

Представление отношений между словами естественного языка в виде разложения на симметрический и кососимметрический линейный оператор

Кощенко Екатерина

Научный руководитель: к.ф.-м.н. Булычев Д.Ю.

Консультант: к.ф.-м.н. Кураленок И.Е.

НИУ ВШЭ

June 6, 2019

- Векторизация — представление слова в виде вещественного вектора (*word embedding*).
- Важные свойства векторов:
 - расстояния и углы между представлениями слов
$$e(\text{Красный}) - e(\text{Желтый}) < e(\text{Красный}) - e(\text{Умный})$$
 - словесные аналогии [1]
$$e(\text{Швеция}) - e(\text{Стокгольм}) \approx e(\text{Норвегия}) - e(\text{Осло})$$

[1] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. “Linguistic Regularities in Continuous Space Word Representations.” In: 2013.

Обработка естественного языка. Примеры задач:

- Индексирование документов [2]
- Анализ тональности текста [3]
- Вопросно-ответная система [4]

Векторизация также используется в языковом моделировании — другом подходе к решению задач обработки естественного языка.

[2] F. Sebastiani. “Machine learning in automated text categorization”. In: 2002.

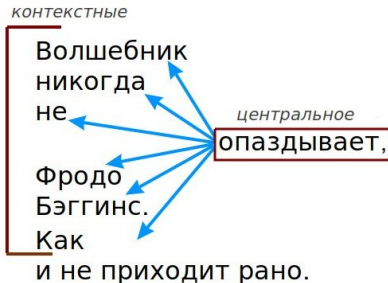
[3] William L. Hamilton et al. “Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora”. In: 2016.

[4] Yikang Shen et al. “Word Embedding Based Correlation Model for Question/Answer Matching”. In: 2017.

Основные подходы

Общая идея векторизации:

- Корпус обучения сканируется скользящим окном
- Слова в окне: центральное и контекстные



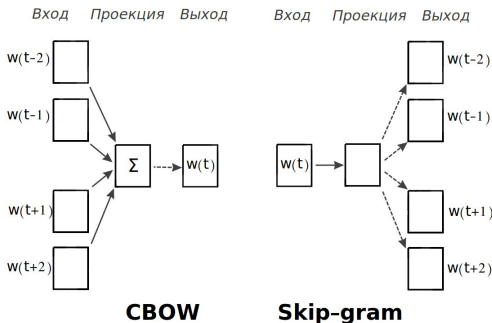
- Считается вероятность появления слов в контексте

Есть два популярных подхода: Word2Vec и GloVe.

Основные подходы: Word2Vec

Mikolov et al. (2013) [5]

- Две модели: CBOW и Skip-gram
- Слово представляется двумя векторами разных ролей
- Сохраняется свойство словесных аналогий



[5] Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: 2013.

Manning et al. (2014) [6]

- Отношения между словами – матрица попарных встречаемостей
- Слово представляется двумя векторами
- Сохраняется свойство словесных аналогий

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. “GloVe: Global Vectors for Word Representation”. In: 2014.

- Учет свойства асимметрии отношений слов.
- Вычисление двух векторов разных ролей на слово: центральный и контекстный.
- Моделирование отношения скалярным произведением.
- Использование целевого и контекстного векторов равной длины.

Цель: построить модель, контролирующую влияние асимметрической информации на векторизацию.

Задачи:

- Проанализировать модель GloVe
- Построить на основе GloVe свою модель, контролирующую влияние асимметрии
- Сравнить модель с существующими на задаче словесных аналогий
- Изучить влияние асимметрии слов на результирующую векторизацию

Можно провести параллель с формулой взаимной информации.

$$J^* = \sum_{y,x} p(x,y|D) \cdot \left(\log \frac{p(x,y|F)}{p(x|F)p(y|F)} - \log \frac{p(x,y|D)}{p(x|D)p(y|D)} \right)^2,$$

где F — модель, D — данные корпуса.

$$\frac{p(x,y|F)}{p(x|F)p(y|F)} \Rightarrow e^{u_x^T v_y}$$

$$\log \frac{p(x,y|D)}{p(x|D)p(y|D)} \Rightarrow \log X_{xy} - b_x - b_y,$$

где u_x и v_y — контекстный и центральный вектора, b_x и b_y - переменные смещения.

Любой линейный оператор можно разложить на сумму симметричной и кососимметричной матриц. [7]

$$u_i^T v_j = w_i U^T V w_j$$

$$L = U^T V = S + K$$

Раскладываем симметричную на квадрат матрицы меньшего ранга. Аналогично с кососимметричной.

$$l_{ij} = s_{ij} + k_{ij} = a_i^T a_j + \xi_{ij} \cdot c_i^T c_j,$$

где $\xi_{ij} = -1$, если $i > j$, иначе $\xi_{ij} = 1$

[7] F.R. Gantmacher. "The theory of matrices". In: 1960.

Целевая функция:

$$Q = \sum_{i,j=1}^{|V|} p(w_i, w_j|D) \cdot (a_i^T a_j + \xi_{ij} \cdot c_i^T c_j + \log p(w_i|D) + \\ + \log p(w_j|D) - \log p(w_i, w_j|D))^2,$$

где

- D — корпус обучения,
- V — словарь, построенный по корпусу,
- a_i — симметричный вектор слова w_i ,
- c_i — кососимметрический вектор слова w_i ,
- $\xi_{ij} = -1$, если $i > j$, иначе $\xi_{ij} = 1$.

SSDE — Symmetric Skew-symmetric Decomposition Embedding

SSDE — модель векторизации слов, позволяющая отдельно контролировать количество параметров, отвечающих за влияние симметричной и асимметричной информации на результат.

За основу SSDE была взята модель GloVe, т.к.

- меньшее время обучения
- лучшие результаты на задаче словесных аналогий.

Метрики словесных аналогий, собранные GloVe.

Примеры:

- (Сем) "Father" - "Man" + "Woman" = "Mother"
- (Сем) "Dollar" - "USA" + "Russia" = "Ruble"
- (Синт) "Calmly" - "Calm" + "Happy" = "Happily"
- (Синт) "Worst" - "Bad" + "Good" = "Best"

Корпус обучения — выделенный из датасета "2014 Wikipedia dump" корпус 100Мб.

Одинаковая длина вектора представления

Model	Average sem score	Average synt score	Average total score	Sec per iter
GloVe-80	25.3%	40.2%	31.3%	27
SSDE-80-5	24.3%	39.8%	30.6%	17
SSDE-80-20	25.5%	41.4%	31.9%	20
SSDE-80-50	24.2%	39.3%	29.6%	30
SSDE-80-80	25.1%	39.5%	30.9%	32
GloVe-120	27.6%	44.0%	34.2%	34
SSDE-120-10	25.3%	48.1%	34.5%	23
SSDE-120-20	26.3%	47.8%	35.1%	24

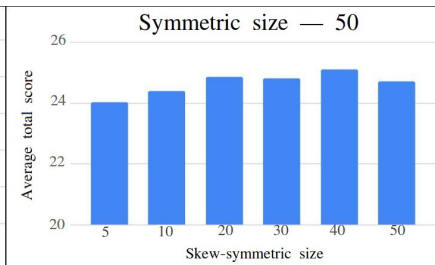
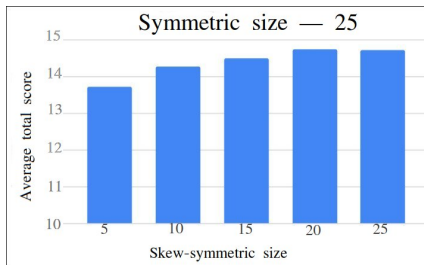
Результаты сравнения

Сравнимое количество обучаемых параметров.

Model	Average sem score	Average synt score	Average total score	Sec per iter
GloVe-25	12.9%	12.3%	12.5%	13
SSDE-25-20	16.5%	13.6%	14.6%	13
SSDE-40-5	19.1%	22.5%	20.4%	11
SSDE-40-10	21.6%	23.5%	22.8%	12
SSDE-50-5	22.5%	28.7%	24.0%	12
GloVe-50	20.3%	27.4%	23.9%	18
SSDE-50-40	21.5%	32.1%	25.1%	19
SSDE-80-5	24.3%	39.8%	30.6%	17
SSDE-80-20	25.5%	41.4%	31.9%	20

Анализ влияния асимметрии

Увеличение кососимметрической части до некоторого момента улучшает результаты. Но кососимметрическая часть длины, сравнимой с симметрической, работает не лучше, чем половина симметрической.



- По итогам анализа GloVe была построена модель векторизации SSDE, позволяющая контролировать влияние асимметрической информации.
- Сравнение SSDE и GloVe показало:
 - Длина векторизации — 80.
Результаты GloVe на 10 сек/итерация быстрее.
 - Количество параметров — 50.
Результаты на 12% лучше GloVe за то же время.
- Результаты были представлены на конференции SEIM.

ДОПОЛНЕНИЕ

- Применить для разных задач окна разных типов
- Использовать информацию, полученную в асимметрических векторах
- Научиться различать многозначность

Bojanowski et al. (2017)

- Слово представляется двумя векторами
- Сохраняется информация о строении слова через n -грамм

$$u_i^T v_j \rightarrow \sum_{g \in G_{w_i}} u_g^T v_j,$$

где G_{w_i} – слово w_i и все его n -граммы.

- Используется для классификации текстов
- Одна итерация за $\mathcal{O}(|T| \cdot \text{const})$

Главная идея: отношение слов w_i и w_j выражается через их отношения с другими словами (w_k):

$$F((u_i - u_j)^T v_k) = \frac{P(w_i, w_k)}{P(w_j, w_k)}, \quad (1)$$

где

- F – тренируемая модель
- u – центральный вектор
- v – контекстный вектор

С точки зрения матрицы встречаемостей, роль слова не имеет значения. Поэтому функция модели должна быть гомоморфизмом:

$$F((u_i - u_j)^T v_k) = \frac{F(u_i^T v_k)}{F(u_j^T v_k)} \quad (2)$$

Тогда F – экспоненциальна, в сочетании с (2):

$$u_i^T v_k = \log P_{ik} = \log X_{ik} - \log X_i \quad (3)$$

Дальше в формулу вносятся смещения:

$$u_i^T v_k + b_i^u + b_k^v = \log X_{ik} \quad (4)$$

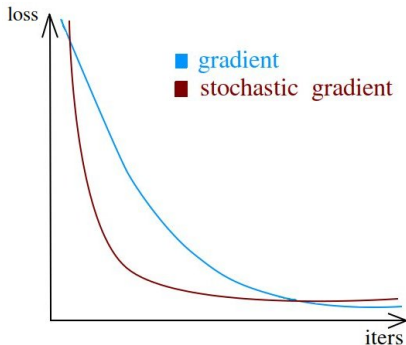
Целевая функция GloVe:

$$J = \sum_{i,j=1}^{|V|} f(X_{ij}) \cdot (u_i^T v_j + b_i^u + b_j^v - \log X_{ij})^2 \quad (5)$$
$$f(x) = \begin{cases} (x/x_{max})^\alpha, & \text{if } x < x_{max} \\ 1, & \text{otherwise} \end{cases}$$

где

- X — матрица встречаемостей,
- $|V|$ — размер словаря,
- u_i и b_i^u — вектор-кодер и смещение слова w_i ,
- v_j и b_j^v — вектор-декодер и смещение слова w_j ,
- $f(x)$ — обрезает частотные встречаемости.

GloVe использовал метод стохастической оптимизации AdaGrad [8] . На одной его итерации значения мало менялись, что позволило распараллелить процесс.



[8] John Duchi, Elad Hazan, and Yoram Singer. “Adaptive subgradient methods for online learning and stochastic optimization”. In: 2011.