Математические методы анализа текстов

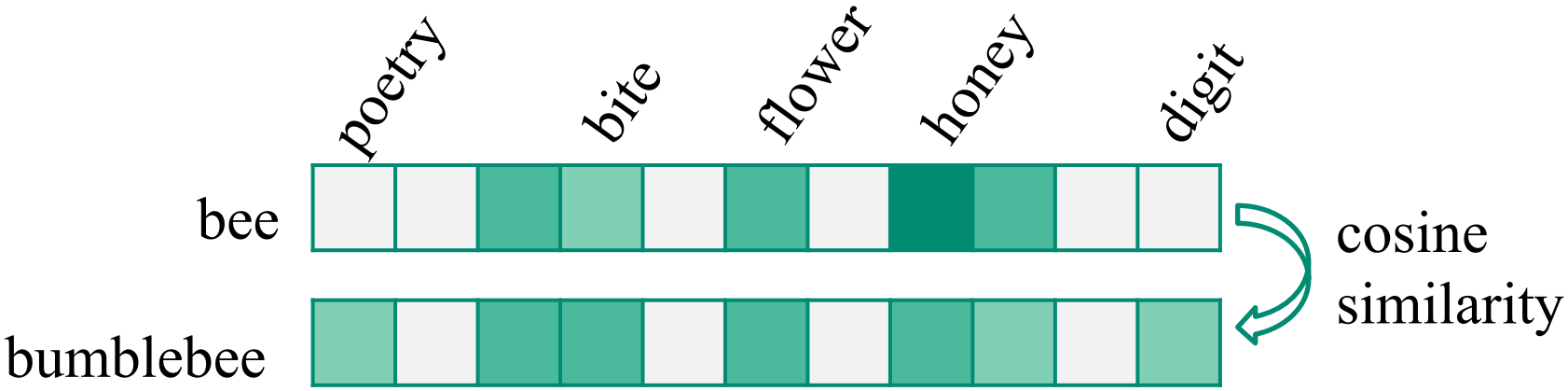
Векторные представления слов (word embeddings)

Потапенко Анна Александровна

24 октября 2018

Семантическая близость слов

- Co-встречаемости первого порядка
  syntagmatic associates / relatedness (bee and honey)
- Co-встречаемости второго порядка
  paradigmatic parallels / similarity (bee and bumblebee)



Schutze, H., & Pedersen, J. (1993). A vector model for syntagmatic and paradigmatic relatedness. In Making Sense of Words: Proceedings of the Conference, pp. 104-113, Oxford, England.

Дистрибутивная гипотеза

“You shall know a word by the company it keeps.”
— Firth, 1957.

- Скользящее окно фиксированной ширины
- Счетчики со-встречаемости n_{uv}

Дистрибутивная гипотеза

“You shall know a word by the company it keeps.”
— Firth, 1957.

- Скользящее окно фиксированной ширины
- Счетчики со-встречаемости n_{uv}
- **Лучше:** Pointwise Mutual Information:

$$PMI = log \frac{p(u,v)}{p(u)p(v)} = log \frac{n_{uv} n}{n_u n_v}$$

Дистрибутивная гипотеза

"You shall know a word by the company it keeps."
— Firth, 1957.

- Скользящее окно фиксированной ширины
- Счетчики со-встречаемости n_{uv}
- **Лучше:** Pointwise Mutual Information:

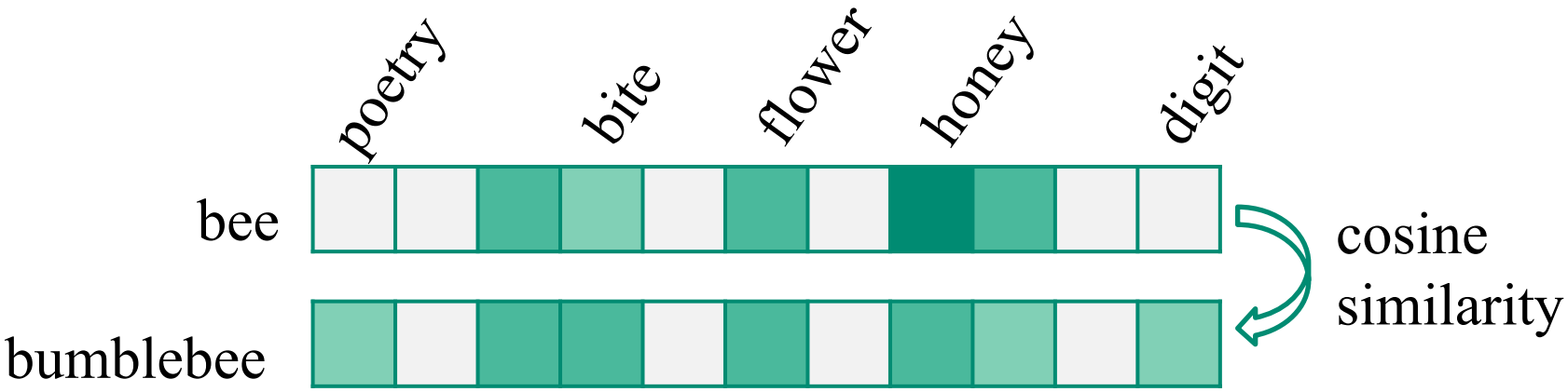
$$PMI = log\frac{p(u,v)}{p(u)p(v)} = log\frac{n_{uv}n}{n_u n_v}$$

- **Еще лучше:** positive Pointwise Mutual Information:

$$pPMI = \max(0, PMI)$$

Проблемы?

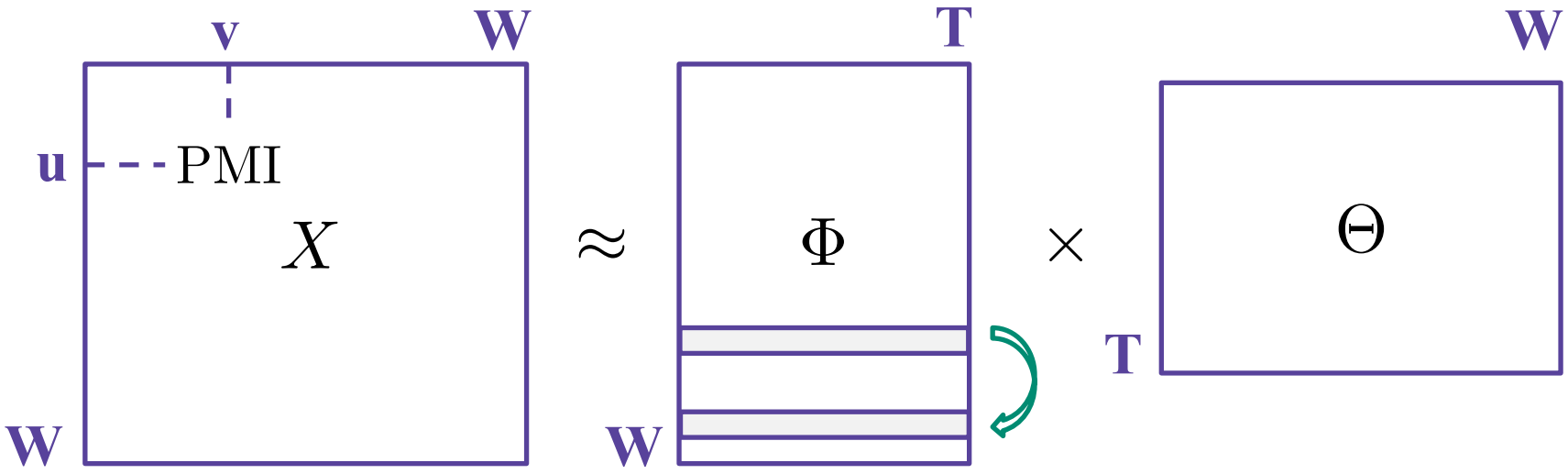
- Co-встречаемости первого порядка
  syntagmatic associates / relatedness (bee and honey)
- Co-встречаемости второго порядка
  paradigmatic parallels / similarity (bee and bumblebee)



Schutze, H., & Pedersen, J. (1993). A vector model for syntagmatic and paradigmatic relatedness. In Making Sense of Words: Proceedings of the Conference, pp. 104-113, Oxford, England.

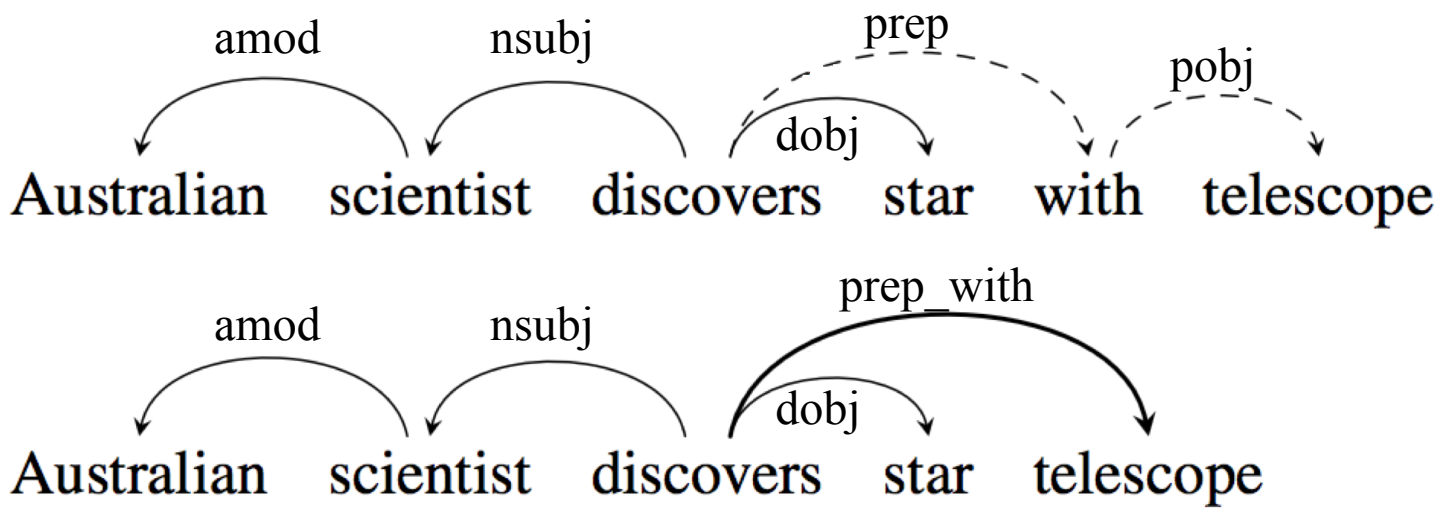
Векторные модели семантики

- **Вход:** со-встречаемости слов (counts, PMI, …)
- **Метод:** понижение размерности (SVD, …)
- **Выход:** близость между векторами слов



Turnay, P.D., Pantel, P.: from Frequency to Meaning: Vector Space Models of Semantics, 2010.

Что такое контекст?

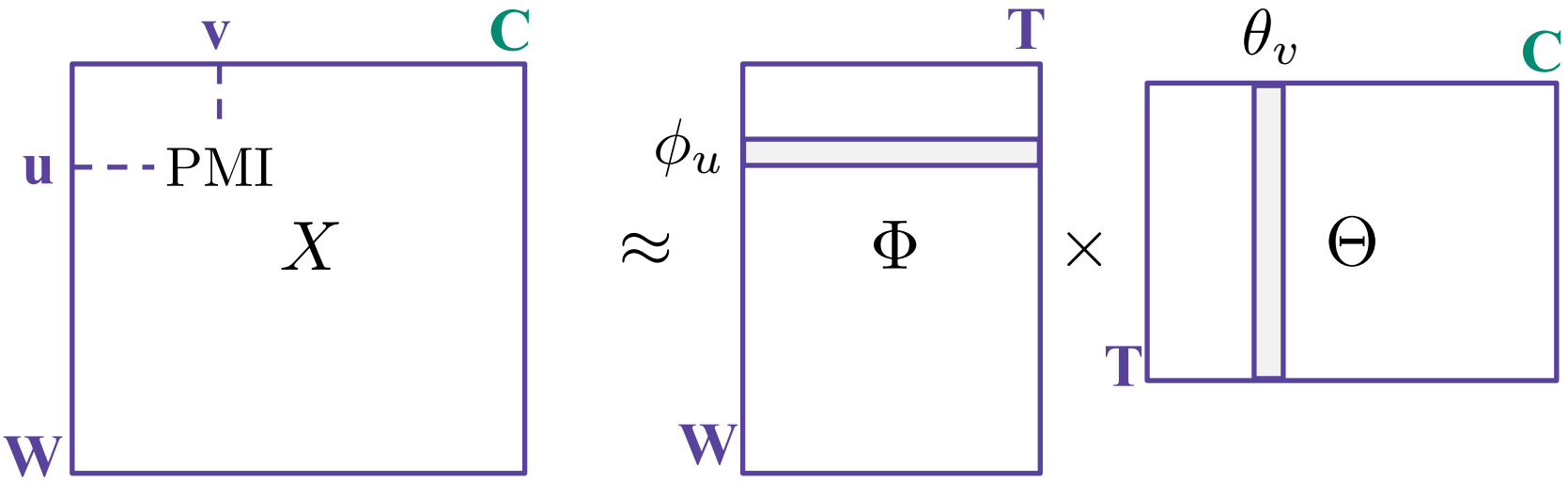


WORD	CONTEXTS
australian	scientist/amod^{-1}
scientist	australian/amod, discovers/nsubj^{-1}
discovers	scientist/nsubj, star/dobj, telescope/prep_with
star	discovers/dobj^{-1}
telescope	discovers/prep_with^{-1}

Omer Levy, Yoav Goldber, Dependency-Based Word Embeddings, ACL-2014.

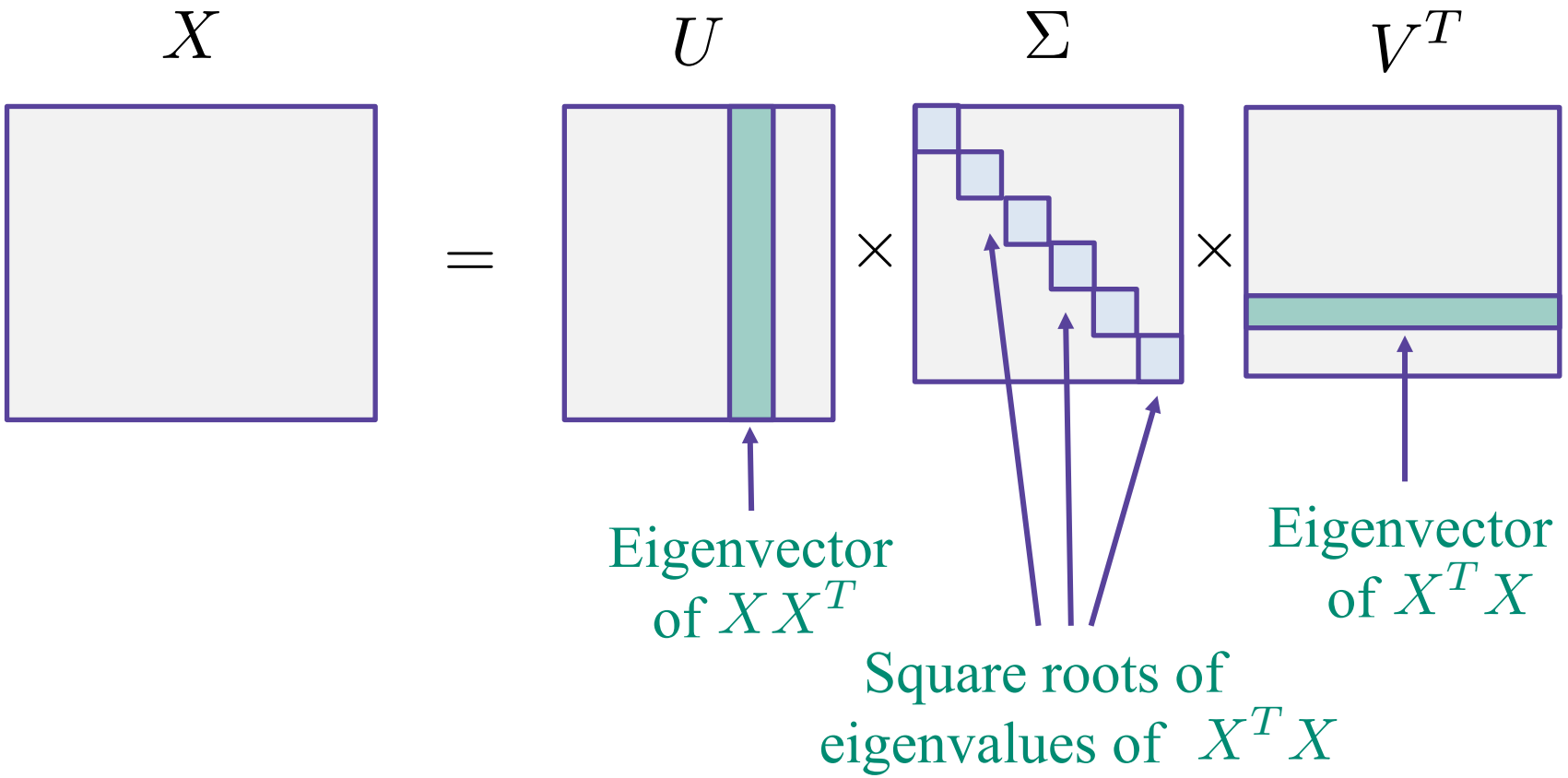
Что такое контекст?

- **C** – это словарь контекстов (напр., australian/amod)
- Но обычно контексты это слова из окна
- Тогда **W = C** и X – это симметричная матрица



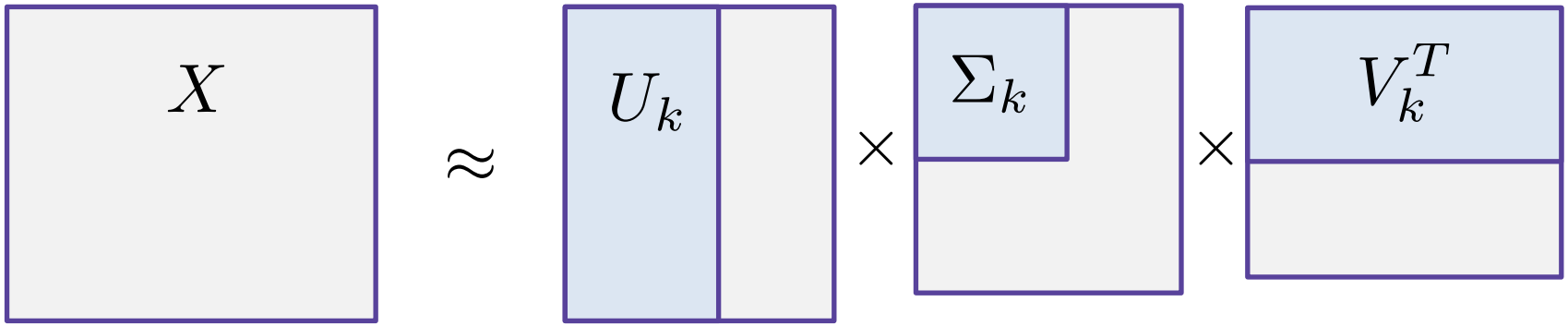
Низкоранговые матричные разложения

Сингулярное разложение (SVD)



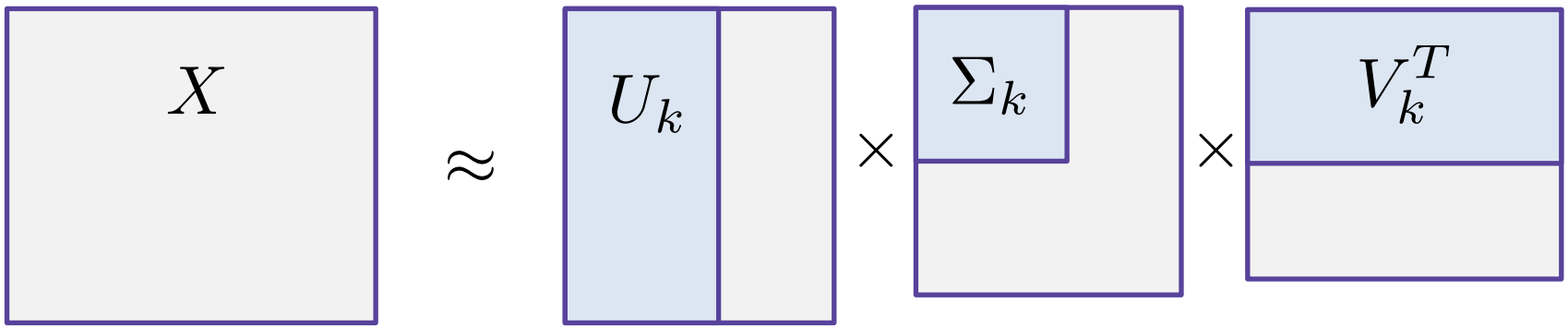
Truncated SVD

Оставляем только k компонент $\hat{X}_k = U_k \Sigma_k V_k^T$



Truncated SVD

Оставляем только k компонент $\hat{X}_k = U_k \Sigma_k V_k^T$

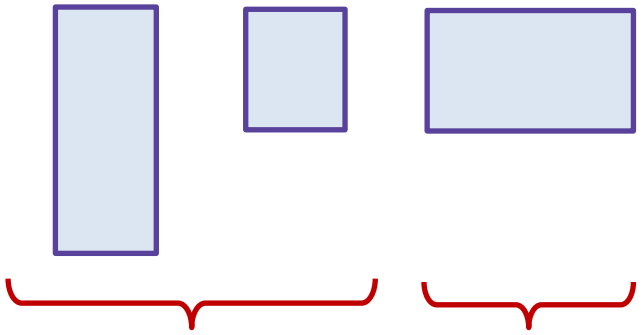


Лучшая аппроксимация ранга k в смысле нормы Фробениуса:

$$||X - \hat{X}||_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \hat{x}_{ij})^2}$$

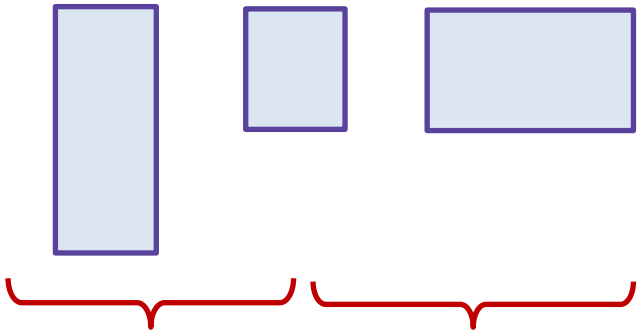
Как использовать SVD разложение?

Опция 1:



$$\Phi = U_k \Sigma_k \quad \Theta = V_k^T$$

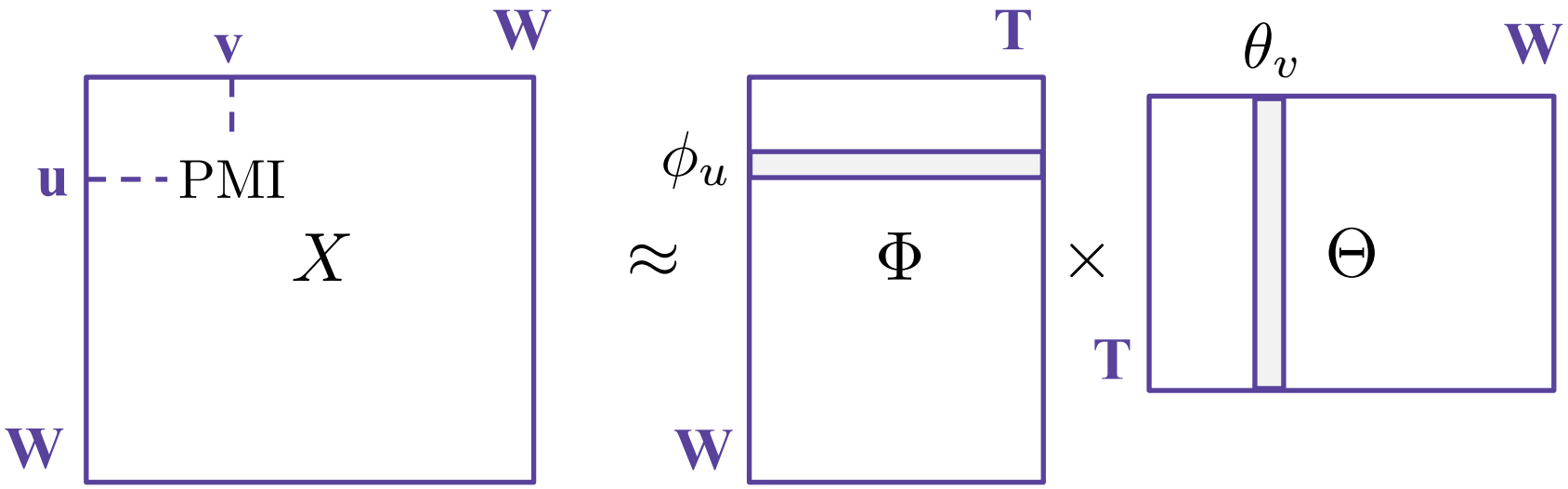
Опция 2:



$$\Phi = U_k \sqrt{\Sigma_k} \quad \Theta = \sqrt{\Sigma_k} V_k^T$$

Векторные модели семантики

- **Вход:** со-встречаемости слов (counts, PMI, …)
- **Метод:** понижение размерности (SVD, …)
- **Выход:** близость между векторами слов

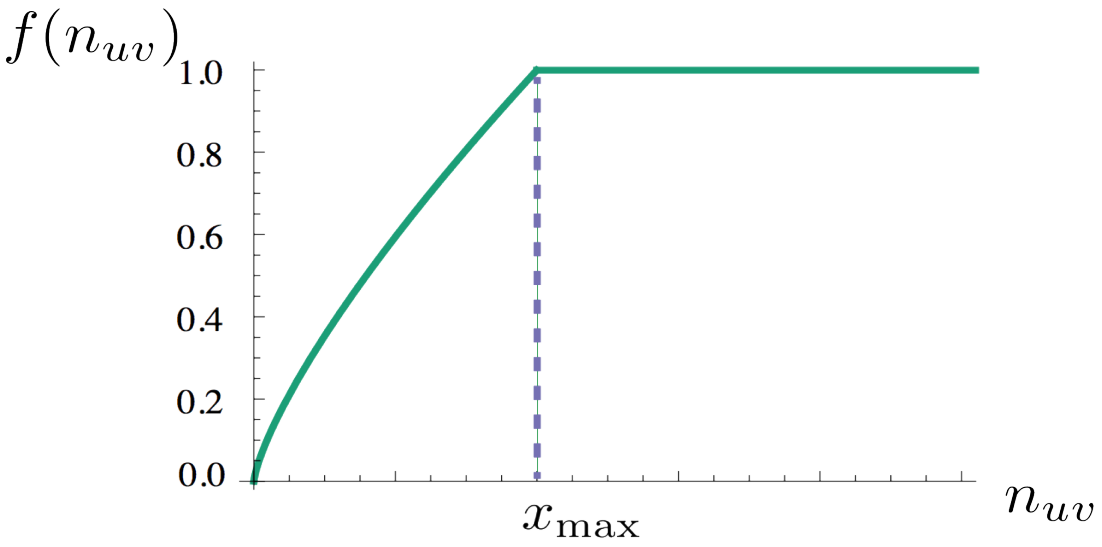


Turnay, P.D., Pantel, P.: from Frequency to Meaning: Vector Space Models of Semantics, 2010.

Взвешенные квадратичные потери: GloVe

Составим X из $\log n_{uv}$ и рассмотрим функцию потерь:

$$\sum_{u \in W} \sum_{v \in W} f(n_{uv}) \left(\langle \phi_u, \theta_v \rangle + b_u + b'_v - \log n_{uv} \right)^2 \to \min_{\phi_u, \theta_v, b_u, b'_v}$$



Pennington et. al. GloVe: Global Vectors for Word Representation, 2014.

Языковая модель: Skip-Gram

Предсказываем *контексты* по слову:

$$p(w_{i-h}, \dots w_{i+h}|w_i) = \prod_{-h \le k \le h,\, k \ne 0} p(w_{i+k}|w_i)$$

Каждую вероятность моделируем как *softmax*:

$$p(u|v) = \frac{\exp \langle \phi_u, \theta_v \rangle}{\sum_{u' \in W} \exp \langle \phi_{u'}, \theta_v \rangle}$$

Две матрицы параметров (как и в других методах).

Как обучать модель?

Максимизация *лог-правдоподобия*:

$$\mathcal{L} = \sum_{u \in W} \sum_{v \in W} n_{uv} \log p(u|v)$$

со-встречаемость

Метод:

- SGD по парам слов в корпусе

Проблема:

- *softmax* по словарю считать медленно.

Skip-gram Negative Sampling (SGNS)

Вместо того чтобы предсказывать слова,

будем предсказывать «да» и «нет» для пар слов:

$$\sum_{u \in W} \sum_{v \in W} n_{uv} \log \sigma \left(\langle \phi_u, \theta_v \rangle \right) +$$

$$k \, \mathbb{E}_{\bar{v}} \log \sigma \left(- \langle \phi_u, \theta_{\bar{v}} \rangle \right) \to \max_{\phi_u, \theta_v}$$

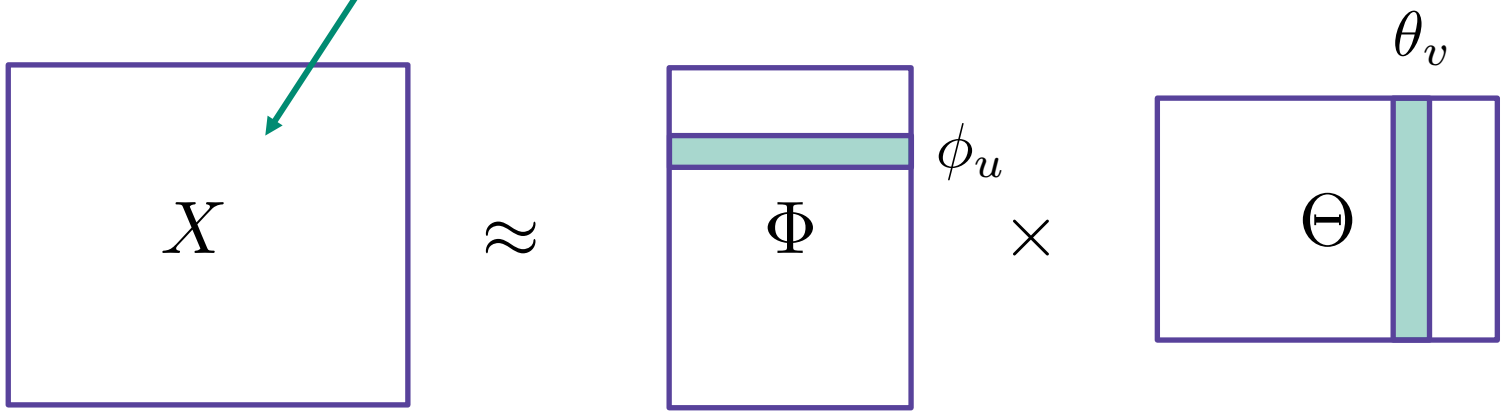
- Положительные примеры: со-встретившиеся пары u , v
- Отрицательные примеры: k случайных $\bar{v}$ для каждого u

Обучается все еще стохастическим градиентным спуском.

SGNS как неявное матричное разложение

Функция потерь в SGNS принимает максимум, когда $\langle \phi_u, \theta_v \rangle$ равно shifted Pointwise Mutual Information:

$$\text{sPMI} = \log \frac{n_{uv} n}{n_u n_v} - \log k$$



Levy and Goldberg. Neural Word Embedding as Implicit Matrix Factorization, 2014.

Оценивание моделей

Оценивание моделей: близость слов

Как проверить что *близкие слова* имеют *близкие вектора*?

- Лингвисты выделяют несколько типов близости
- Можно использовать *экспертные оценки* пар слов.
- Сравнивать *корреляцию Спирмена* для двух списков:

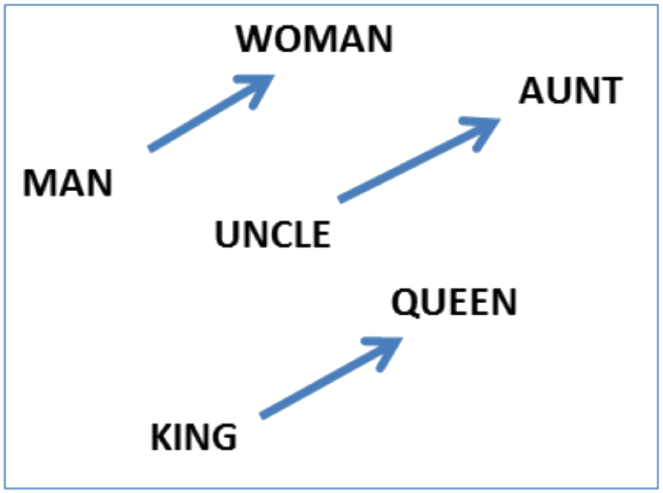
tiger	tiger	10.00
media	radio	7.42
tiger	cat	7.37
train	car	6.31
...	…	…

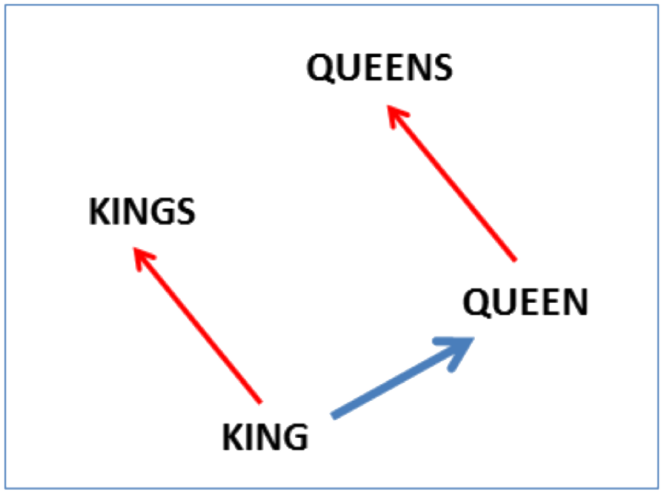
tiger	tiger	$\cos(\phi_u, \phi_v)$
media	radio	…
tiger	cat	…
train	car	…
...	…	…

Оценивание моделей: задача аналогий

- В когнитивных науках известна как *relational similarity* (vs. *attributional similarity*).
- a : a' как b : b' (man : woman как king : ?)

$$\cos(b - a + a', x) \to \max_x$$





Gentner, D. Structure-mapping: A theoretical framework for analogy. Cognitive Science, 1983.
Mikolov et. al. Linguistic Regularities in Continuous Space Word Representations, 2013.

Сравнение моделей

- *Count-based* подходы не уступают *predictive* подходам

win	Method	WordSim Similarity	WordSim Relatedness	Bruni et al. MEN	Radinsky et al. M. Turk
2	PPMI	.732	**.699**	.744	.654
	SVD	.772	.671	**.777**	.647
	SGNS	**.789**	.675	.773	**.661**
	GloVe	.720	.605	.728	.606
5	PPMI	.732	**.706**	.738	**.668**
	SVD	.764	.679	**.776**	.639
	SGNS	**.772**	.690	.772	.663
	GloVe	.745	.617	.746	.631

Levy et. al. Improving distributional similarity with lessons learned from word embeddings, 2015.

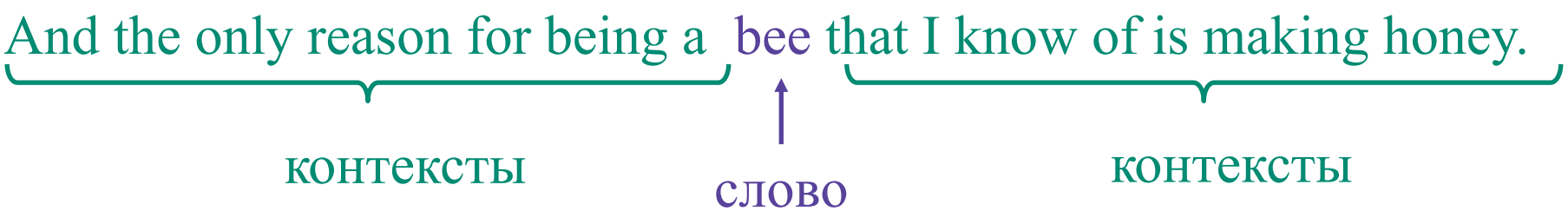
Сравнение моделей

- Задача аналогий решается с точностью 70%

win	Method	Google Add / Mul	MSR Add / Mul
2	PPMI	.552 / .677	.306 / .535
	SVD	.554 / .591	.408 / .468
	SGNS	.676 / **.689**	.617 / **.644**
	GloVe	.649 / .666	.540 / .591
5	PPMI	.518 / .649	.277 / .467
	SVD	.532 / .569	.369 / .424
	SGNS	.692 / **.714**	.605 / **.645**
	GloVe	.700 / .712	.541 / .599

Levy et. al. Improving distributional similarity with lessons learned from word embeddings, 2015.

Paragraph2vec, он же doc2vec



DM (Distributed Memory):

$$p(w_i | w_{i-h}, \dots w_{i+h}, d)$$

DBOW (Distributed Bag Of Words):

$$p(w_{i-h}, \dots w_{i+h} | d)$$

Оценивание моделей: близости документов

Как оценить что *близкие документы* имеют *близкие вектора*?

- ArXiv: статья A, похожая статья B, не похожая статья C
- Точность угадывания, какая из статей B и C не похожа

http://arxiv.org/pdf/1206.5743	http://arxiv.org/pdf/cond-mat/0403258	http://arxiv.org/pdf/1408.0189
http://arxiv.org/pdf/1209.0268	http://arxiv.org/pdf/1307.7598	http://arxiv.org/pdf/math/0504051
http://arxiv.org/pdf/hep-ph/9908436	http://arxiv.org/pdf/nucl-th/9707019	http://arxiv.org/pdf/1112.3014
http://arxiv.org/pdf/1111.2905	http://arxiv.org/pdf/1303.2538	http://arxiv.org/pdf/1109.1922
http://arxiv.org/pdf/nucl-ex/0112013	http://arxiv.org/pdf/physics/9704013	http://arxiv.org/pdf/1408.4595
http://arxiv.org/pdf/0709.3419	http://arxiv.org/pdf/quant-ph/0611134	http://arxiv.org/pdf/0902.0616
http://arxiv.org/pdf/hep-th/9609148	http://arxiv.org/pdf/solv-int/9710009	http://arxiv.org/pdf/astro-ph/0508060

Andrew Dai, Cristopher Olah, Quoc Le. Document Embedding with Paragraph Vectors, CoRR, 2015.

Оценивание моделей: близости документов

Integral formula of Minkowski type and new characterization of the Wulff shape

Yijun He * Haizhong Li †

Abstract

Given a positive function F on S^n which satisfies a convexity condition, we introduce the r-th anisotropic mean curvature M_r for hypersurfaces in $\mathbb{R}^{n+1}$ which is a generalization of the usual r-th mean curvature H_r. We get integral formulas of Minkowski type for compact hypersurfaces in R^{n+1}. We give some new characterizations of the Wulff shape by use of our integral formulas of Minkowski type, in case $F = 1$ which reduces to some well-known results.

2000 Mathematics Subject Classification: Primary 53C42, 53A30; Secondary 53B25.

Key words and phrases: Wulff shape, F-Weingarten operator, anisotropic principal curvature, r-th anisotropic mean curvature, integral formula of Minkowski type.

xiv.org/pdf/1408.0189

xiv.org/pdf/math/0504051

xiv.org/pdf/1112.3014

xiv.org/pdf/1109.1922

xiv.org/pdf/1408.4595

xiv.org/pdf/0902.0616

xiv.org/pdf/astro-ph/0508060

Оценивание моделей: близости документов

Integral formula of Minkowski type and new characterization of the Wulff shape

COMPLEX CURVES IN ALMOST-COMPLEX MANIFOLDS AND MEROMORPHIC HULLS

Sergei IVASHKOVICH – Vsevolod SHEVCHISHIN

Preface

Chapter I. Local Properties of Complex Curves.

Lecture 1. Complex Curves in Almost-Complex Manifolds. . . . pp. 1–12
1.1. Almost-Complex Manifolds, Hermitian Metrics, Associated (1,1)-Forms. 1.2. Existence of Calibrating and Tame Structures. 1.3. Almost-Complex Submanifold, Complex Curves, Energy and Area. 1.4. Symplectic Surfaces. 1.5. Adjunction Formula for Immersed Symplectic Surfaces.

Оценивание моделей: близости документов

Accepted for publication in *Solar Physics*, waiting for the authoritative version and a DOI which will be available at http://www.springerlink.com/content/0038-0938

Time-dependent Stochastic Modeling of Solar Active Region Energy

M. Kanazir and M. S. Wheatland[1]

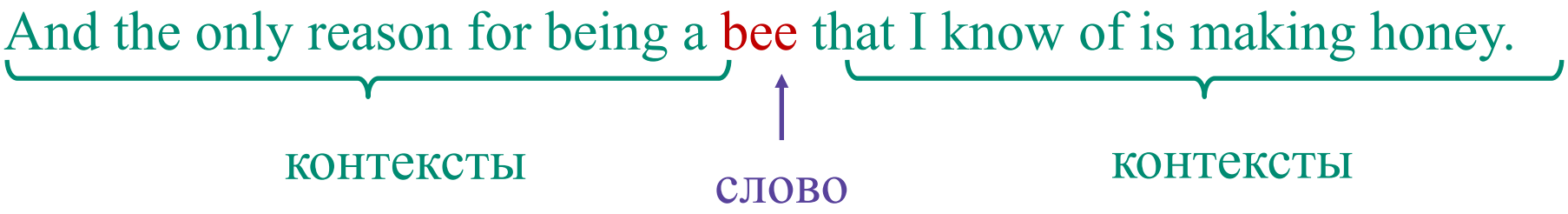
Received: 7 July 2010 / Accepted: 31 July 2010 / Published online: ••••••••••

Abstract A time-dependent model for the energy of a flaring solar active region

Lecture 1. Complex Curves in Almost-Complex Manifolds. . . . pp. 1–12
1.1. Almost-Complex Manifolds, Hermitian Metrics, Associated (1,1)-Forms. 1.2. Existence of Calibrating and Tame Structures. 1.3. Almost-Complex Submanifold, Complex Curves, Energy and Area. 1.4. Symplectic Surfaces. 1.5. Adjunction Formula for Immersed Symplectic Surfaces.

Задача аналогий без магии:
king + man – woman != queen

Магическое свойство модели word2vec



Вектора слов учатся по предсказаниям контекстов.

Получаются вектора, которые решают задачу аналогий:

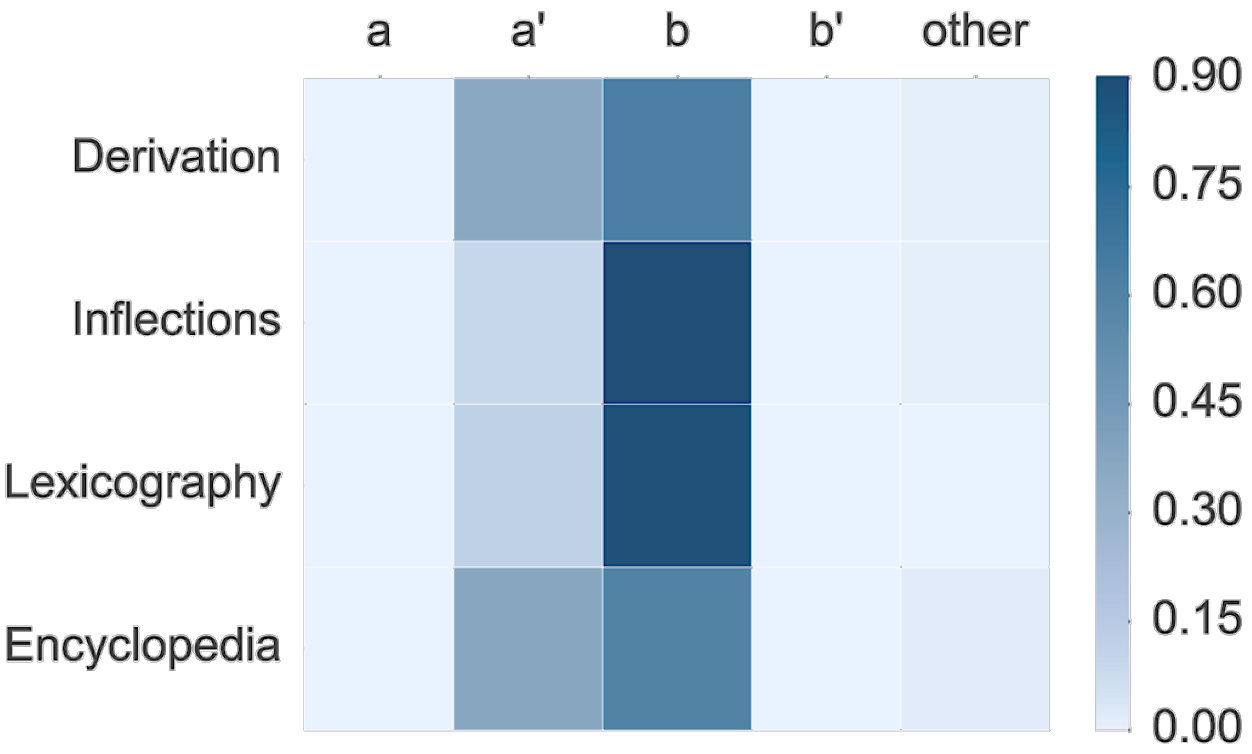
- king – man + woman = queen
- Moscow – Russia + France = Paris

Демо: rare-technologies.com/word2vec-tutorial/

Присмотримся внимательнее

$$\cos(b - a + a', x) \rightarrow \max_{x \notin \{a, a', b\}}$$

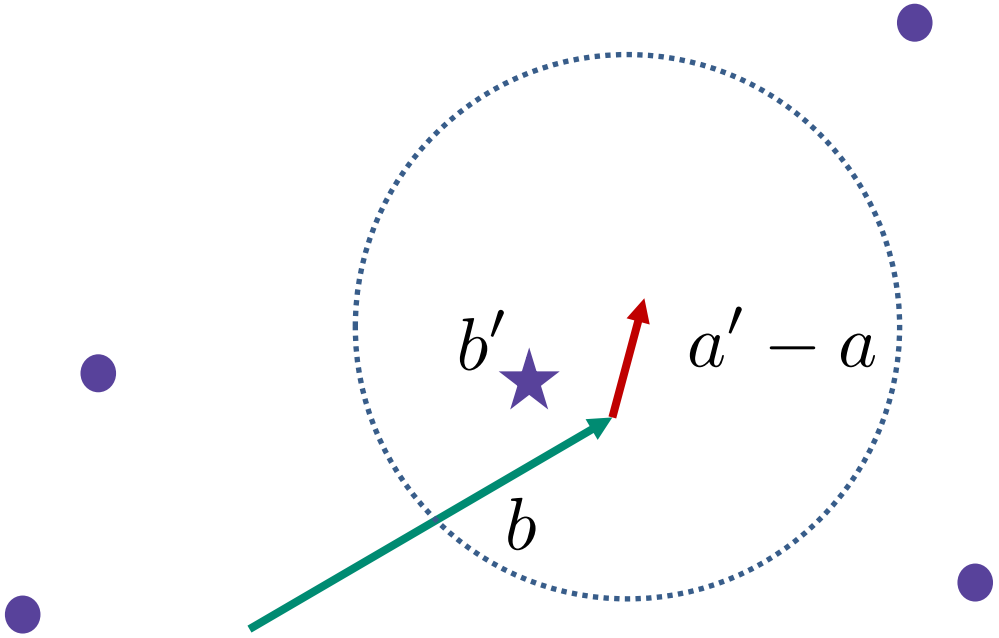
king – man + woman = king:



Rogers et. al. The (Too Many) Problems of Analogical Reasoning with Word Vectors, 2017.

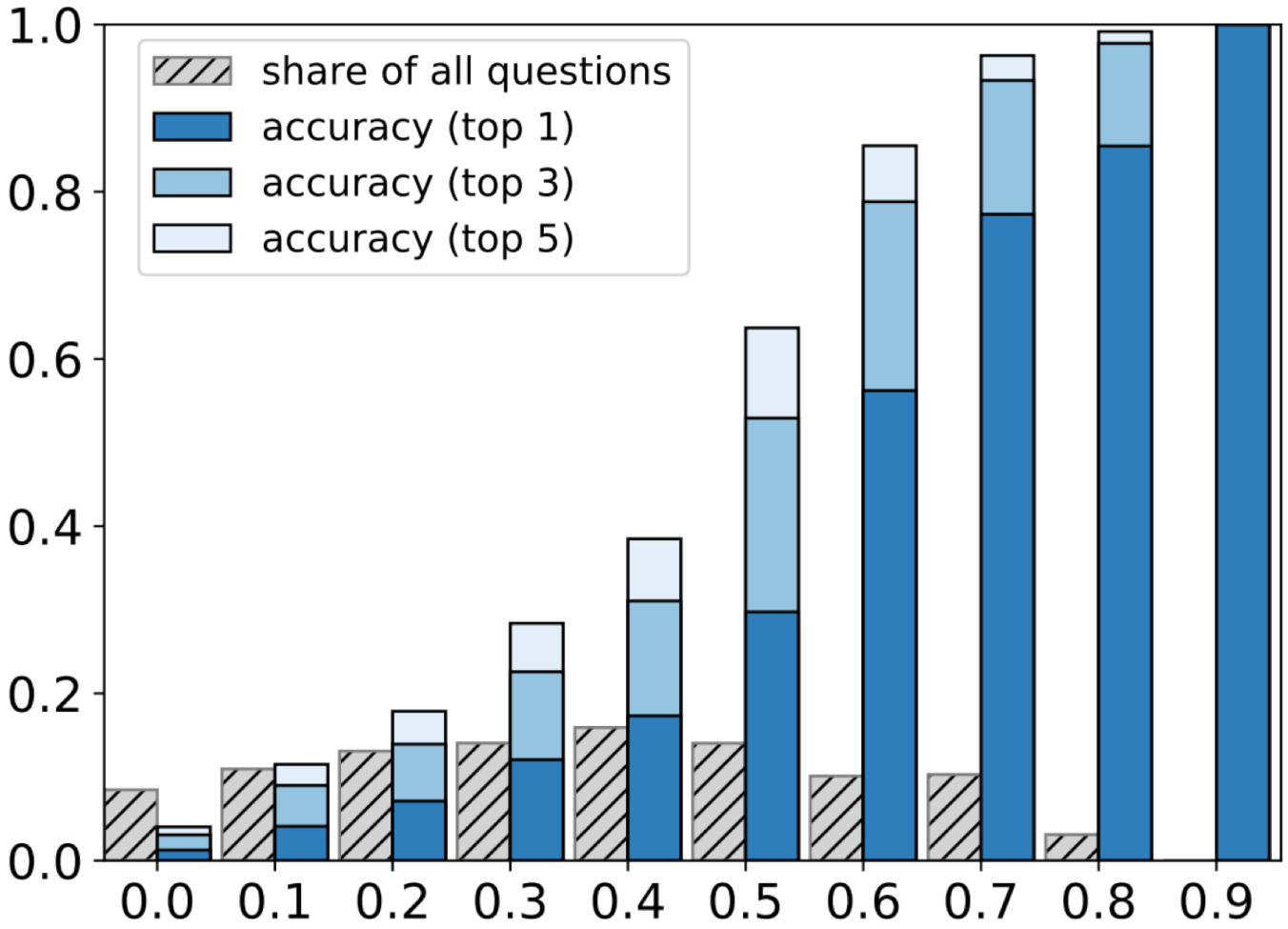
Достаточно взять ближайшего соседа для b?

Для категории множественного числа в датасете Google алгоритм, выдающий ближайшего соседа b, дает точность 70%. Как это возможно?



Linzen. Issues in evaluating semantic spaces using word analogies, 2016.

Точность высокая, когда b и b' близки



Корзинки основаны на близости между b и b'.

BATS (Bigger analogies test set)

Инфлекционная морфология:

- *student:students, strong:stronger, follow:following, …*

Деривационная морфология:

- *bake:baker, edit:editable, home:homeless, …*

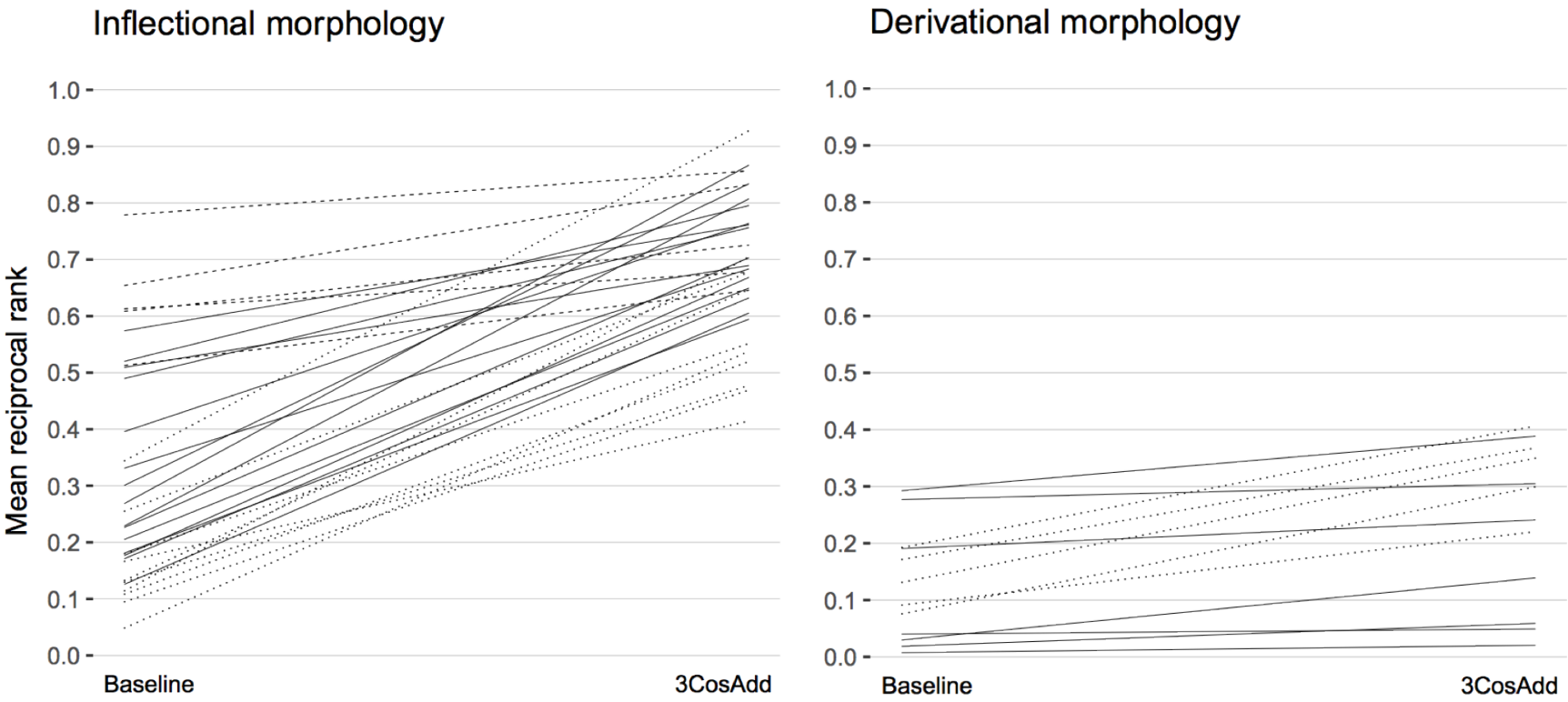
Лексикографическая семантика:

- Гипонимы, меронимы: *peach:fruit, sea:water, player:team, …*
- Антонимы, синонимы: *up:down, clean:dirty, cry:scream, …*

Энциклопедическая семантика:

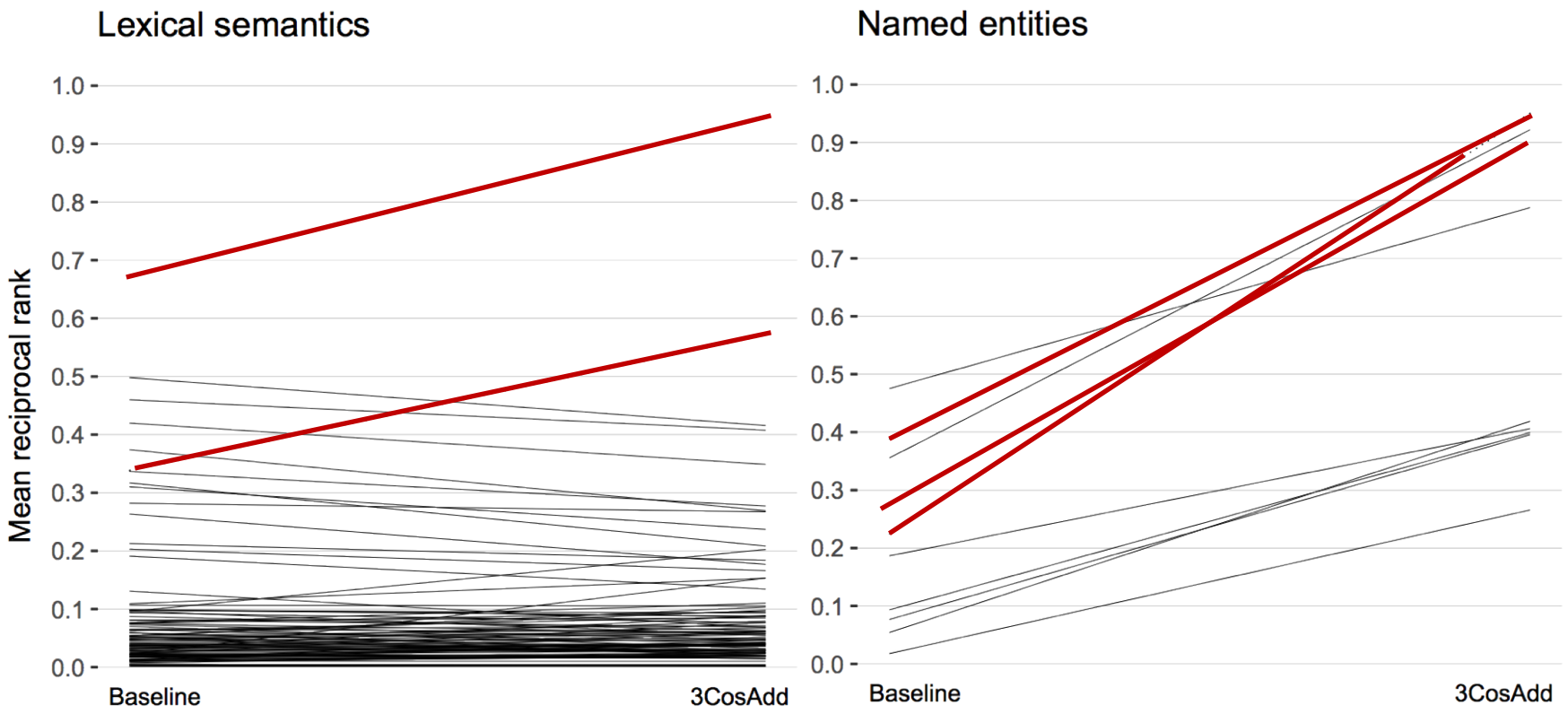
- Животные: *cat:kitten, dog:bark, …*
- География: *Athens:Greece, Peru:Spanish, …*
- Люди: *Lincoln:president, Lincoln:American, …*
- Другое: *blood:red, actor:actress, …*

Качество по категориям



Finley et. al. What Analogies Reveal about Word Vectors and their Compositionality, 2017.

Аналогии на пол и на страны – лучшие



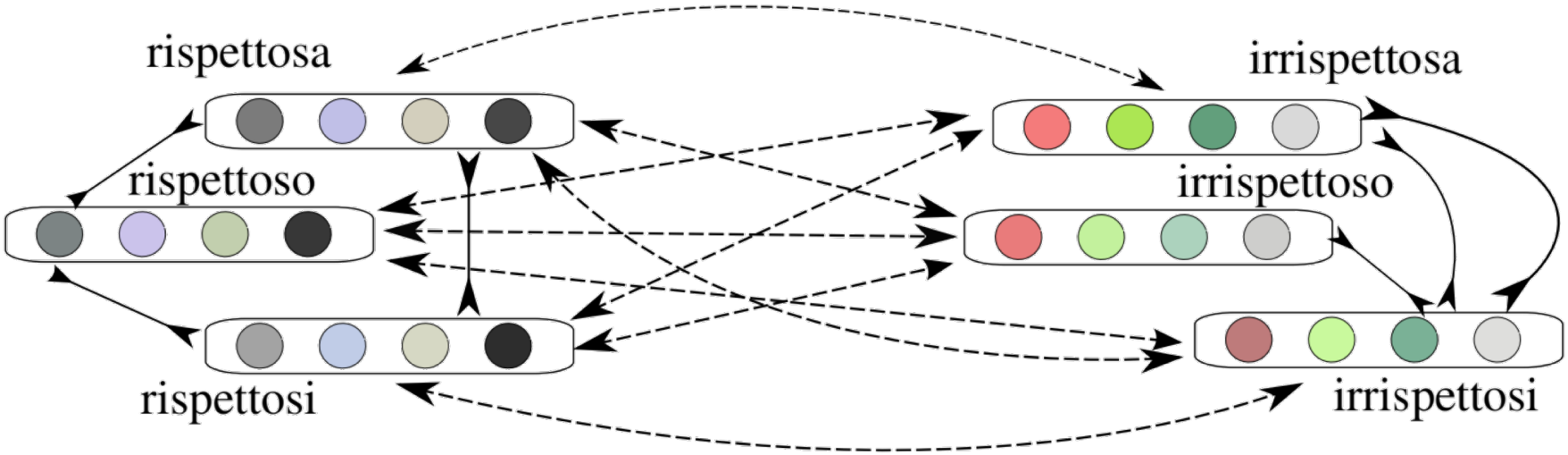
Finley et. al. What Analogies Reveal about Word Vectors and their Compositionality, 2017.

Почему только слова?
Вектора для символов и предложений

Морфология

Проблемы:

- Языки с богатой морфологией
- Редкие слова, OOV слова



Использование морфологии для улучшения векторов слов.

I. Vulic et. al. Morph-fitting: Fine-Tuning Word Vector Spaces with Simple Language-Specific Rules, 2017.

FastText

Представим слово мешком буквенных n-грамм, для n = 3:

G_{where} : ***_wh, whe, her, ere, re_, _where_***

Моделируем вектор слова как сумму векторов n-грамм:

SGNS:

$$sim(u, v) = \langle \phi_u, \theta_v \rangle$$

FastText:

$$sim(u, v) = \sum_{g \in G_v} \langle \phi_u, \theta_g \rangle$$

Код и вектора: https://fasttext.cc/

P. Bojanowsky et al. Enriching Word Vectors with Subword Information, 2016.

Sent2vec

Первые идеи:

- Усреднить *предобученные* вектора слов *(word2vec, GloVe)*.
- Возможно, использовать TF-IDF веса.

Sent2vec:

- *Обучить* вектора предложений как усреднение векторов:

$$sim(u, s) = \frac{1}{|G_s|} \sum_{g \in G_s} \langle \phi_u, \theta_g \rangle$$

где G_s это множество n-грамм для предложения *s*.

Код и вектора: https://github.com/epfml/sent2vec

Pagliardini et. al. Unsupervised Learning of Sentence Embeddings using Compositional n-Gram Features, 2017.

StarSpace

Общий инструмент:

сущности (предложения) и *признаки* (слова)

Множество приложений:

- Классификация текстов, например, сентимент
- Ранжирование, например, выдачи поиска
- Рекомендательные системы
- Представления графов
- ….

Код и тьюториал: github.com/facebookresearch/Starspace

Wu et. al. StarSpace: Embed All The Things! 2017

StarSpace

Mode 3 (вектора для предложений):

Задача: обучить вектора предложений, используя коллекцию пар близких предложений.

Формат данных: близкие предложения в одной строке:

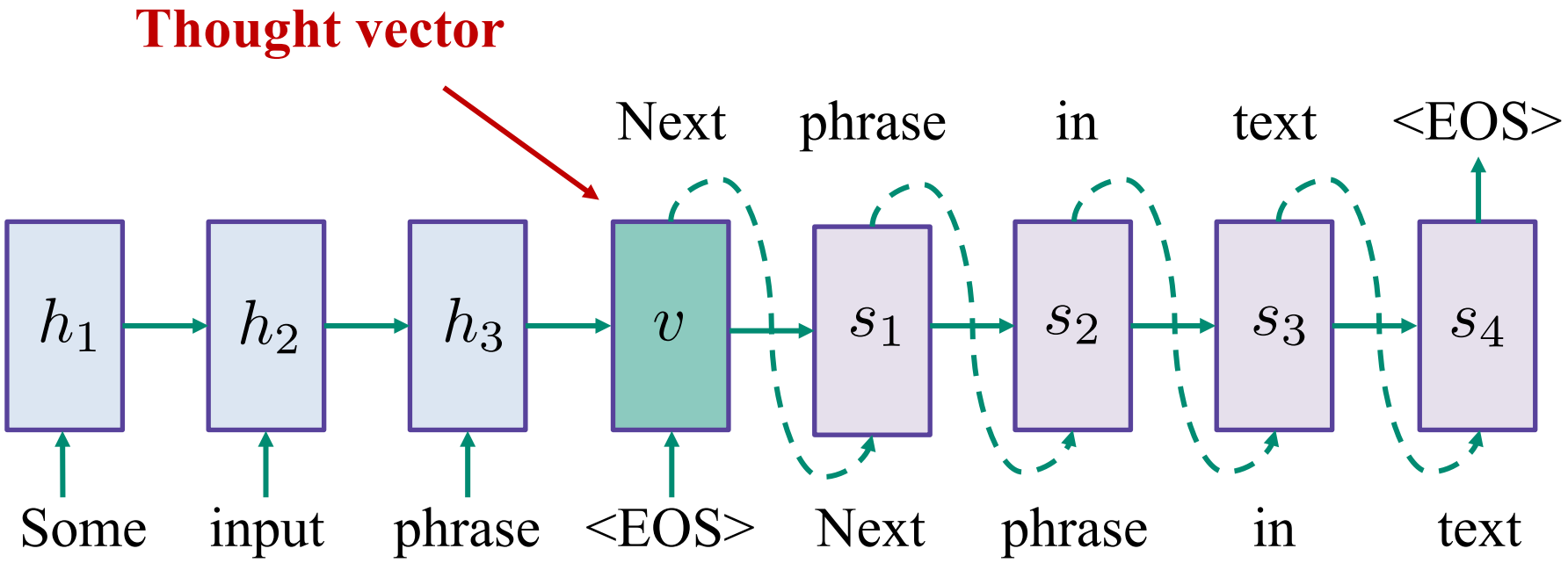
```
sent1_word1 sent1_word2 … <tab> sent2_word1 sent2_word2 …
```

Обучение:

- Каждое предложение представляется как мешок признаков (слов и n-грамм).
- Предсказывается попарная близость предложений.
- Негативные примеры пар сэмплируются случайно.

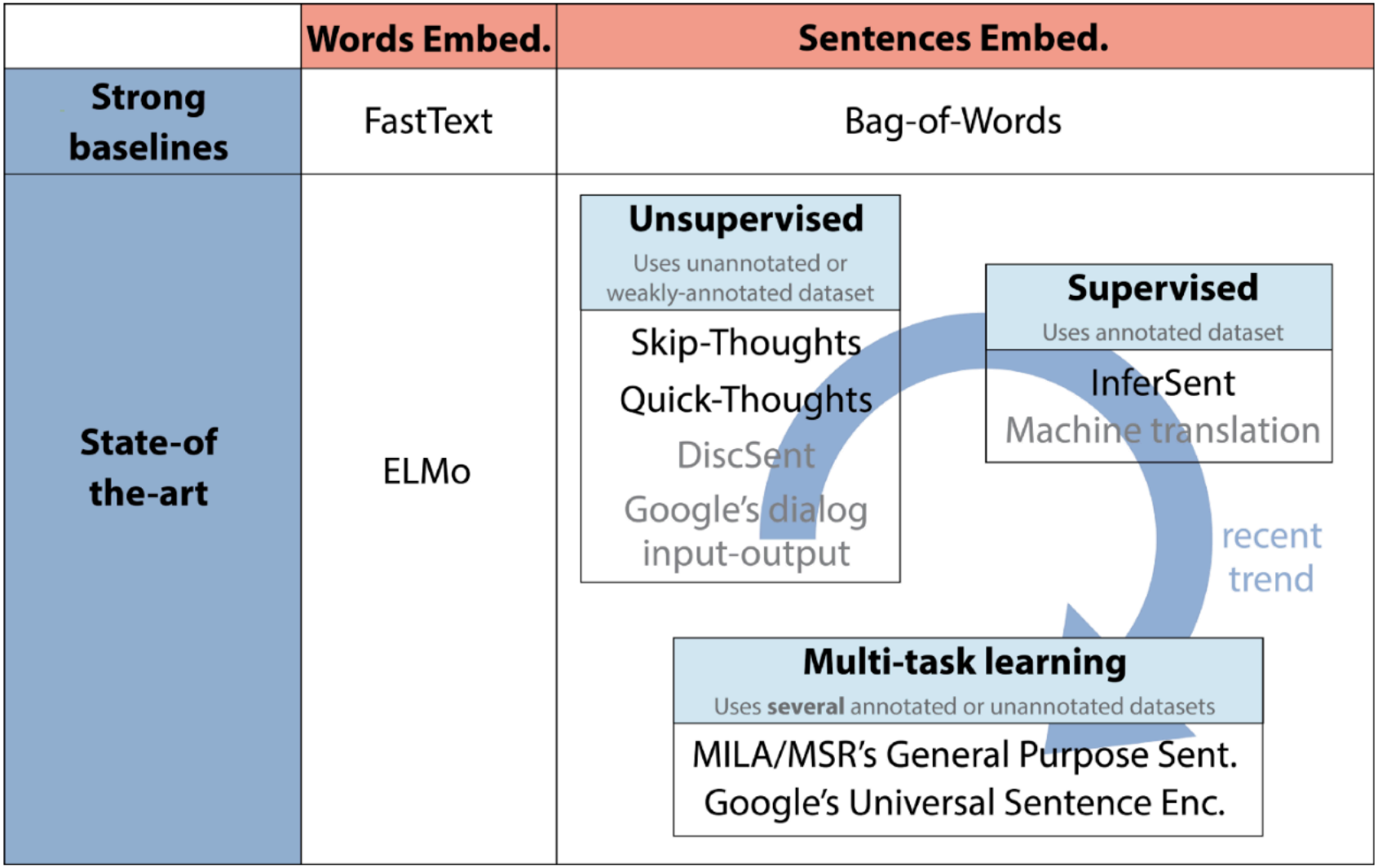
Skip-thought vectors

- Предсказываем предыдущее и следующее предложение
- Архитектура: RNN encoder-decoder



Kiros et. al. Skip-Thought Vectors, 2015, https://github.com/ryankiros/skip-thoughts.

SOTA ушла за два года вперед…



https://medium.com/huggingface/universal-word-sentence-embeddings-ce48ddc8fc3a