Word2vec

Евгений Соколов, Виктор Кантор

Похожие слова

- «Идти» и «шагать» — синонимы
- Для компьютера это разные строки
- Как понять, что они похожи?

Похожие слова

- «Идти» и «шагать» — синонимы
- Для компьютера это разные строки
- Как понять, что они похожи?

- На основе данных!
- Слова со схожим смыслом часто идут в паре с одними и теми же словами
- У них похожие контексты

Дистрибутивная семантика

Term-context matrix

	C1	C2	C3	C4	C5	C6	C7
dog	5	0	11	2	2	9	1
cat	4	1	7	1	1	7	2
bread	0	12	0	0	9	1	9
pasta	0	8	1	2	14	0	10
meat	0	7	1	1	11	1	8
mouse	4	0	8	0	1	8	1

Term-context matrix

	dog	cat	computer	animal	mouse
dog	0	4	0	2	1
cat	4	0	0	3	5
computer	0	0	0	0	3
animal	2	3	0	0	2
mouse	1	5	3	2	0

Векторные представления слов

Хотим каждое слово представить как вещественный вектор:

$$w \to \vec{w} \in \mathbb{R}^d$$

Какие требования?

- Размерность d должна быть не очень велика
- Похожие слова должны иметь близкие векторы
- Арифметические операции над векторами должны иметь смысл

word2vec

- Будем обучать представления слов так, чтобы они хорошо предсказывали свой контекст
- Выборка состоит из текстов, каждый представляет собой последовательность слов $w_1, \ldots, w_i, \ldots, w_n$

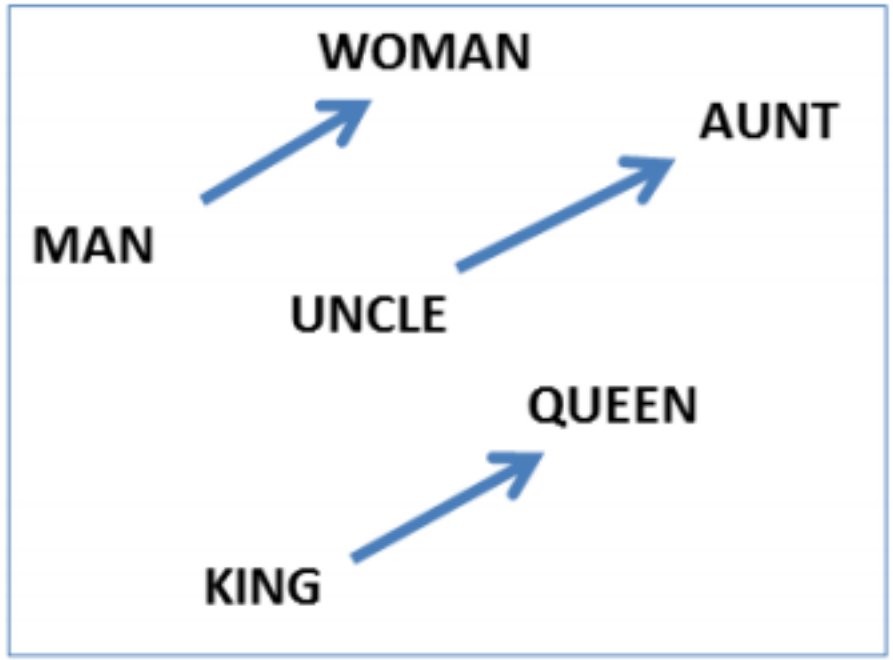
$$\sum_{i=1}^{n} \sum_{\substack{j=-k \\ j \neq 0}}^{k} \log p(w_{i+j} \mid w_i) \to \max,$$

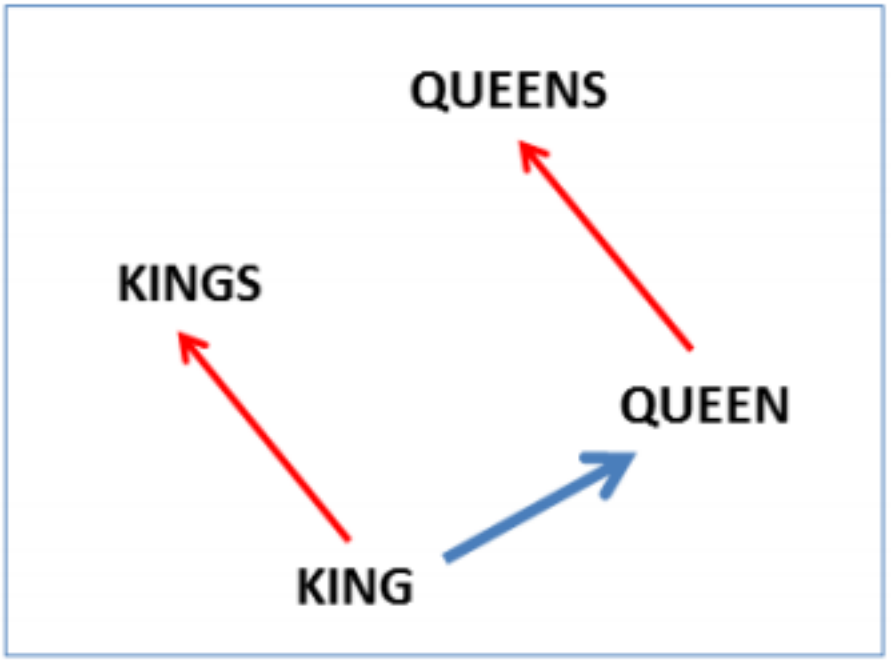
где вероятность вычисляется через soft-max:

$$p(w_i \mid w_j) = \frac{\exp(\langle \vec{w}_i, \vec{w}_j \rangle)}{\sum_w \exp(\langle \vec{w}, \vec{w}_j \rangle)}$$

(сумма в знаменателе — по всем словам из словаря)

Самый популярный пример на word2vec



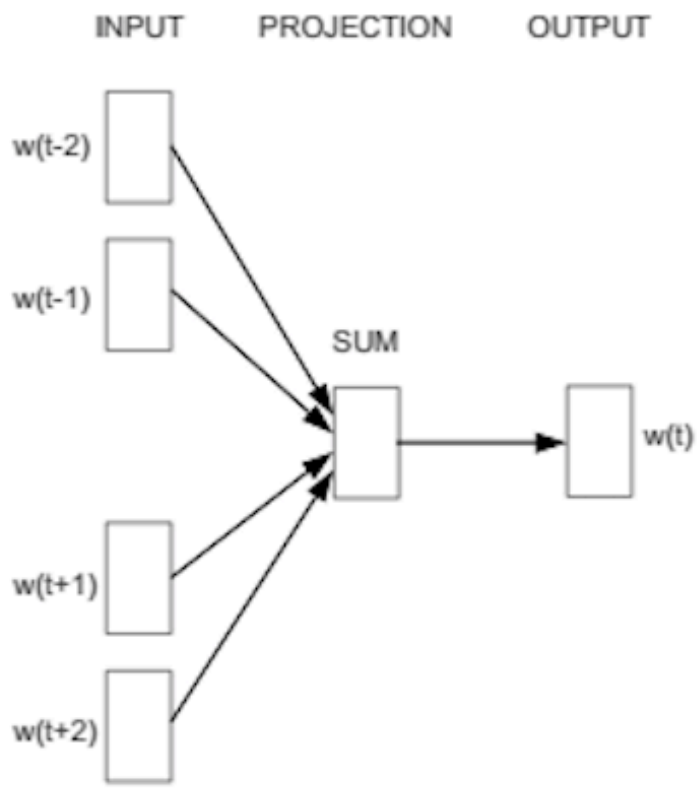


(Mikolov et al., NAACL HLT, 2013)

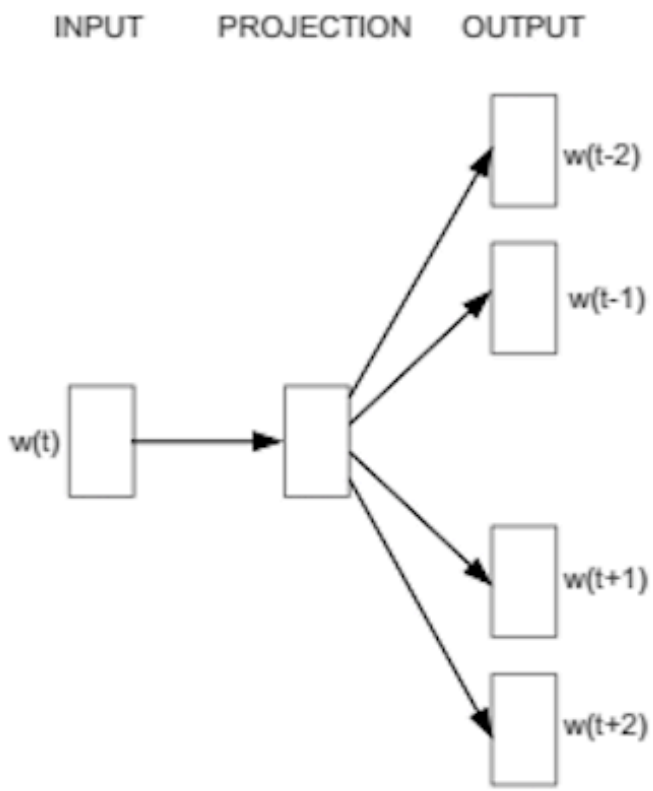
Свойства представлений

- Косинусное расстояние хорошо отражает схожесть слов по тематике (в зависимости от корпуса)
- $\vec{\text{king}} - \vec{\text{man}} + \vec{\text{woman}} \approx \vec{\text{queen}}$
- $\vec{\text{Moscow}} - \vec{\text{Russia}} + \vec{\text{England}} \approx \vec{\text{London}}$
- Перевод: $\vec{\text{one}} - \vec{\text{uno}} + \vec{\text{four}} \approx \vec{\text{quatro}}$
- Среднее представление по всем словам в тексте — хорошее признаковое описание

Традиционная картинка про word2vec

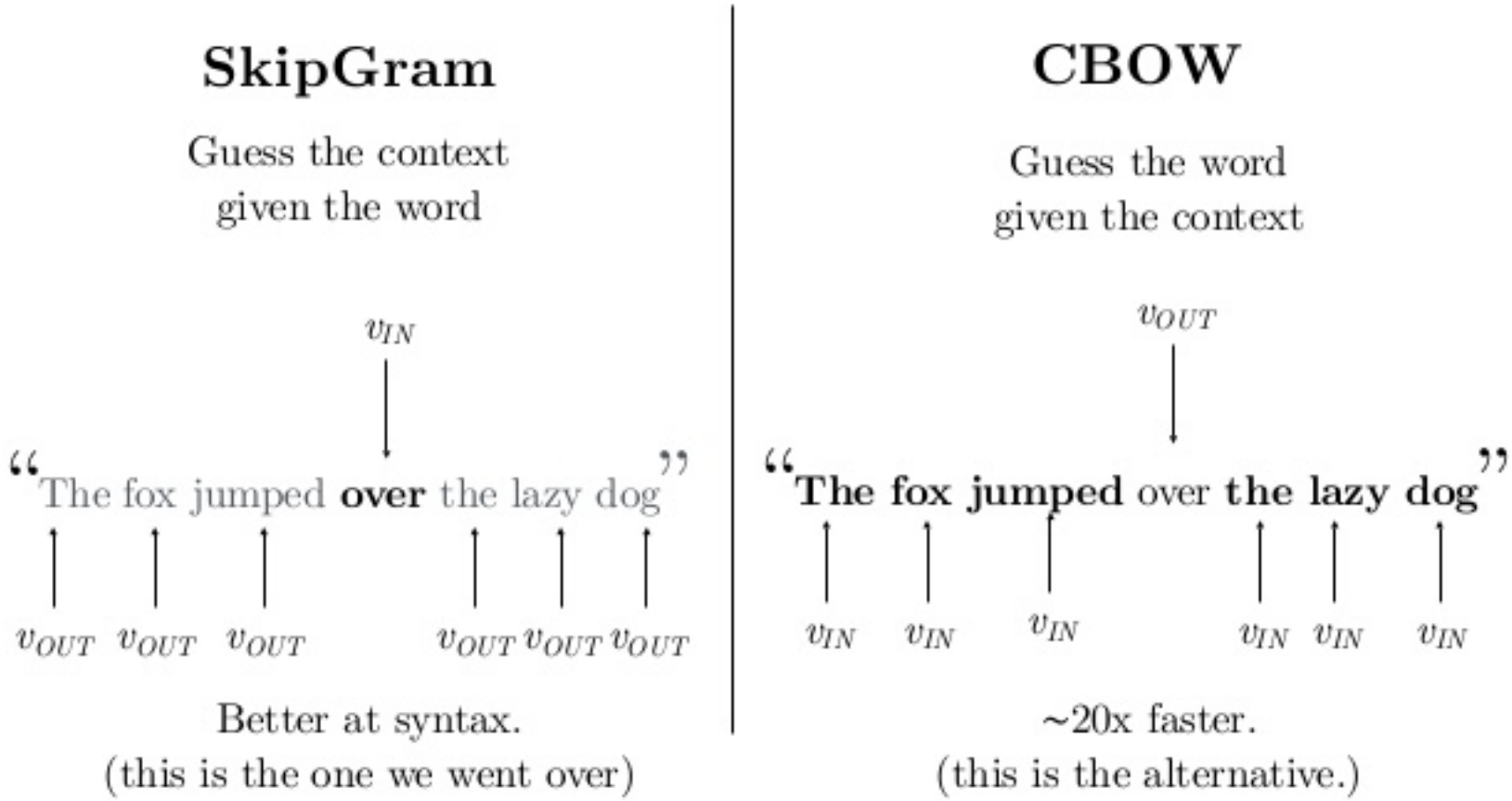


CBOW



Skip-gram

CBOW и Skip-gram



Еще одна постановка в word2vec

$$P(D = 1|w, c) = \sigma(\vec{w} \cdot \vec{c}) = \frac{1}{1 + e^{-\vec{w} \cdot \vec{c}}}$$

Еще одна постановка в word2vec

$$P(D = 1|w, c) = \sigma(\vec{w} \cdot \vec{c}) = \frac{1}{1 + e^{-\vec{w} \cdot \vec{c}}}$$

$$\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} \left[\log \sigma(-\vec{w} \cdot \vec{c}_N) \right]$$

Еще одна постановка в word2vec

$$P(D = 1|w, c) = \sigma(\vec{w} \cdot \vec{c}) = \frac{1}{1 + e^{-\vec{w} \cdot \vec{c}}}$$

$$\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} \left[\log \sigma(-\vec{w} \cdot \vec{c}_N) \right]$$

Negative sampling

Еще одна постановка в word2vec

$$P(D = 1|w, c) = \sigma(\vec{w} \cdot \vec{c}) = \frac{1}{1 + e^{-\vec{w} \cdot \vec{c}}}$$

$$\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} \left[\log \sigma(-\vec{w} \cdot \vec{c}_N) \right]$$

$$P_D(c) = \frac{\#(c)}{|D|}$$

Еще одна постановка в word2vec

$$P(D = 1|w, c) = \sigma(\vec{w} \cdot \vec{c}) = \frac{1}{1 + e^{-\vec{w} \cdot \vec{c}}}$$

$$\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} \left[\log \sigma(-\vec{w} \cdot \vec{c}_N) \right]$$

$$P_D(c) = \frac{\#(c)}{|D|}$$

$$\ell = \sum_{w \in V_W} \sum_{c \in V_C} \#(w, c) \left(\log \sigma(\vec{w} \cdot \vec{c}) + k \cdot \mathbb{E}_{c_N \sim P_D} \left[\log \sigma(-\vec{w} \cdot \vec{c}_N) \right] \right)$$

word2vec и обучение с учителем

- Проблема мешка слов — слишком большое количество признаков
- Средний word2vec-вектор позволяет получить компактное признаковое описание
- При размерности вектора 100-500 можно обучать композиции деревьев

Связь word2vec и матричных разложений

Neural Word Embedding as Implicit Matrix Factorization - Omer Levy

Резюме

- В дистрибутивной семантике предполагается, что смысл слова описывается контекстами, в которых оно встречается
- Как правило в качестве множества контекстов рассматривают множество слов
- Word2vec позволяет описать каждое слово вектором
- Есть разные постановки оптимизационной задачи для word2vec
- Похожие слова имеют близкие векторы
- Неплохие для текста — средний вектор по всем словам
- Word2vec (в одной из постановок) фактически решает задачу разложения матрицы некоторого специального вида