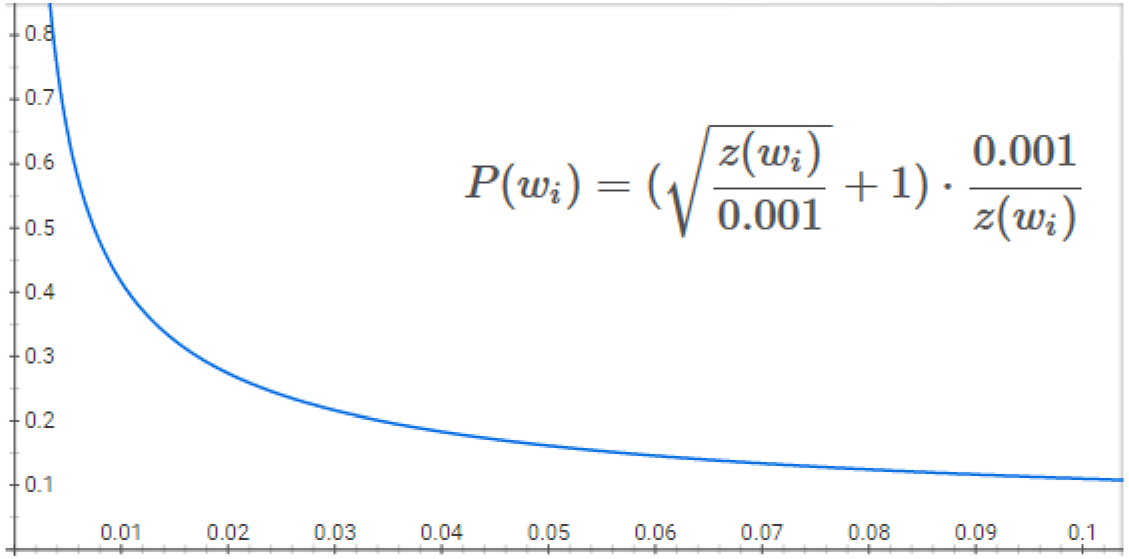
word2vec

Распространённые фразы – одно слово

White_Spunner_Construction
Bad_Habits
Toxics_Alliance

Частые слова – реже выбираются при обучении

вероятность быть выбранным от частоты:



«Negative Sampling»

y («открыл») = OHE(«дверь»)
чтобы не править много выходов, соответствующим нулям, выбираем несколько случайных (5–20)

word2vec – немного математики

Последовательность слов $x_1,\ldots,x_T$

Правдоподобие

$$\prod_{t=1}^{T}\prod_{c\in C_t} p(x_c \mid x_t) \sim \sum_{t=1}^{T}\sum_{c\in C_t} \log p(x_c \mid x_t) \to \max$$

(второе произведение по окрестности – индексы соседних слов)

Можно: $p(x_c \mid x_t) = \dfrac{\exp(s(x_t, x_c))}{\sum_x \exp(s(x_t, x))}$

Такая модель подходила бы,
если бы для каждого слова один правильный ответ
хотя тоже используется

word2vec: Negative Sampling

Как делаем… «skipgram model with negative sampling» [Mikolov]

Используем «negative log-likehood»

$$\log\left(1+\exp(-s(x_t,x_c))\right)+\sum_{x\in N_{t,c}}\log\left(1+\exp(s(x_t,x))\right)$$

$N_{t,c}$ **– выборка негативных примеров**

Если logloss $l(z)=\log\left(1+\exp(-z)\right)$**, то**

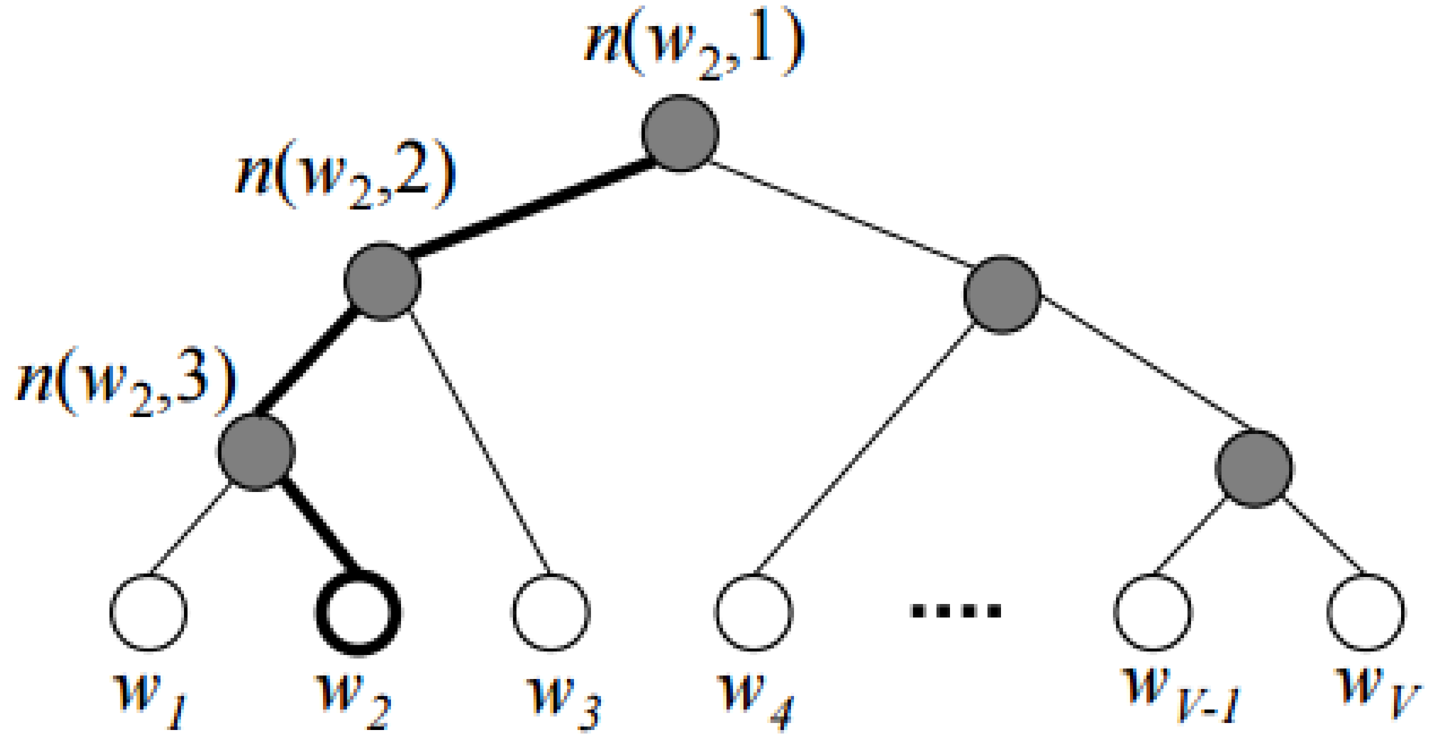
$$\sum_{t=1}^{T}\left[\sum_{c\in C_t} l(s(x_t,x_c))+\sum_{x\in N_{t,c}} l(-s(x_t,x))\right]\to\min$$

Скоринговая функция: $s(x_t,x_c)=\mathrm{vec}(x_t)^{\mathrm{T}}\cdot\mathrm{vec}(x_c)$

тут нужны будут негативные примеры

Hierarchical Softmax

softmax-слой представляется так (специальная кодировка Хаффмана)



листья – слова

вероятность = произведение вероятностей в вершинах пути

Ближайшие соседи

peace	**Path**	**Stop**
Peaceful	**Paths**	**Quit**
Friendship	**Approach**	**Stopped**
Nonviolence	**Titled**	**Avoid**
	Pathway	**Resist**
	Way	

http://bionlp-www.utu.fi/wv_demo/

Операции над представлениями слов

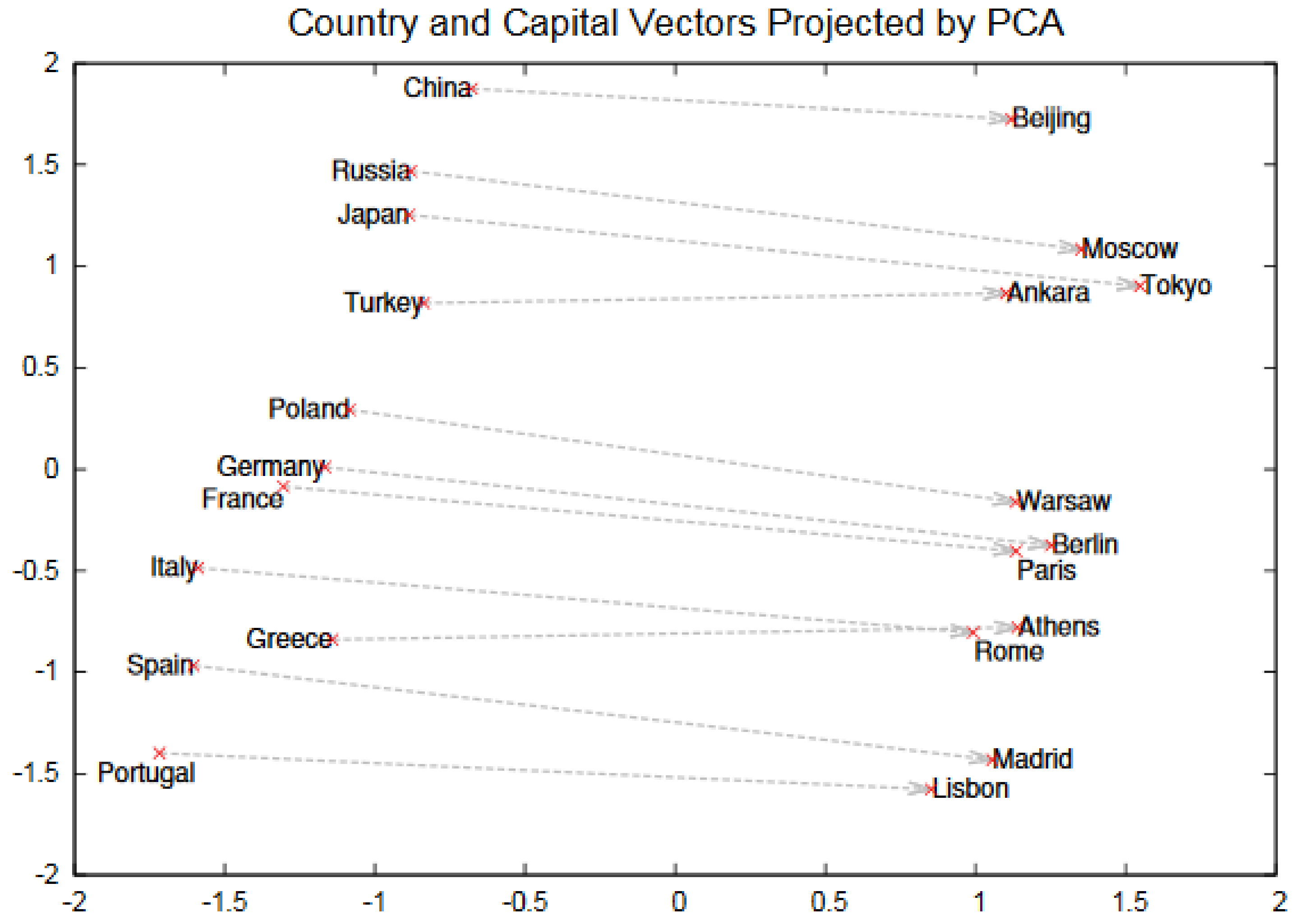


Figure 2: Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. The figure illustrates ability of the model to automatically organize concepts and learn implicitly the relationships between them, as during the training we did not provide any supervised information about what a capital city means.

[Mikolov et al., 2013] https://arxiv.org/pdf/1310.4546.pdf

Другие представления: fasttext

тоже «слово → контекст»

попытка учесть морфологию слов

раньше «сеть», «сетевой», «сетью» разные векторы…

+ использовать n-граммные представления слова

«where» ~ <wh, whe, her, ere, re>

n-граммы хэшируются;)

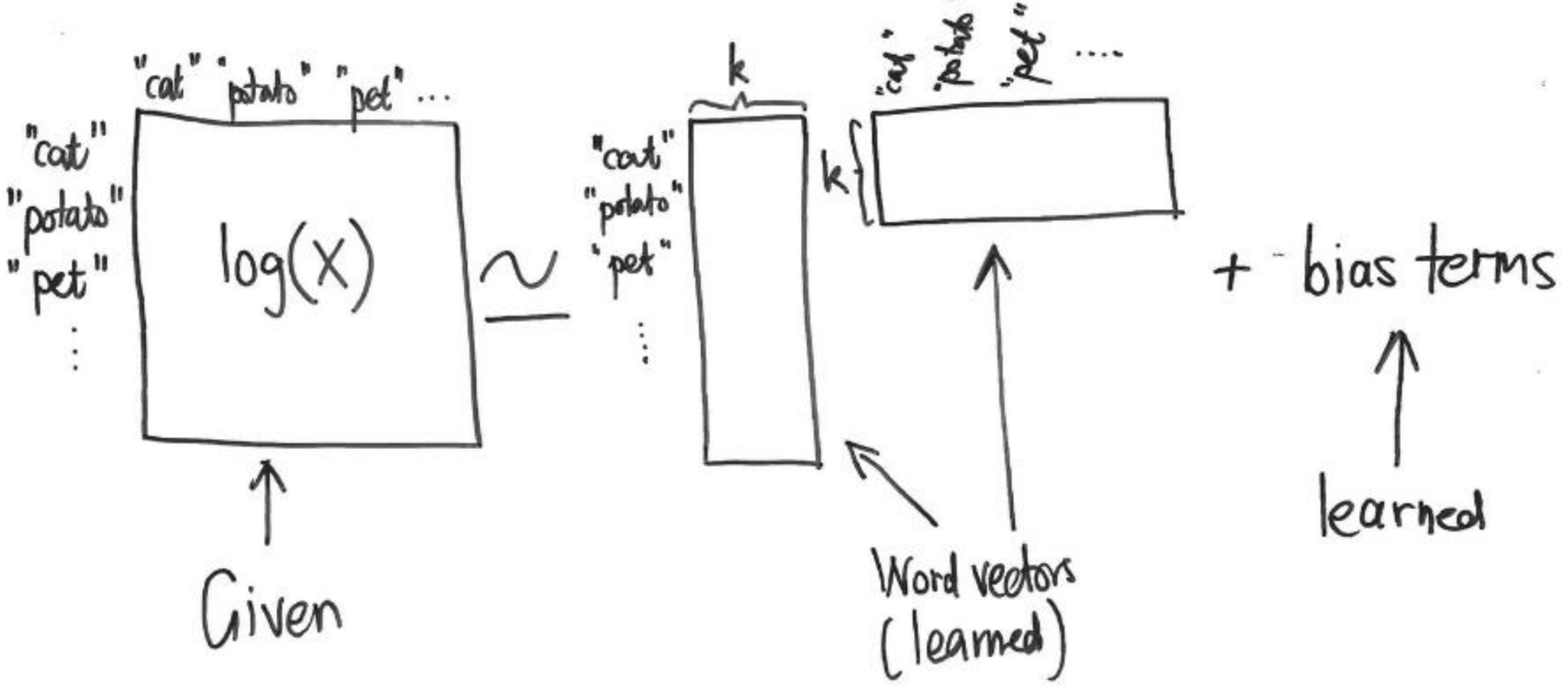
код = сумма кодов для n-грамм

Решается проблема новых слов

«Enriching Word Vectors with Subword Information» [Bojanowski P. et al., 2017 https://arxiv.org/pdf/1607.04606.pdf]

https://fasttext.cc – тут есть все ссылки!!!

Glove: Global Vectors for Word Representation



идея в разложении матрицы

http://building-babylon.net/2015/07/29/glove-global-vectors-for-word-representations/

https://nlp.stanford.edu/projects/glove/

Glove: Global Vectors for Word Representation

$\#ij$ **– сколько раз слово j в контексте слова i**

(на расстоянии ≤k слов) есть и другие варианты

$$\sum_{i,j} f(\#ij)(w_i^{\mathrm{T}}\tilde{w}_j + b_i + \tilde{b}_j - \log(\#ij))^2 \to \min$$

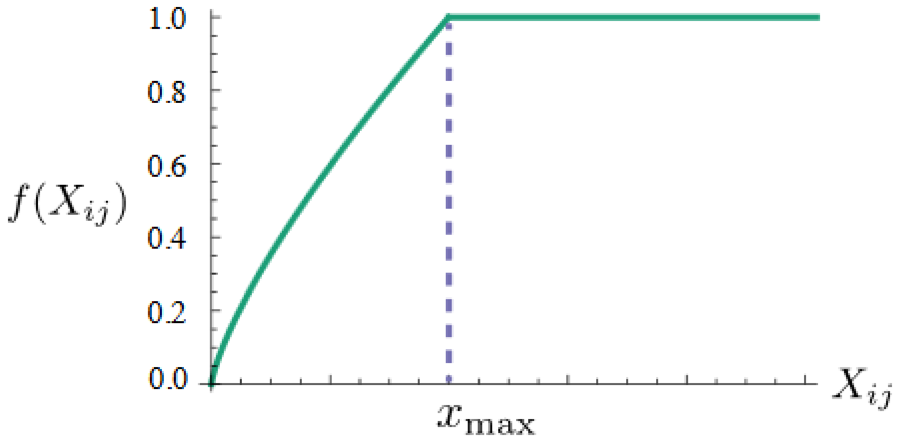
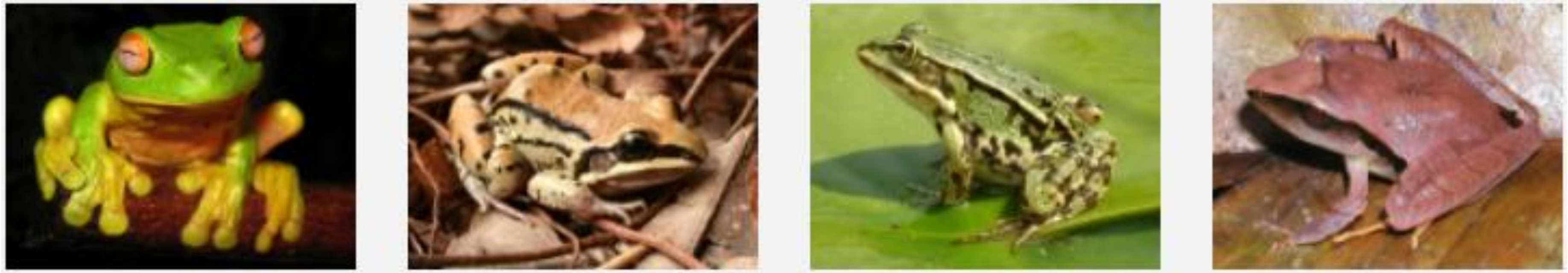


Figure 1: Weighting function f with $\alpha = 3/4$.

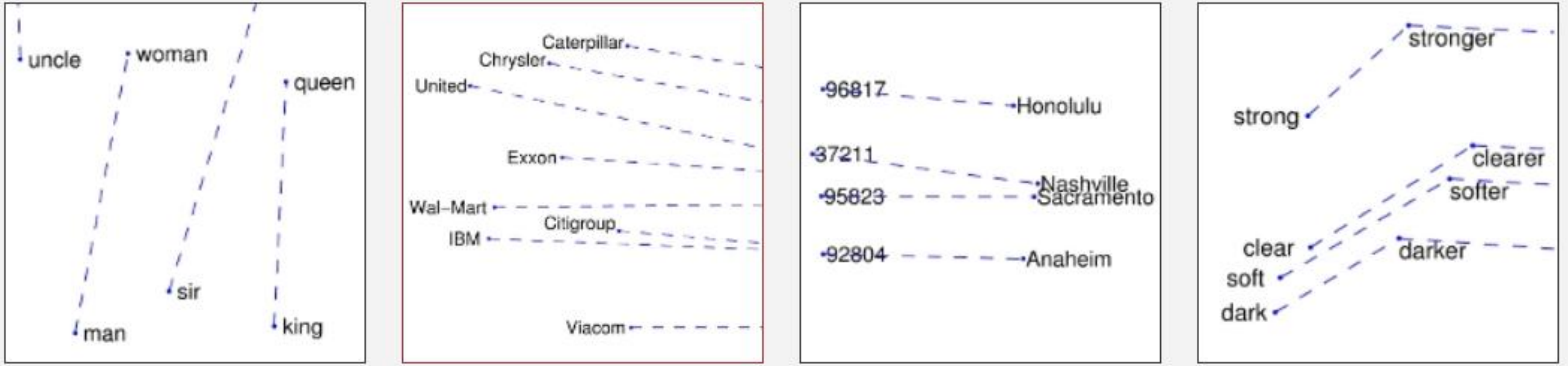
$$f(x) = \begin{cases} \left(\dfrac{x}{x_{\max}}\right)^{\alpha}, & x < x_{\max}, \\ 1, & x \geq x_{\max}. \end{cases}$$

Glove: ближайшие соседи

frog
frogs
toad
litoria
leptodactylidae
rana
lizard
leutherodactylus



3. litoria 4. leptodactylidae 5. rana 7. eleutherodactylus



man - woman | company - ceo | city - zip code | comparative - superlative

Contextualized Word Embeddings

недостатки предыдущих вложений – не учитывают контекст

«Рискую всем банком»
«В банке не работал кондиционер»
«Хранить деньги в банках не стоит»
«На банке сидела муха»

«The bank will not be accepting cash on Saturdays»
«The river overflowed the bank»

Выход:
языковые модели

- embeddings in Tag LM
- CoVe
- ELMo
- Flair

Embeddings in Tag LM

Одна из первых работ с идеей, что недостаточно просто представлений слов

Используются
предобученные представления слов
предобученная нейронная LM
оба представления используются
решалась задача простановки тегов

Matthew E. Peters et. al. «Semi-supervised sequence tagging with bidirectional language models» // https://arxiv.org/pdf/1705.00108.pdf

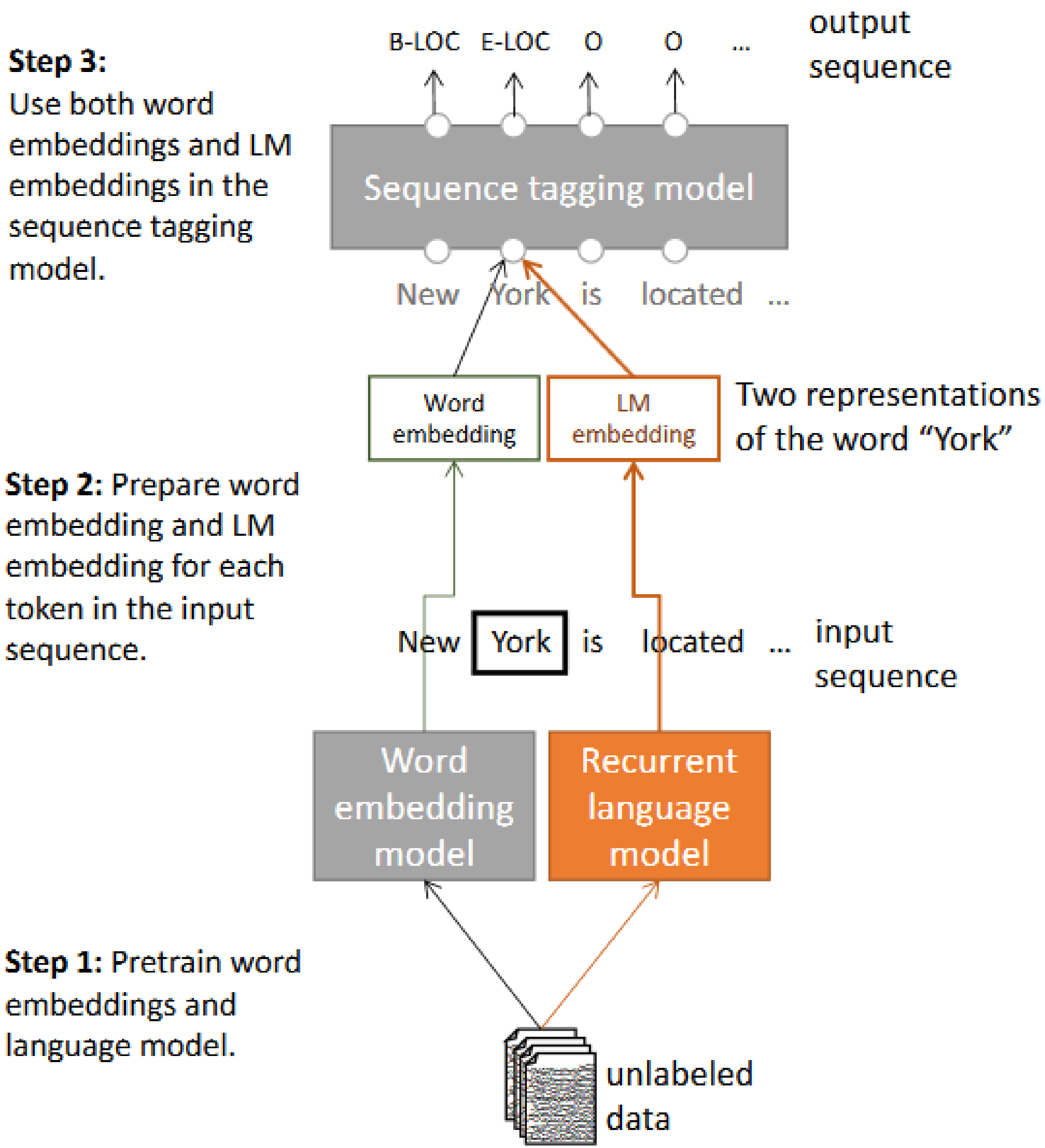
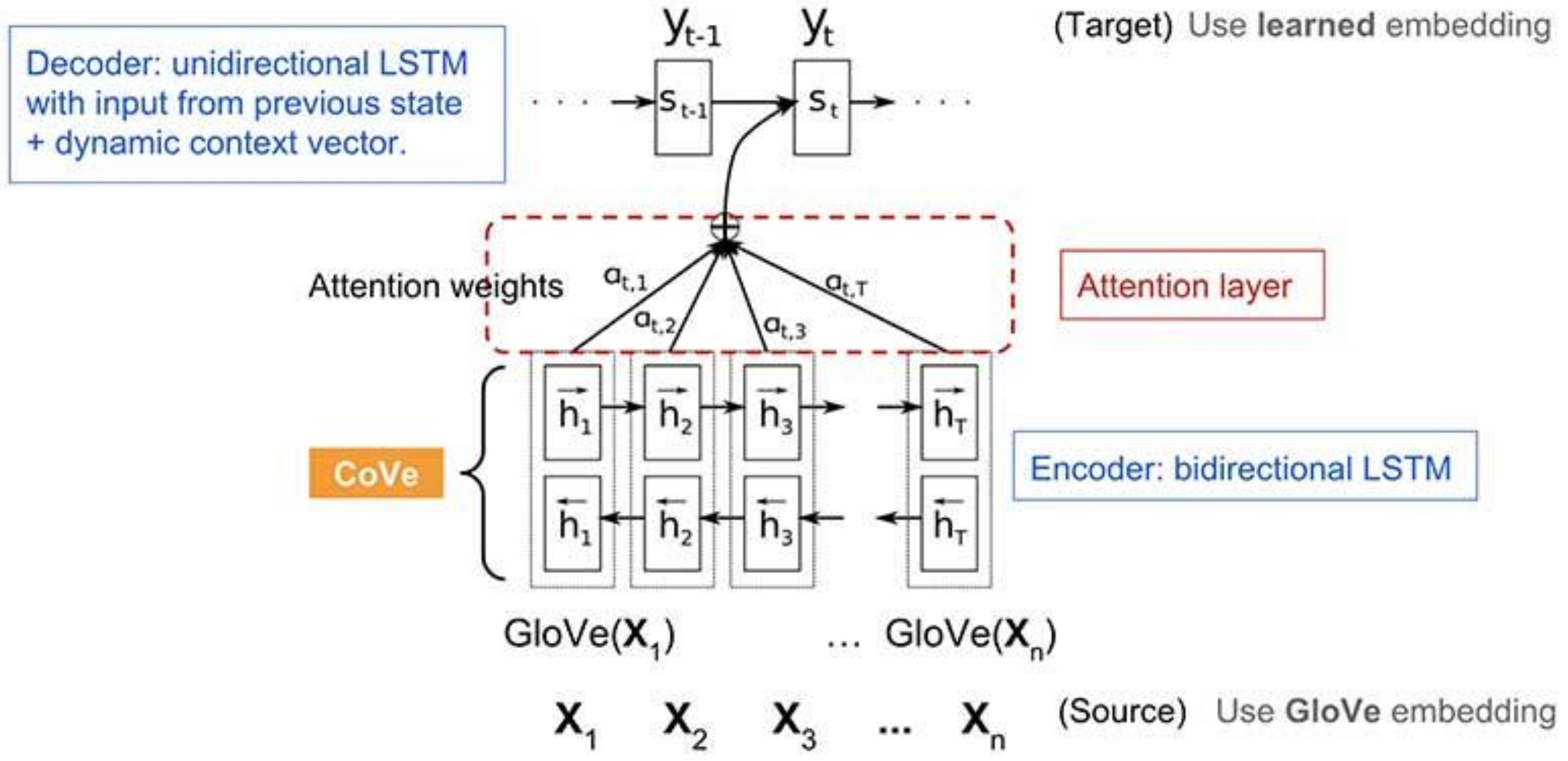


Figure 1: The main components in TagLM, our language-model-augmented sequence tagging system. The language model component (in orange) is used to augment the input token representation in a traditional sequence tagging models (in grey).

CoVe = Contextual Word Vectors

В отличие от классических представлений выводим кодирование слова, зависящее от контекста (всего предложений)

Например, то что выучивает кодировщик в attentional seq-to-seq в NMT



https://www.topbots.com/generalized-language-models-cove-elmo/

CoVe = Contextual Word Vectors

CoVe(x) = MT-biLSTM(GloVe(x))

конкатенация скрытых состояний слова [h←, h→]

в изначальной работе предлагалось потом в задачах классификации конкатенировать [GloVe(x), CoVe(x)]

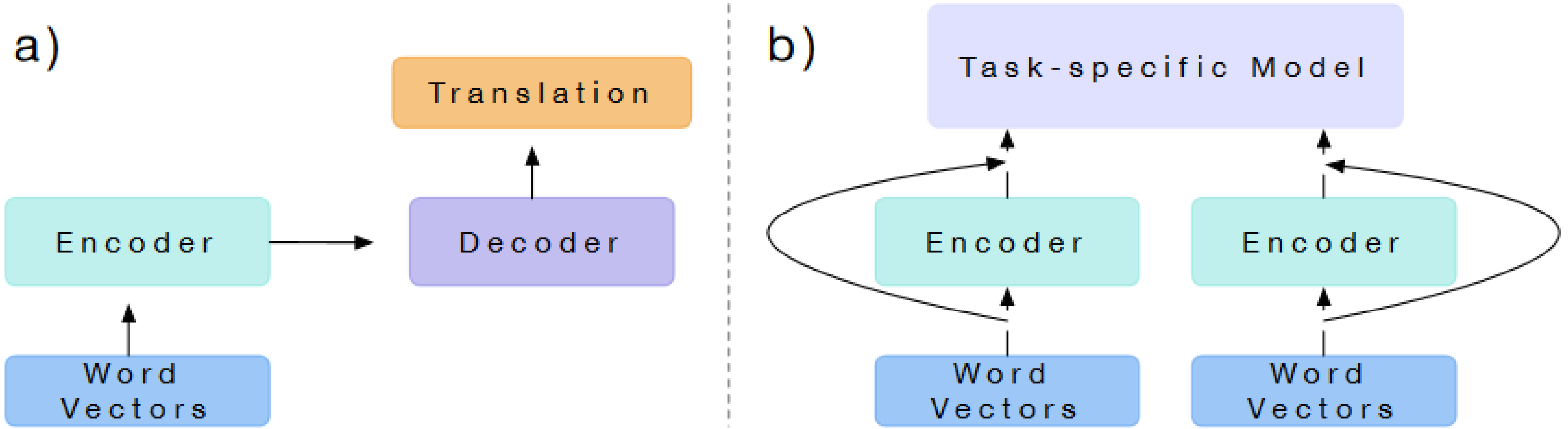
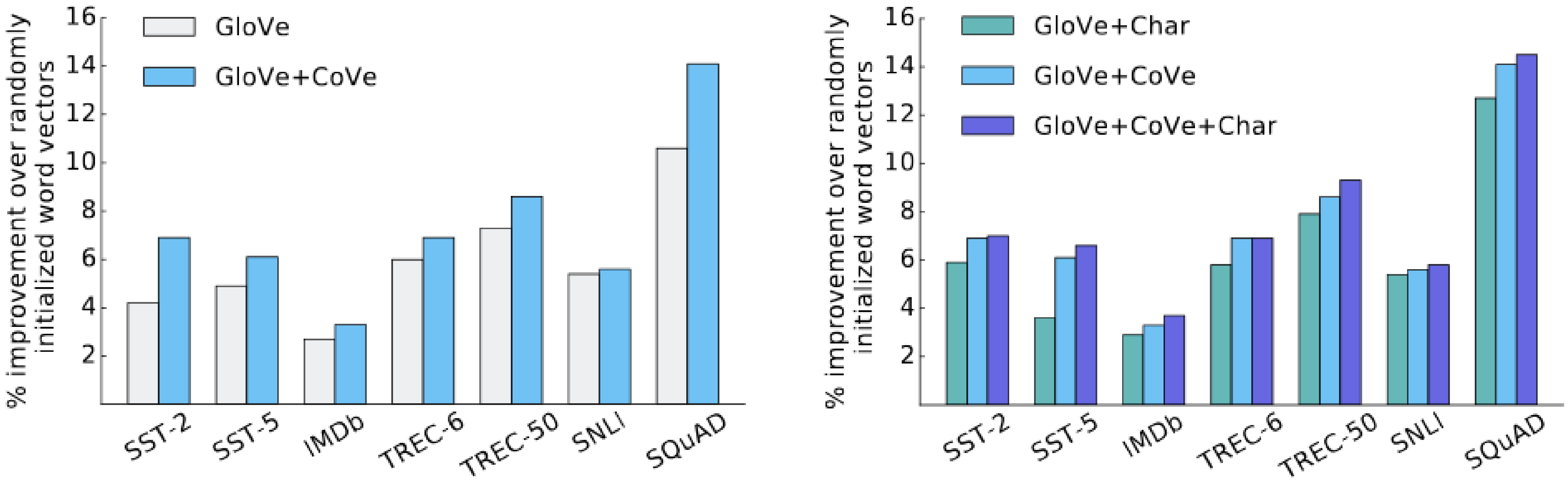


Figure 1: We a) train a two-layer, bidirectional LSTM as the encoder of an attentional sequence-to-sequence model for machine translation and b) use it to provide context for other NLP models.

термин введён в Bryan McCann et. al. «Learned in Translation: Contextualized Word Vectors»
// https://arxiv.org/pdf/1708.00107.pdf

CoVe = Contextual Word Vectors



(a) CoVe and GloVe (b) CoVe and Characters

Figure 3: The Benefits of CoVe

Char = character n-gram embeddings

результат не супер, как ожидалось…

м.б. машинный перевод более сложная задача, чем моделирование языка (что успешнее использовалось в других техниках)

ELMo: Embeddings from Language Models

представление с помощью предтренировки без учителя
biLM обучена на большом корпусе текстов

новое предложение в нашей задаче пропускается через biLM
представление слоя = лк состояний слова

$\Rightarrow$

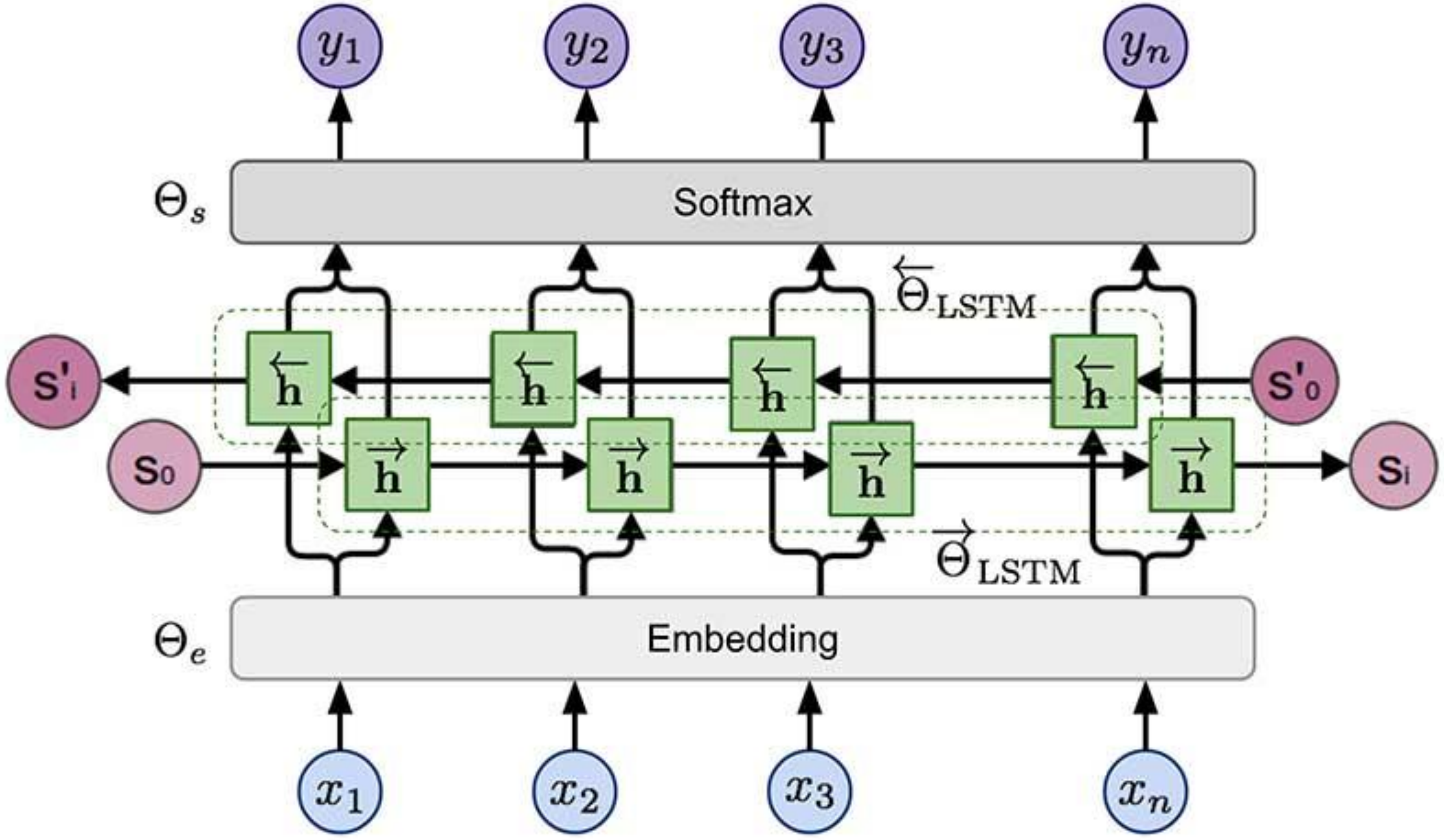
- **зависит от всего предложения**
- **глубокое (зависит от всех слоёв)**
- **есть возможность его обучать (т.к. лк)**

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer Deep contextualized word representations // https://arxiv.org/abs/1802.05365

ELMo: Embeddings from Language Models

строим biLM (Bidirectional language model):

$$\sum_k \log p(t_k \mid t_1,\ldots,t_{k-1};\Theta_x,\theta^{\rightarrow}_{\mathrm{LSTM}},\Theta_s)) + \log p(t_k \mid t_{k+1},\ldots,t_n;\Theta_x,\theta^{\leftarrow}_{\mathrm{LSTM}},\Theta_s))$$



Θ_x **– представление токенов**

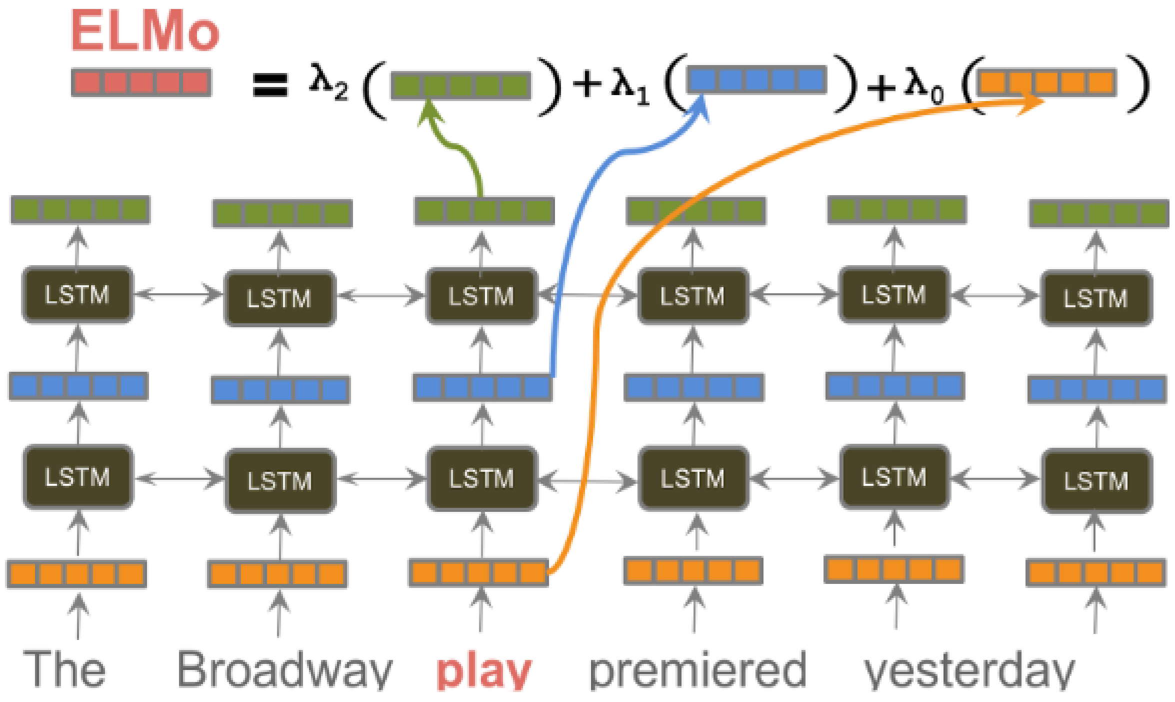
Θ_s **– softmax-слой**

https://www.topbots.com/generalized-language-models-cove-elmo/

ELMo: Embeddings from Language Models

$$\sum_k \log p(t_k \mid t_1,\ldots,t_{k-1};\Theta_x,\theta_{\mathrm{LSTM}}^{\rightarrow},\Theta_s)) + \log p(t_k \mid t_{k+1},\ldots,t_n;\Theta_x,\theta_{\mathrm{LSTM}}^{\leftarrow},\Theta_s))$$

можно затачивать представление под конкретную задачу –
– такую л/к скрытых состояний



$$\mathrm{ELMO}_k = \gamma^{\mathrm{task}} \sum_{l\in\mathrm{layers}} s_j^{\mathrm{task}} [\vec{h}_{k,j}^{\mathrm{LM}}, \overleftarrow{h}_{k,j}^{\mathrm{LM}}]$$

сюда ещё добавляют и выход embedding-слоя

разные слои – разный уровень абстракции
низкие ~ части речи
высокие ~ ответы на вопросы

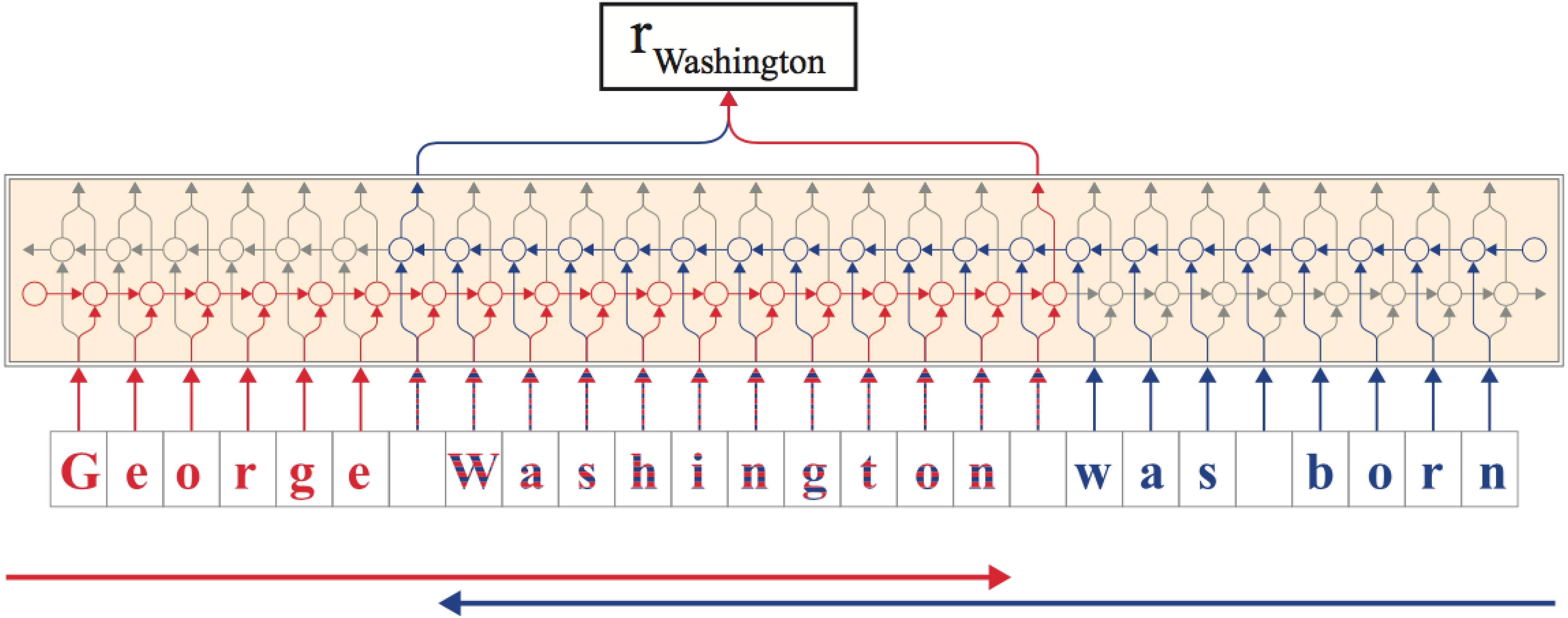
ELMo: Embeddings from Language Models

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

Table 4: Nearest neighbors to "play" using GloVe and the context embeddings from a biLM.

FLAIR: Contextual String Embeddings for Sequence Labelling

учим посимвольную двунаправленную LM (Character-level Language Model)
конкатенируем скрытое состояние последней буквы LM→, первой LM←



Alan Akbik, Duncan Blythe, Roland Vollgraf «Contextual String Embeddings for Sequence Labeling» https://www.aclweb.org/anthology/C18-1139/

FLAIR: Contextual String Embeddings for Sequence Labelling

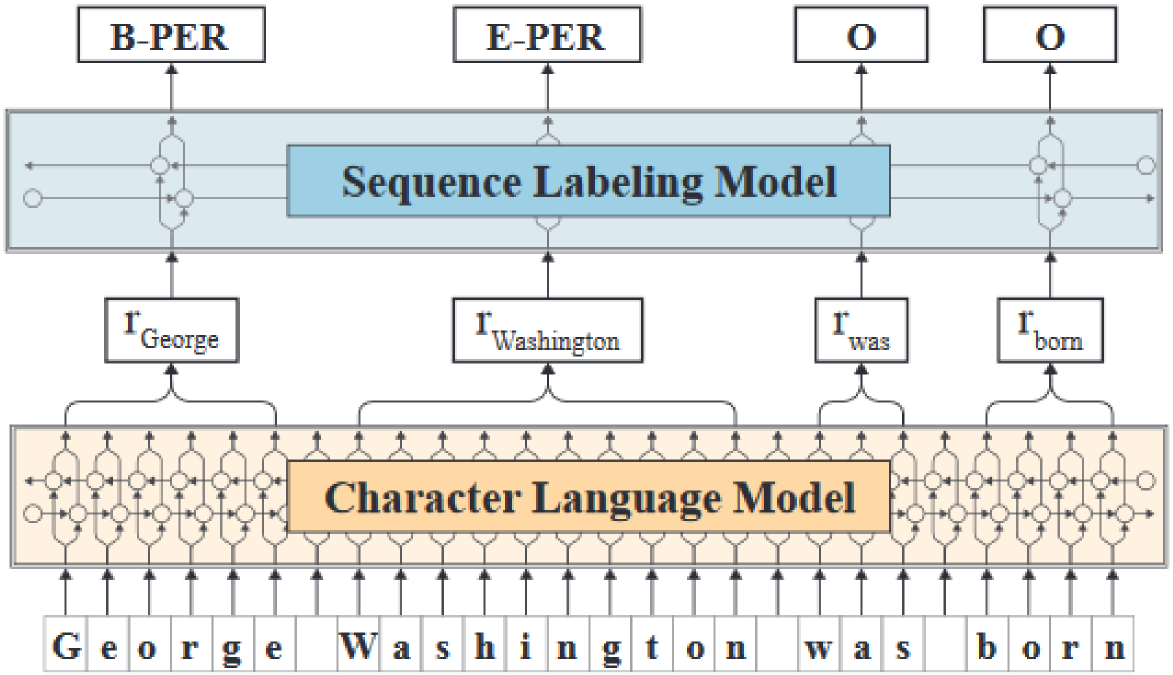


Figure 1: High level overview of proposed approach. A sentence is input as a character sequence into a pre-trained bidirectional character language model. From this LM, we retrieve for each word a contextual embedding that we pass into a vanilla BiLSTM-CRF sequence labeler, achieving robust state-of-the-art results on downstream tasks (NER in Figure).

FLAIR: Contextual String Embeddings for Sequence Labelling

word	context	selected nearest neighbors
Washington	(a) ***Washington*** *to curb support for [..]*	(1) ***Washington*** *would also take [..] action [..]* (2) ***Russia*** *to clamp down on barter deals [..]* (3) ***Brazil*** *to use hovercrafts for [..]*
Washington	(b) *[..] Anthony* ***Washington*** *(U.S.) [..]*	(1) *[..] Carla* ***Sacramento*** *(Portugal) [..]* (2) *[..] Charles* ***Austin*** *(U.S.) [..]* (3) *[..] Steve* ***Backley*** *(Britain) [..]*
Washington	(c) *[..] flown to* ***Washington*** *for [..]*	(1) *[..] while visiting* ***Washington*** *to [..]* (2) *[..] journey to New York City and* ***Washington*** *[..]* (14) *[..] lives in* ***Chicago*** *[..]*
Washington	(d) *[..] when* ***Washington*** *came charging back [..]*	(1) *[..] point for victory when* ***Washington*** *found [..]* (4) *[..] before* ***England*** *struck back with [..]* (6) *[..] before* ***Ethiopia*** *won the spot kick decider [..]*
Washington	(e) *[..] said* ***Washington*** *[..]*	(1) *[..] subdue the never-say-die* ***Washington*** *[..]* (4) *[..] a private school in* ***Washington*** *[..]* (9) *[..] said* ***Florida*** *manager John Boles [..]*

Table 4: Examples of the word “Washington” in different contexts in the CoNLL03 data set, and nearest neighbors using cosine distance over our proposed embeddings. Since our approach produces different embeddings based on context, we retrieve different nearest neighbors for each mention of the same word.

Совместное использование представлений

можно конкатенировать разные представления

использовать одни как инициализации для вычисления других

Другие решения

BERT

не просто контекст слева и справа
а сразу всё!

Раньше

Кот сидел на крыше около трубы

Потом

Кот сидел на крыше около трубы

Представление текстов

умеем представлять (вкладывать) слова
как быть с предложениями / абзацами / текстами?

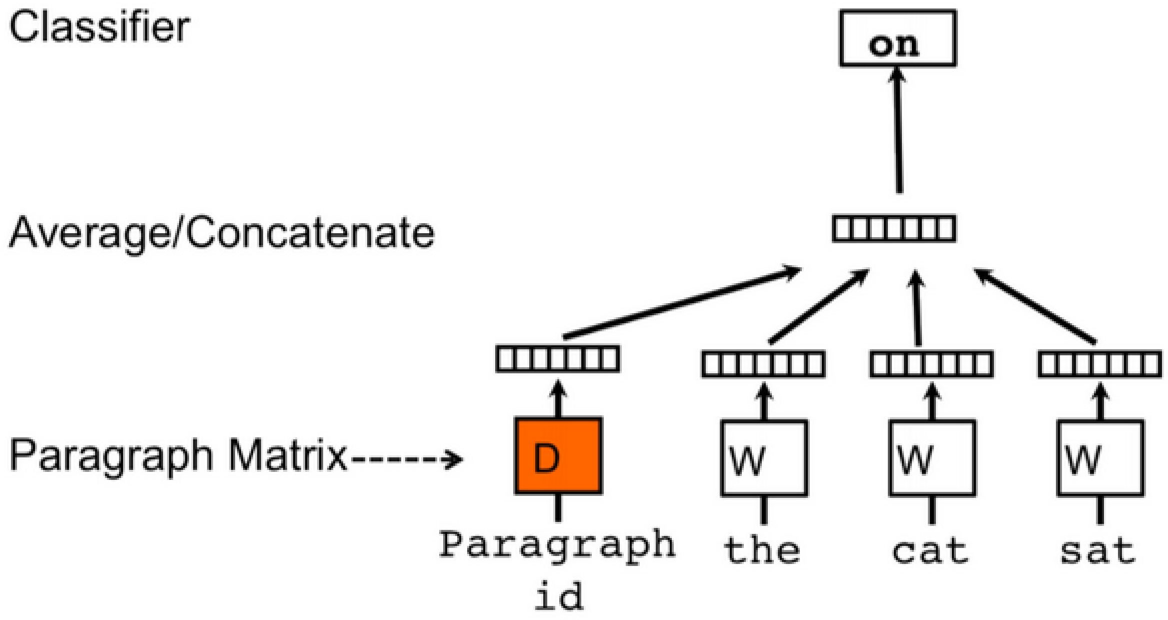
текст ~ «среднее» векторов входящих слов
~ сумма с весами – вероятностями слов

уже было в seq2seq

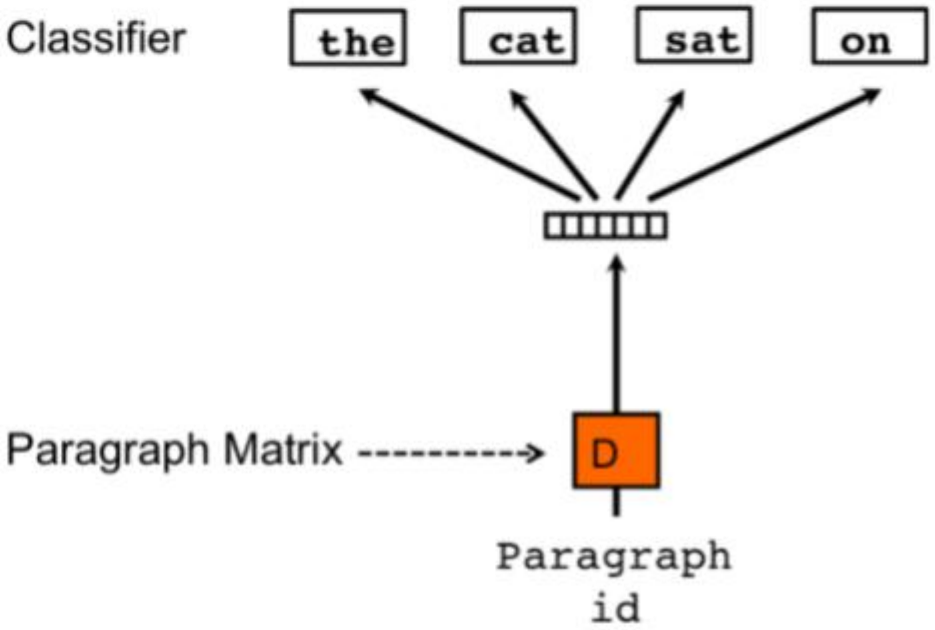
Представление текстов: Paragraph Vector (Doc2Vec / paragraph2vec)

По аналогии с word2 vec

PV-DM (Distributed Memory)



Distributed Bag Of Words (DBOW)



предсказываем случайно выбранные слова

Quoc V. Le, Tomas Mikolov Distributed Representations of Sentences and Documents // https://arxiv.org/abs/1405.4053

Представление предложений: The skip-thoughts model

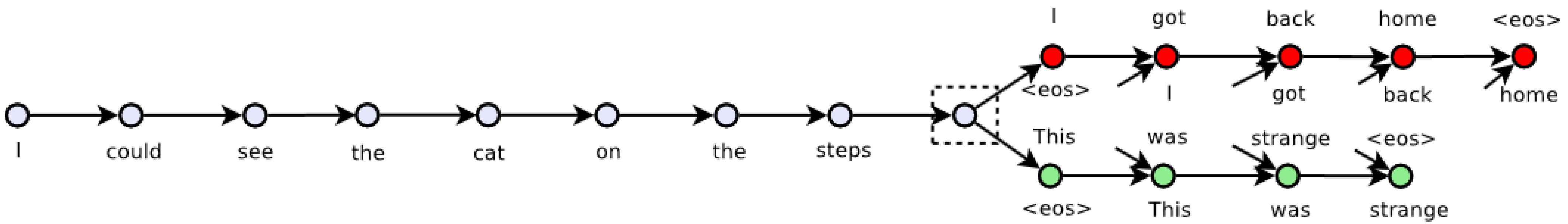


Figure 1: The skip-thoughts model. Given a tuple (s_{i-1}, s_i, s_{i+1}) of contiguous sentences, with s_i the i-th sentence of a book, the sentence s_i is encoded and tries to reconstruct the previous sentence s_{i-1} and next sentence s_{i+1}. In this example, the input is the sentence triplet *I got back home. I could see the cat on the steps. This was strange.* Unattached arrows are connected to the encoder output. Colors indicate which components share parameters. ⟨eos⟩ is the end of sentence token.

Последовательность предложений:

I got back home. I could see the cat on the steps. This was strange.

пытаемся по среднему предсказать первое и третье

один цвет – разделение параметров

$$\sum_t \log P(w_{i+1}^t | w_{i+1}^{<t}, \mathrm{h}_i) + \sum_t \log P(w_{i-1}^t | w_{i-1}^{<t}, \mathrm{h}_i)$$

кодировщик-декодировщик

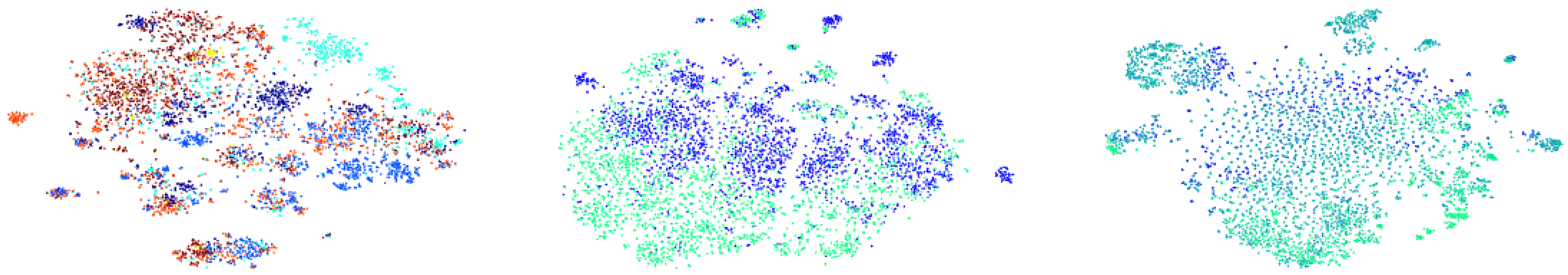
довольно долгий, но качество высокое

The skip-thoughts model: ближайшие соседи

Query and nearest sentence
he ran his hand inside his coat , double-checking that the unopened letter was still there . he slipped his hand between his coat and his shirt , where the folded copies lay in a brown envelope .
im sure youll have a glamorous evening , she said , giving an exaggerated wink . im really glad you came to the party tonight , he said , turning to her .
although she could tell he had n't been too invested in any of their other chitchat , he seemed genuinely curious about this . although he had n't been following her career with a microscope , he 'd definitely taken notice of her appearances .
an annoying buzz started to ring in my ears , becoming louder and louder as my vision began to swim . a weighty pressure landed on my lungs and my vision blurred at the edges , threatening my consciousness altogether .
if he had a weapon , he could maybe take out their last imp , and then beat up errol and vanessa . if he could ram them from behind , send them sailing over the far side of the levee , he had a chance of stopping them .
then , with a stroke of luck , they saw the pair head together towards the portaloos . then , from out back of the house , they heard a horse scream probably in answer to a pair of sharp spurs digging deep into its

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, Sanja Fidler Skip-Thought Vectors // https://arxiv.org/abs/1506.06726

The skip-thoughts model: ближайшие соседи



(a) TREC (b) SUBJ (c) SICK

Figure 2: t-SNE embeddings of skip-thought vectors on different datasets. Points are colored based on their labels (question type for TREC, subjectivity/objectivity for SUBJ). On the SICK dataset, each point represents a sentence pair and points are colored on a gradient based on their relatedness labels. Results best seen in electronic form.

Предтренировка автокодировщика (Autoencoder pretraining)

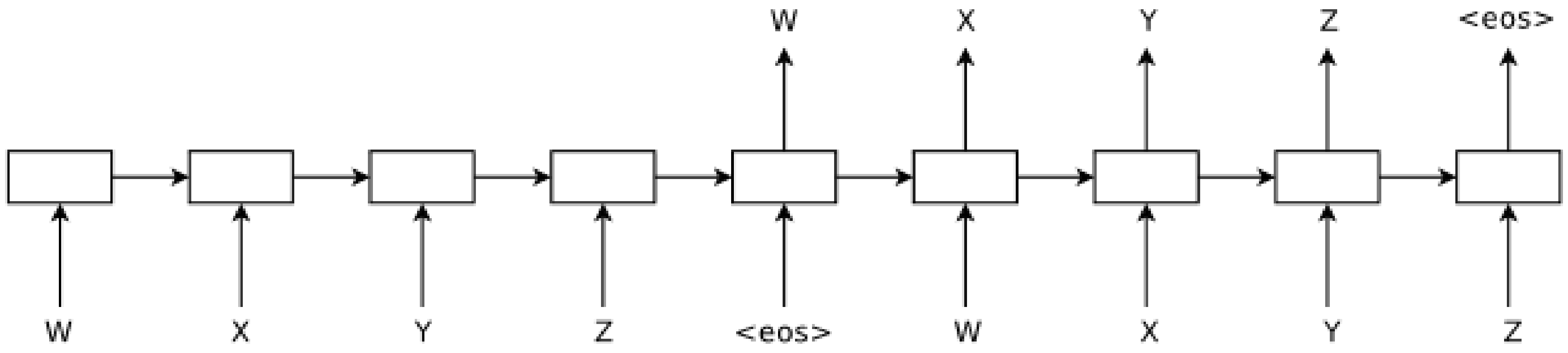


Figure 1: The sequence autoencoder for the sequence "WXYZ". The sequence autoencoder uses a recurrent network to read the input sequence in to the hidden state, which can then be used to reconstruct the original sequence.

хотим, чтобы автокодировщик воспроизводил входную последовательность!

Andrew M. Dai, Quoc V. Le «Semi-supervised Sequence Learning» // https://arxiv.org/abs/1511.01432

Supervised sentence embeddings

- **Paragram-phrase: uses paraphrase database for supervision, best for paraphrase and semantic similarity** (Wieting et al. 2016)

- **InferSent: bi-LSTM trained on SNLI + MNLI** (Conneau et al. 2017)

дальше есть…

- **GenSen: multitask training (skip-thought, machine translation, NLI, parsing)** (Subramanian et al. 2018)

Это рассказывать не будем!

Представление слов/предложений/текстов: StarSpace

название: $*\to$ «space» (пространтсво)

Метод оперирует с объектами, которые описываются наборами признаков из фиксированного множества

Пример: предложение = набор слов
(или = набор n-грамм)

$$\sum_{\substack{(a,b)\in K^+\\ b^-\in K^-}} L^{\text{batch}}(\text{sim}(a,b),\text{sim}(a,b_1^-),\ldots,\text{sim}(a,b_k^-))\to\min$$

~ генерация позитивных и негативных пар

под решение конкретной задачи

представление всех сущностей (документы, картинки и т.п.) в едином пространстве

Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, Jason Weston StarSpace: Embed All The Things! // https://arxiv.org/abs/1709.03856
https://github.com/facebookresearch/StarSpace

Представление слов/предложений/текстов: StarSpace

Input Query	StarSpace result	fastText result
She is the 1962 Blue Swords champion and 1960 Winter Universiade silver medalist.	**Article**: Eva Groajov. **Paragraph**: Eva Groajov , later Bergerov-Groajov , is a former competitive figure skater who represented Czechoslovakia. She placed 7th at the 1961 European Championships and 13th at the 1962 World Championships. She was coached by Hilda Mdra.	**Article**: Michael Reusch. **Paragraph**: Michael Reusch (February 3, 1914April 6 , 1989) was a Swiss gymnast and Olympic Champion. He competed at the 1936 Summer Olympics in Berlin, where he received silver medals in parallel bars and team combined exercises...
The islands are accessible by a one-hour speedboat journey from Kuala Abai jetty, Kota Belud, 80 km north-east of Kota Kinabalu, the capital of Sabah.	**Article**: Mantanani Islands. **Paragraph**: The Mantanani Islands form a small group of three islands off the north-west coast of the state of Sabah, Malaysia, opposite the town of Kota Belud, in northern Borneo. The largest island is Mantanani Besar; the other two are Mantanani Kecil and Lungisan...	**Article**: Gum-Gum **Paragraph**: Gum-Gum is a township of Sandakan, Sabah, Malaysia. It is situated about 25km from Sandakan town along Labuk Road.
Maggie withholds her conversation with Neil from Tom and goes to the meeting herself, and Neil tells her the spirit that contacted Tom has asked for something and will grow upset if it does not get done.	**Article**: Stir of Echoes **Paragraph**: Stir of Echoes is a 1999 American supernatural horror-thriller released in the United States on September 10 , 1999 , starring Kevin Bacon and directed by David Koepp . The film is loosely based on the novel ”A Stir of Echoes” by Richard Matheson...	**Article**: The Fabulous Five **Paragraph**: The Fabulous Five is an American book series by Betsy Haynes in the late 1980s . Written mainly for preteen girls , it is a spin-off of Haynes ’ other series about Taffy Sinclair...

Table 8: StarSpace predictions for some example Wikipedia Article Search (Task 1) queries where StarSpace is correct.

Представление слов/предложений/текстов: StarSpace

Task	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	SICK-R	SICK-E	STS14
Unigram-TFIDF*	73.7	79.2	90.3	82.4	-	85.0	73.6 / 81.7	-	-	0.58 / 0.57
ParagraphVec (DBOW)*	60.2	66.9	76.3	70.7	-	59.4	72.9 / 81.1	-	-	0.42 / 0.43
SDAE*	74.6	78.0	90.8	86.9	-	78.4	73.7 / 80.7	-	-	0.37 / 0.38
SIF(GloVe+WR)*	-	-	-	82.2	-	-	-	-	84.6	0.69 / -
word2vec*	77.7	79.8	90.9	88.3	79.7	83.6	72.5 / 81.4	0.80	78.7	0.65 / 0.64
GloVe*	78.7	78.5	91.6	87.6	79.8	83.6	72.1 / 80.9	0.80	78.6	0.54 / 0.56
fastText (public Wikipedia model)*	76.5	78.9	91.6	87.4	78.8	81.8	72.4 / 81.2	0.80	77.9	0.63 / 0.62
StarSpace [word]	73.8	77.5	91.53	86.6	77.2	82.2	73.1 / 81.8	0.79	78.8	0.65 / 0.62
StarSpace [sentence]	69.1	75.1	85.4	80.5	72.0	63.0	69.2 / 79.7	0.76	76.2	0.70 / 0.67
StarSpace [word + sentence]	72.1	77.1	89.6	84.1	77.5	79.0	70.2 80.3	0.79	77.8	0.69/0.66
StarSpace [ensemble w+s]	76.6	80.3	91.8	88.0	79.9	85.2	71.8 / 80.6	0.78	82.1	0.69 / 0.65

Table 9: Transfer test results on SentEval. * indicates model results that have been extracted from (Conneau et al. 2017). For MR, CR, SUBJ, MPQA, SST, TREC, SICK-R we report accuracies; for MRPC, we report accuracy/F1; for SICK-R we report Pearson correlation with relatedness score; for STS we report Pearson/Spearman correlations between the cosine distance of two sentences and human-labeled similarity score.

Представление предложений: Deep Averaging Network (DAN)

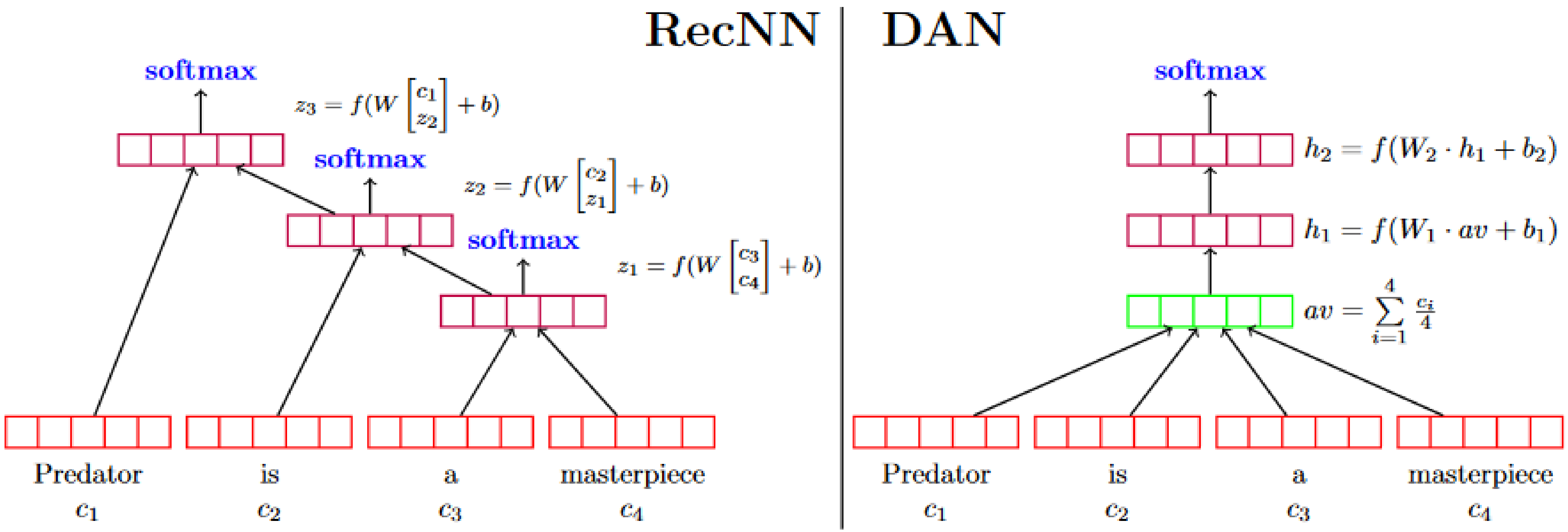


Figure 1: On the left, a RecNN is given an input sentence for sentiment classification. Softmax layers are placed above every internal node to avoid vanishing gradient issues. On the right is a two-layer DAN taking the same input. While the RecNN has to compute a nonlinear representation (purple vectors) for every node in the parse tree of its input, this DAN only computes two nonlinear layers for every possible input.

Простое усреднение…

Подумать – по сути это классификация

M. Iyyer, etc. Deep Unordered Composition Rivals Syntactic Methods for Text Classification, 2015 // http://www.aclweb.org/anthology/P15-1162

Представление предложений: Deep Averaging Network (DAN)

1. **Task**: map an input sequence of tokens X to one of k labels
2. **Composition** function g averages word embeddings:

$$z = g(w \in X) = \frac{1}{|X|} \sum_{w \in X} v_w,$$

   where v_w is a word embedding of word w
3. Estimate **probabilities** for each output label: $\hat{y} = \mathrm{softmax}(W_s \times z + b)$ and **predict** the label with highest probability
4. **Training**: minimize cross-entropy error: $\sum_{p=1}^{k} y_p \log \hat{y}_p$

Add more layers:

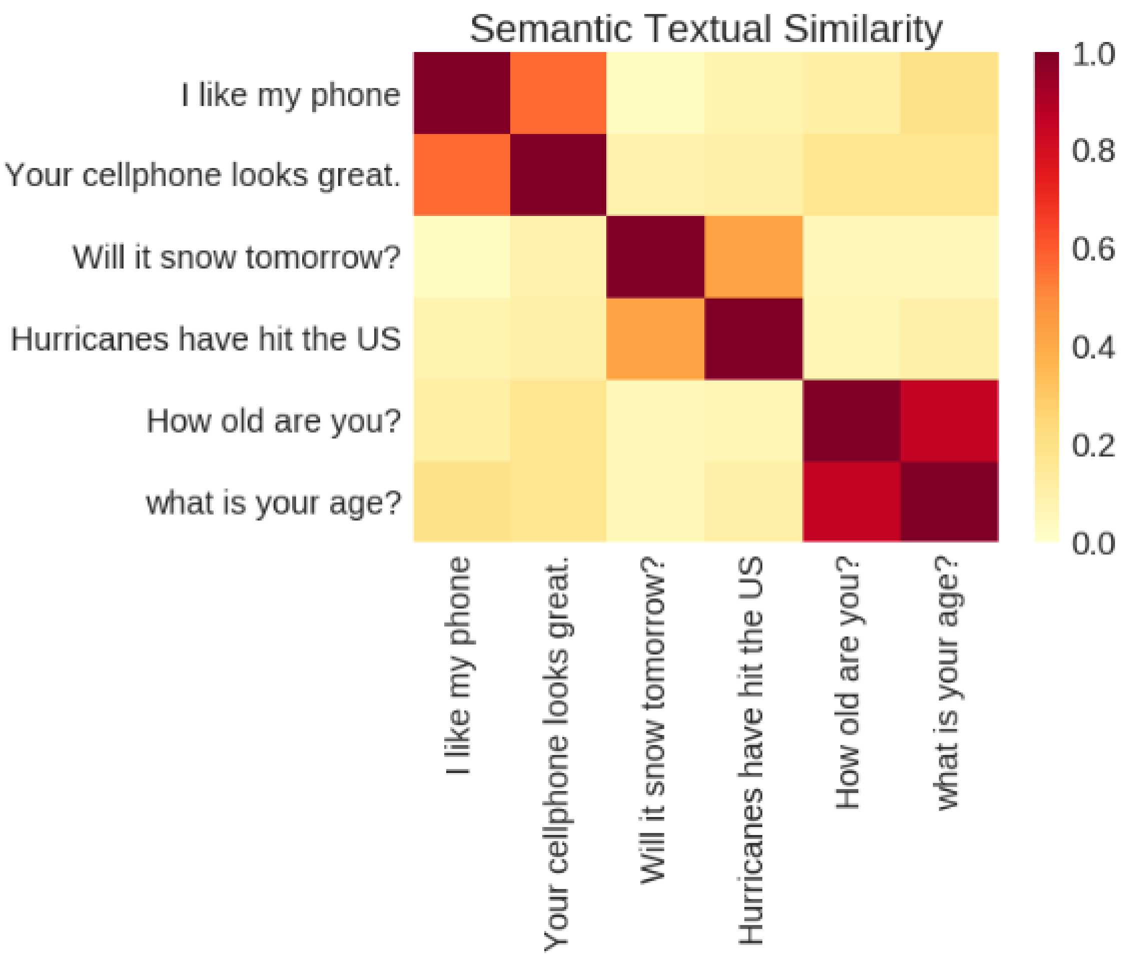
$$z_i = g(z_{i-1}) = f(W_i \times z_{i-1} + b_i)$$

Word dropout: drop word tokens' entire word embeddings from the vector average

Sentence	DAN	DRecNN	Ground Truth
a lousy movie that's not merely unwatchable, but also unlistenable	negative	negative	negative
if you're not a prepubescent girl, you'll be laughing at britney spears' movie-starring debut whenever it does n't have you impatiently squinting at your watch	negative	negative	negative
blessed with immense physical prowess he may well be, but ahola is simply not an actor	positive	neutral	negative
who knows what exactly godard is on about in this film, but his words and images do n't have to add up to mesmerize you.	positive	positive	positive
it's so good that its relentless, polished wit can withstand not only inept school productions, but even oliver parker's movie adaptation	negative	positive	positive
too bad, but thanks to some lovely comedic moments and several fine performances, it's not a total loss	negative	negative	positive
this movie was not good	negative	negative	negative
this movie was good	positive	positive	positive
this movie was bad	negative	negative	negative
the movie was not bad	negative	negative	positive

Table 3: Predictions of DAN and DRecNN models on real (top) and synthetic (bottom) sentences that contain negations and contrastive conjunctions. In the first column, words colored red individually predict the negative label when fed to a DAN, while blue words predict positive. The DAN learns that the negators *not* and *n't* are strong negative predictors, which means it is unable to capture double negation as in the last real example and the last synthetic example. The DRecNN does slightly better on the synthetic double negation, predicting a lower negative polarity.

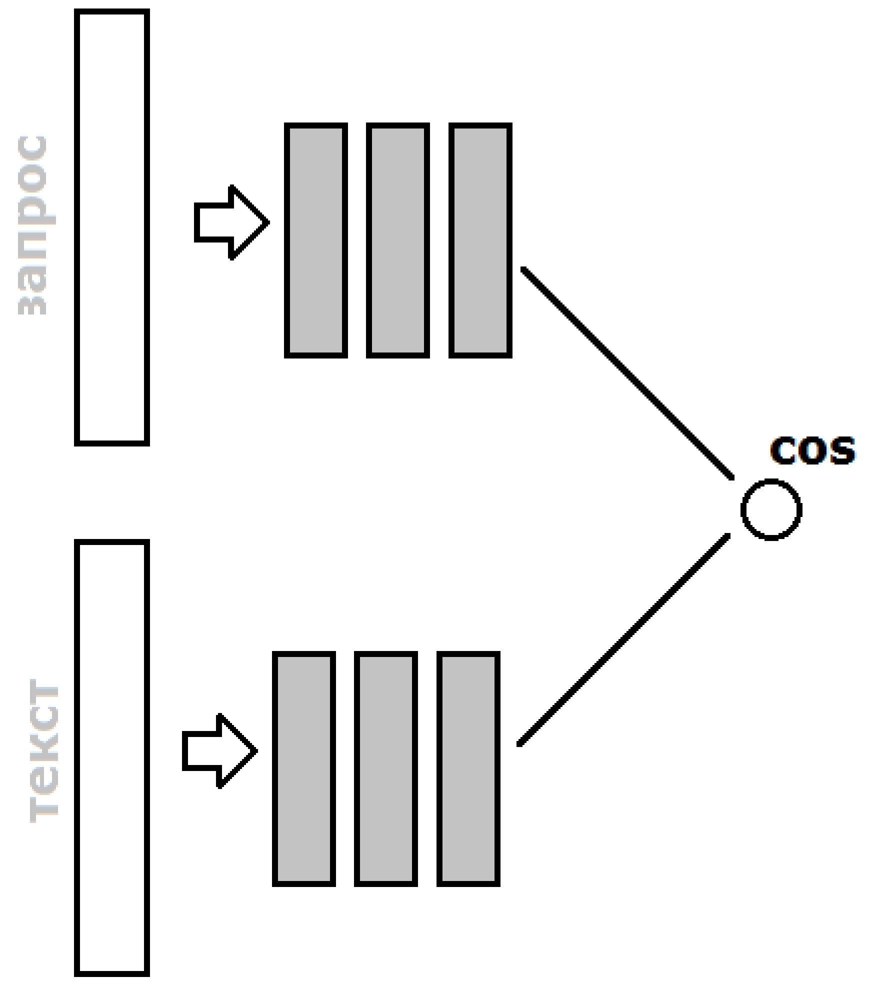
Universal Sentence Encoder



использовали 1) Transformer 2) DAN

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil Universal Sentence Encoder // https://arxiv.org/abs/1803.11175

DSSM = Deep Structured Semantic Model



сиамская сеть

DSSM = Deep Structured Semantic Model

https://www.researchgate.net/publication/262289160_Learning_deep_structured_semantic_models_for_web_search_using_clickthrough_data

вход – не только слова, но и n-граммы (вместе с ними – конкатенация)
https://habr.com/company/yandex/blog/314222

часто легко найти положительные примеры
отрицательные
1) берутся случайные – обучаются сети
2) берутся те, у которых высокая вероятность класса +, но они –
3) повторяется п. 2

Ещё подходы

Чем проще агрегация кодировок слов, тем нехуже

Dinghan Shen, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao, Lawrence Carin Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms // https://arxiv.org/abs/1805.09843

Обзор (полный, хороший)

Christian S. Perone, Roberto Silveira, Thomas S. Paula Evaluation of sentence embeddings in downstream and linguistic probing tasks // https://arxiv.org/abs/1806.06259

Table 6: Results from downstream classification tasks results using a MLP. Values in this table are accuracies for the test set.

Approach	CR	MPQA	MR	MRPC	SICK-E	SST-2	SST-5	SUBJ	TREC
Baseline									
Random Embedding	61.16	68.41	48.75	64.35	54.94	49.92	24.48	49.83	18.00
Experiments									
ELMo (BoW, all layers, 5.5B)	83.95	**91.02**	**80.91**	72.93	82.36	**86.71**	47.60	**94.69**	93.60
ELMo (BoW, all layers, original)	85.11	89.55	79.72	71.65	81.86	86.33	**48.73**	94.32	93.40
ELMo (BoW, top layer, original)	84.13	89.30	79.36	70.20	79.64	85.28	47.33	94.06	93.40
Word2Vec (BoW, google news)	79.23	88.24	77.44	73.28	79.09	80.83	44.25	90.98	83.60
p-mean (monolingual)	80.82	89.09	78.34	73.22	83.52	84.07	44.89	92.63	88.40
FastText (BoW, common crawl)	79.63	87.99	78.03	74.49	79.28	83.31	44.34	92.19	86.20
GloVe (BoW, common crawl)	78.67	87.90	77.63	73.10	79.01	81.55	45.16	91.48	84.00
USE (DAN)	80.50	83.53	74.03	71.77	80.39	80.34	42.17	91.93	89.60
USE (Transformer)	**86.04**	86.99	80.20	72.29	83.32	86.05	48.10	93.74	**93.80**
InferSent (AllNLI)	83.58	89.02	80.02	**74.55**	**86.44**	83.91	47.74	92.41	89.80
SkipThought	81.03	87.06	76.60	73.22	84.33	81.77	44.80	93.33	91.00

Общий подход и случайный кодировщик

Вложение предложения ищется в виде $h = f_\theta(e_1, \ldots, e_n)$

$e_1, \ldots, e_n$ **– вложения слов. Обучаем параметры** θ**.**

IferSent	$\max(\mathrm{BiLSTM}(e_1, \ldots, e_n))$ **Обучаем предсказывая метки «entailment», «neutral», «contradictive»** **cross-entropy**
SkipThought	$\mathrm{GRU}_n(e_1, \ldots, e_n)$ **Декодируем следующее и предыдущее** **negative log-likelihood**
Случайные кодировщики	
BOPER	$\mathrm{pool}(We_1, \ldots, We_n)$ $W \in \mathrm{rand}([-1/\sqrt{d}, +1/\sqrt{d}] \mid \mathbb{R}^{D \times d})$
RANDOM LSTM	$\mathrm{pool}(\mathrm{random\_BiLSTM}(e_1, \ldots, e_n))$
Echo State Networks (ESNs)	$\max(\mathrm{ESN}(e_1, \ldots, e_n))$

Случайный кодировщик не сильно хуже!

Model	Dim	MR	CR	MPQA	SUBJ	SST2	TREC	SICK-R	SICK-E	MRPC	STSB
BOE	300	77.3(.2)	78.6(.3)	87.6(.1)	91.3(.1)	80.0(.5)	81.5(.8)	80.2(.1)	78.7(.1)	72.9(.3)	70.5(.1)
BOREP	4096	77.4(.4)	79.5(.2)	88.3(.2)	91.9(.2)	81.8(.4)	**88.8(.3)**	85.5(.1)	82.7(.7)	73.9(.4)	68.5(.6)
RandLSTM	4096	77.2(.3)	78.7(.5)	87.9(.1)	91.9(.2)	81.5(.3)	86.5(1.1)	85.5(.1)	81.8(.5)	**74.1(.5)**	72.4(.5)
ESN	4096	**78.1(.3)**	**80.0(.6)**	**88.5(.2)**	**92.6(.1)**	**83.0(.5)**	87.9(1.0)	**86.1(.1)**	**83.1(.4)**	73.4(.4)	**74.4(.3)**
InferSent-1 = paper, G	4096	81.1	86.3	90.2	92.4	84.6	88.2	88.3	86.3	76.2	75.6
InferSent-2 = fixed pad, F	4096	79.7	84.2	89.4	92.7	84.3	90.8	88.8	86.3	76.0	78.4
InferSent-3 = fixed pad, G	4096	79.7	83.4	88.9	92.6	83.5	90.8	88.5	84.1	76.4	77.3
Δ InferSent-3, BestRand	-	*1.6*	*3.4*	*0.4*	*0.0*	*0.5*	*2.0*	*2.4*	*1.0*	*2.3*	*2.9*
ST-LN	4800	79.4	83.1	89.3	93.7	82.9	88.4	85.8	79.5	73.2	68.9
Δ ST-LN, BestRand	-	*1.3*	*3.1*	*0.8*	*1.1*	*-0.1*	*0.5*	*-0.3*	*-3.6*	*-0.9*	*-5.5*

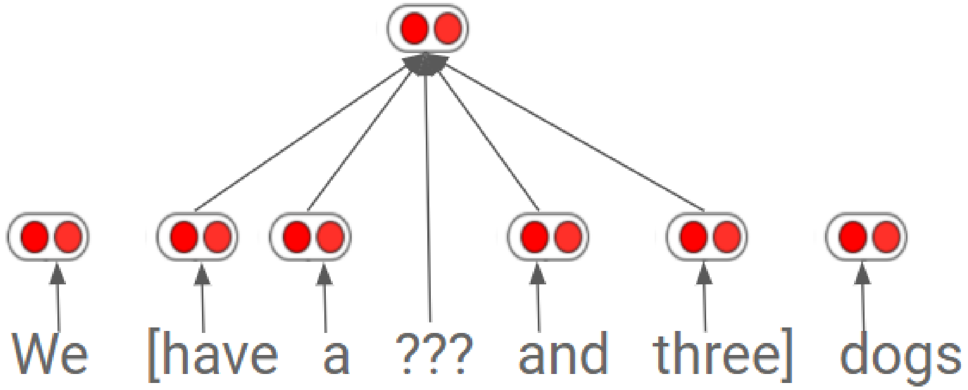
Table 1: Performance (accuracy for all tasks except SICK-R and STSB, for which we report Pearson's r) on all ten downstream tasks where all models have 4096 dimensions with the exception of BOE (300) and ST-LN (4800). Standard deviations are show in parentheses. InferSent-1 is the paper version with GloVe (G) embeddings, InferSent-2 has fixed padding and uses FastText (F) embeddings, and InferSent-3 has fixed padding and uses GloVe embeddings. We also show the difference between the best random architecture (BestRand) and InferSent-3 and ST-LN, respectively. The average performance difference between the best random architecture and InferSent-3 and ST-LN is 1.7 and -0.4 respectively.

John Wieting, Douwe Kiela No Training Required: Exploring Random Encoders for Sentence Classification

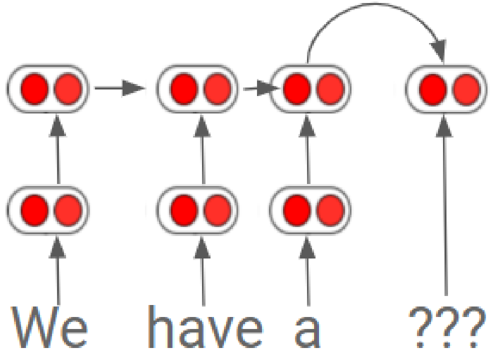
https://arxiv.org/abs/1901.10444

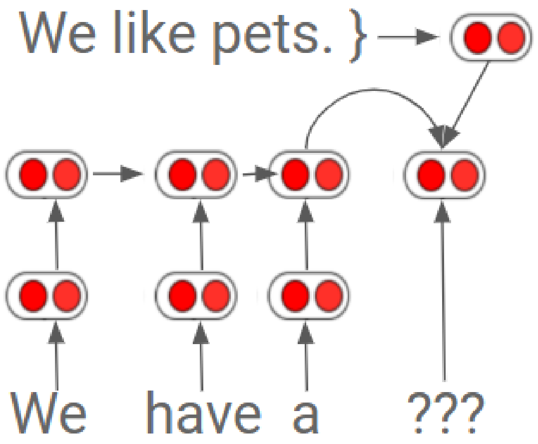
Итог

word2vec, Mikolov et al (2013)



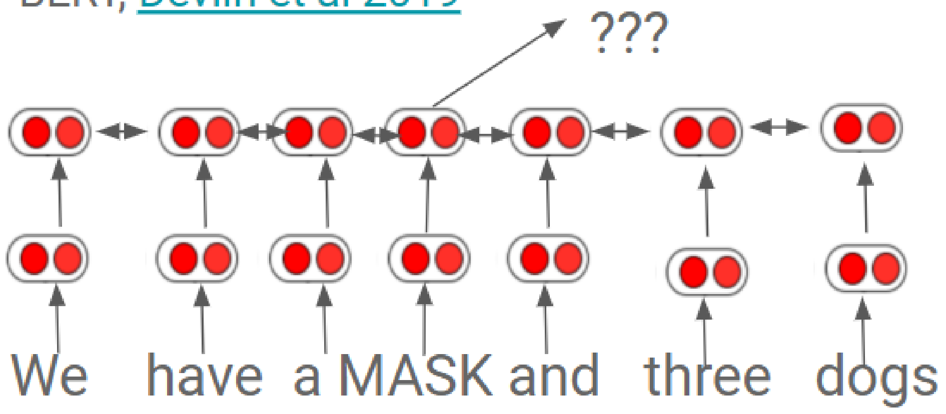
ELMo, Peters et al. 2018, ULMFiT (Howard & Ruder 2018), GPT (Radford et al. 2018)





Skip-Thought (Kiros et al., 2015)

BERT, Devlin et al 2019



http://tiny.cc/NAACLTransfer

Итог

Есть классические испытанные способы

Они используются и для получения более продвинутых представлений

Есть способы учёта контекста

дальше будем ещё с этим работать

Можно получать представления целых предложений / текстов

Ссылки

Поддерживаемый каталог представлений

https://github.com/Separius/awesome-sentence-embedding

хорошо тонкости методов расписаны

https://lilianweng.github.io/lil-log/2017/10/15/learning-word-embedding.html