Дистрибутивная семантика

Андрей Кутузов и Елизавета Кузьменко

Высшая Школа Экономики

Содержание

Дистрибутивная семантика: как моделировать значение?

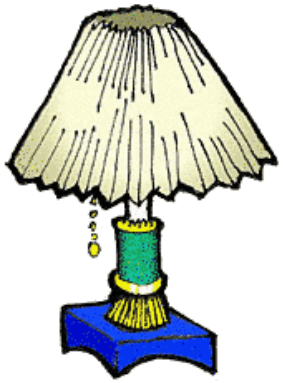
- «светильник»

Дистрибутивная семантика: как моделировать значение?

- «светильник»
- «лампа»

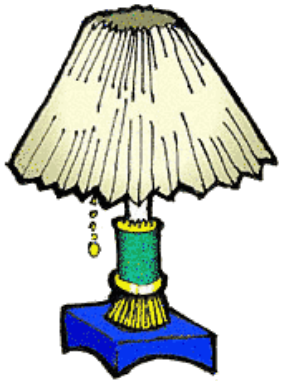
Дистрибутивная семантика: как моделировать значение?

- «светильник»
- «лампа»



Дистрибутивная семантика: как моделировать значение?

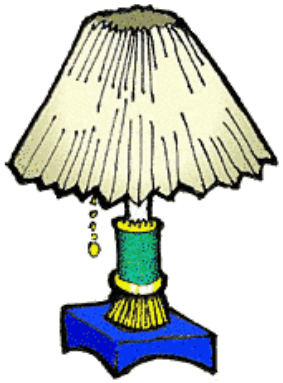
- «светильник»
- «лампа»



Есть много вопросов

Дистрибутивная семантика: как моделировать значение?

- «светильник»
- «лампа»

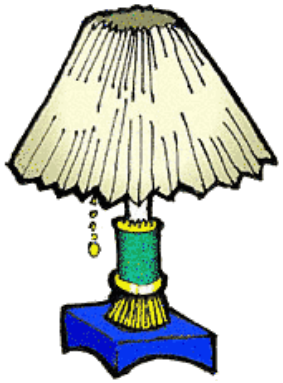


Есть много вопросов

- Внешне эти слова совсем не похожи друг на друга.
- Откуда мы знаем, что у слов *«светильник»* и *«лампа»* схожее значение?

Дистрибутивная семантика: как моделировать значение?

- «светильник»
- «лампа»

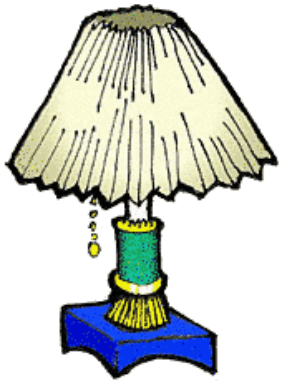


Есть много вопросов

- Внешне эти слова совсем не похожи друг на друга.
- Откуда мы знаем, что у слов *«светильник»* и *«лампа»* схожее значение?
- Как эта информация хранится в мозге?

Дистрибутивная семантика: как моделировать значение?

- «светильник»
- «лампа»

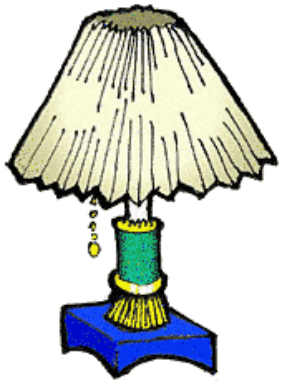


Есть много вопросов

- Внешне эти слова совсем не похожи друг на друга.
- Откуда мы знаем, что у слов *«светильник»* и *«лампа»* схожее значение?
- Как эта информация хранится в мозге?
- ...и главное

Дистрибутивная семантика: как моделировать значение?

- «светильник»
- «лампа»



Есть много вопросов

- Внешне эти слова совсем не похожи друг на друга.
- Откуда мы знаем, что у слов *«светильник»* и *«лампа»* схожее значение?
- Как эта информация хранится в мозге?
- ...и главное – как её моделировать для обработки компьютерами?

Дистрибутивная семантика: как моделировать значение?

Произвольность языкового знака

Дистрибутивная семантика: как моделировать значение?

Произвольность языкового знака

В отличие от многих других знаков, у слов форма и содержание не связаны напрямую друг с другом. Концепт *«лампа»* может выражаться любой последовательностью букв/звуков:

- лампа

Дистрибутивная семантика: как моделировать значение?

Произвольность языкового знака

В отличие от многих других знаков, у слов форма и содержание не связаны напрямую друг с другом. Концепт *«лампа»* может выражаться любой последовательностью букв/звуков:

- лампа
- lantern

Дистрибутивная семантика: как моделировать значение?

Произвольность языкового знака

В отличие от многих других знаков, у слов форма и содержание не связаны напрямую друг с другом. Концепт *«лампа»* может выражаться любой последовательностью букв/звуков:

- лампа
- lantern
- lucerna

Дистрибутивная семантика: как моделировать значение?

Произвольность языкового знака

В отличие от многих других знаков, у слов форма и содержание не связаны напрямую друг с другом. Концепт *«лампа»* может выражаться любой последовательностью букв/звуков:

- лампа
- lantern
- lucerna
- гэрэл

Дистрибутивная семантика: как моделировать значение?

Произвольность языкового знака

В отличие от многих других знаков, у слов форма и содержание не связаны напрямую друг с другом. Концепт *«лампа»* может выражаться любой последовательностью букв/звуков:

- лампа
- lantern
- lucerna
- гэрэл
- ...

Дистрибутивная семантика: как моделировать значение?

Уровни анализа языка

Дистрибутивная семантика: как моделировать значение?

Моделирование – это когда мы можем в сжатом виде представить некие важные свойства изучаемого явления. Например, в предложении *«Лампа стоит на столе»*, слово *«лампа»*:

1. состоит из 5 фонем *[l a m p ə]*;

Дистрибутивная семантика: как моделировать значение?

Моделирование – это когда мы можем в сжатом виде представить некие важные свойства изучаемого явления. Например, в предложении *«Лампа стоит на столе»*, слово *«лампа»*:

1. состоит из 5 фонем *[l a m p ə]*;
2. является существительным женского рода в единственном числе, именительном падеже;

Дистрибутивная семантика: как моделировать значение?

Моделирование – это когда мы можем в сжатом виде представить некие важные свойства изучаемого явления. Например, в предложении *«Лампа стоит на столе»*, слово *«лампа»*:

1. состоит из 5 фонем *[l a m p ə]*;
2. является существительным женского рода в единственном числе, именительном падеже;
3. выполняет в предложении функцию субъекта (подлежащего).

Дистрибутивная семантика: как моделировать значение?

Моделирование – это когда мы можем в сжатом виде представить некие важные свойства изучаемого явления. Например, в предложении *«Лампа стоит на столе»*, слово *«лампа»*:

1. состоит из 5 фонем *[l a m p ə]*;
2. является существительным женского рода в единственном числе, именительном падеже;
3. выполняет в предложении функцию субъекта (подлежащего).

Такими локальными репрезентациями описывается много довольно важных свойств слова *«лампа»*. Но не значение.

Дистрибутивная семантика: как моделировать значение?

А что со значением?

Дистрибутивная семантика: как моделировать значение?

А что со значением?

Следующий шаг: семантика.

Нужно смоделировать, у каких слов **схожее значение**. То есть, придумать такие репрезентации слов, чтобы если слова про одно и то же – то и репрезентации были бы похожими.

Дистрибутивная семантика: как моделировать значение?

А что со значением?

Следующий шаг: семантика.
Нужно смоделировать, у каких слов **схожее значение**. То есть, придумать такие репрезентации слов, чтобы если слова про одно и то же – то и репрезентации были бы похожими.
И вот тут всё оказалось гораздо сложнее. Непонятно, откуда взять значение, и как понять, что *«лампа»* на *«светильник»* похожа, а на *«кипятильник»* – совсем нет?

Дистрибутивная семантика: как моделировать значение?

Где брать данные?

Дистрибутивная семантика: как моделировать значение?

Где брать данные?

Существует два фундаментальных подхода к моделированию семантики:

Дистрибутивная семантика: как моделировать значение?

Где брать данные?

Существует два фундаментальных подхода к моделированию семантики:

- Построение онтологий (knowledge-based approach). Это подход «сверху вниз».

Дистрибутивная семантика: как моделировать значение?

Где брать данные?

Существует два фундаментальных подхода к моделированию семантики:

- Построение онтологий (knowledge-based approach). Это подход «сверху вниз».
- Извлечение значения из употребления слов в текстах (distributional aproach). Это подход «снизу вверх».

Дистрибутивная семантика: как моделировать значение?

Где брать данные?

Существует два фундаментальных подхода к моделированию семантики:

- Построение онтологий (knowledge-based approach). Это подход «сверху вниз».
- Извлечение значения из употребления слов в текстах (distributional aproach). Это подход «снизу вверх».

Подход на онтологиях очень трудоёмкий: нужно силами людей-экспертов строить схему всех понятий, которые мы хотим моделировать. Расширение однажды построенной модели также затруднено.
Мы про него сегодня говорить не будем.

Дистрибутивная семантика: как моделировать значение?

Дистрибутивный подход

Дистрибутивная семантика: как моделировать значение?

Дистрибутивный подход

«Дистрибуция» – это «распределение» (distribution).
Распределение явлений в живой речи.

Дистрибутивная семантика: как моделировать значение?

Дистрибутивный подход

«Дистрибуция» – это «распределение» (distribution). Распределение явлений в живой речи.

Дистрибутивная гипотеза:

Значение лингвистической единицы складывается только из её употребления, использования. В мозге хранится сумма всех тех контекстов, в рамках которых мы слышали или видели то или иное слово. Это и есть его смысл. Без знания типичных соседей никакой семантики нет. Отсюда вывод:

Дистрибутивная семантика: как моделировать значение?

Дистрибутивный подход

«Дистрибуция» – это «распределение» (distribution).
Распределение явлений в живой речи.

Дистрибутивная гипотеза:

Значение лингвистической единицы складывается только из её употребления, использования. В мозге хранится сумма всех тех контекстов, в рамках которых мы слышали или видели то или иное слово. Это и есть его смысл. Без знания типичных соседей никакой семантики нет. Отсюда вывод:
Слова с похожими типичными контекстами имеют схожее значение.

Дистрибутивная семантика: как моделировать значение?

Дистрибутивный подход

«Дистрибуция» – это «распределение» (distribution).
Распределение явлений в живой речи.

Дистрибутивная гипотеза:

Значение лингвистической единицы складывается только из её употребления, использования. В мозге хранится сумма всех тех контекстов, в рамках которых мы слышали или видели то или иное слово. Это и есть его смысл. Без знания типичных соседей никакой семантики нет. Отсюда вывод:
Слова с похожими типичными контекстами имеют схожее значение.
Первые исследования: Зелиг Харрис (40-50 годы XX века), но и до этого идея регулярно появлялась у философов (например, у Витгенштейна).

Дистрибутивная семантика: как моделировать значение?

Тогда что нам нужно, чтобы научить компьютер семантике?

1. Очень много естественных текстов, чем больше, тем лучше. Лингвисты называют такую коллекцию «корпус».

Дистрибутивная семантика: как моделировать значение?

Тогда что нам нужно, чтобы научить компьютер семантике?

1. Очень много естественных текстов, чем больше, тем лучше. Лингвисты называют такую коллекцию «корпус».
2. Модель, описывающая совместную встречаемость слов в этом корпусе.

Дистрибутивная семантика: как моделировать значение?

Тогда что нам нужно, чтобы научить компьютер семантике?

1. Очень много естественных текстов, чем больше, тем лучше. Лингвисты называют такую коллекцию «корпус».
2. Модель, описывающая совместную встречаемость слов в этом корпусе.

Получается очень красиво: гипотетический искусственный интеллект возьмёт семантику прямо из корпуса, без всякого ручного создания сложных карт понятий.

Дистрибутивная семантика: как моделировать значение?

Тогда что нам нужно, чтобы научить компьютер семантике?

1. Очень много естественных текстов, чем больше, тем лучше. Лингвисты называют такую коллекцию «корпус».
2. Модель, описывающая совместную встречаемость слов в этом корпусе.

Получается очень красиво: гипотетический искусственный интеллект возьмёт семантику прямо из корпуса, без всякого ручного создания сложных карт понятий.
Но как представить значение в этой модели?

Содержание

Традиционные счётные модели: матрицы совместной встречаемости

Первый и основной способ представления значения в дистрибутивной семантике – это семантические вектора.
Придумал их американский психолог Чарлз Осгуд, потом идею развивали многие другие.

Традиционные счётные модели: матрицы совместной встречаемости

На выходе у нас примерно такая матрица, где строки и столбцы – это слова из лексикона:

Традиционные счётные модели: матрицы совместной встречаемости

На выходе у нас примерно такая матрица, где строки и столбцы – это слова из лексикона:

	вектор	**значение**	**хомяк**	**семантика**	**суслик**
вектор	0	10	0	8	0
значение	10	0	1	15	0
хомяк	0	1	0	0	20
семантика	8	15	0	0	0
суслик	0	0	20	0	0

Традиционные счётные модели: матрицы совместной встречаемости

На выходе у нас примерно такая матрица, где строки и столбцы – это слова из лексикона:

	вектор	**значение**	**хомяк**	**семантика**	**суслик**
вектор	0	10	0	8	0
значение	10	0	1	15	0
хомяк	0	1	0	0	20
семантика	8	15	0	0	0
суслик	0	0	20	0	0

То есть, слова «вектор» и «значение» в корпусе встретились рядом 10 раз, «хомяк» и «значение» 1 раз, а «вектор» с «хомяком» не стояли рядом ни разу.

Традиционные счётные модели: матрицы совместной встречаемости

На выходе у нас примерно такая матрица, где строки и столбцы – это слова из лексикона:

	вектор	**значение**	**хомяк**	**семантика**	**суслик**
вектор	0	10	0	8	0
значение	10	0	1	15	0
хомяк	0	1	0	0	20
семантика	8	15	0	0	0
суслик	0	0	20	0	0

То есть, слова «вектор» и «значение» в корпусе встретились рядом 10 раз, «хомяк» и «значение» 1 раз, а «вектор» с «хомяком» не стояли рядом ни разу.

Интуитивно вроде бы так и должно быть: вектора в каком-то смысле отражают значение анализируемых слов.

Традиционные счётные модели: матрицы совместной встречаемости

У нас получилось представить семантику в виде численных векторов.

Традиционные счётные модели: матрицы совместной встречаемости

У нас получилось представить семантику в виде численных векторов.

- Многомерное векторное пространство (semantic space).

Традиционные счётные модели: матрицы совместной встречаемости

У нас получилось представить семантику в виде численных векторов.

- Многомерное векторное пространство (semantic space).
- Слова – координатные оси (измерения) этого пространства.

Традиционные счётные модели: матрицы совместной встречаемости

У нас получилось представить семантику в виде численных векторов.

- Многомерное векторное пространство (semantic space).
- Слова – координатные оси (измерения) этого пространства.
- Одновременно они же – вектора или точки в этом пространстве.

Традиционные счётные модели: матрицы совместной встречаемости

У нас получилось представить семантику в виде численных векторов.

- Многомерное векторное пространство (semantic space).
- Слова – координатные оси (измерения) этого пространства.
- Одновременно они же – вектора или точки в этом пространстве.
- В случае с большим корпусом десятки миллионов измерений (осей, слов).

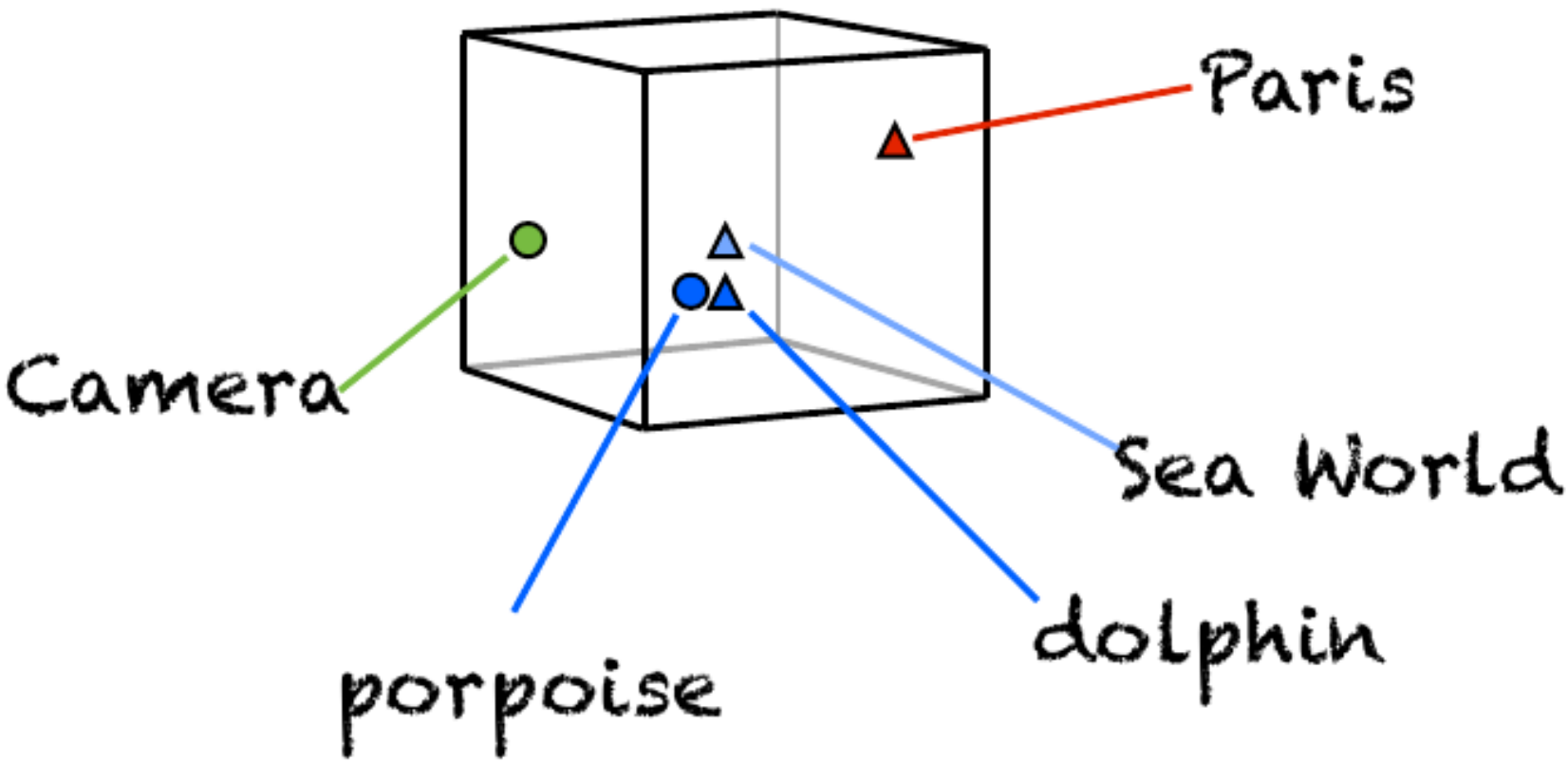
Традиционные счётные модели: матрицы совместной встречаемости

У нас получилось представить семантику в виде численных векторов.

- Многомерное векторное пространство (semantic space).
- Слова – координатные оси (измерения) этого пространства.
- Одновременно они же – вектора или точки в этом пространстве.
- В случае с большим корпусом десятки миллионов измерений (осей, слов).
- Но векторы очень разреженные, большинство компонентов равно нулю.

Традиционные счётные модели: матрицы совместной встречаемости

Слова с похожими соседями будут в этом пространстве рядом:



Традиционные счётные модели: матрицы совместной встречаемости

Для удобства можно использовать не абсолютную частоту совместной встречаемости слов, а как-либо её взвешивать.

Традиционные счётные модели: матрицы совместной встречаемости

Для удобства можно использовать не абсолютную частоту совместной встречаемости слов, а как-либо её взвешивать.Например, коэффициент Дайса:

$$Dice(w,w') = \frac{2c(w,w')}{c(w) + c(w')} \tag{1}$$

где $c(w)$ – абсолютная частота слова w,
$c(w')$ – абсолютная частота слова w'
$c(w,w')$ – частота совместной встречаемости (коллокации) w и w'.

Традиционные счётные модели: матрицы совместной встречаемости

Для удобства можно использовать не абсолютную частоту совместной встречаемости слов, а как-либо её взвешивать.Например, коэффициент Дайса:

$$Dice(w,w') = \frac{2c(w,w')}{c(w) + c(w')} \tag{1}$$

где $c(w)$ – абсолютная частота слова w,
$c(w')$ – абсолютная частота слова w'
$c(w,w')$ – частота совместной встречаемости (коллокации) w и w'.
Есть и другие способы взвешивания: mutual information, log-likelihood и так далее.

Традиционные счётные модели: матрицы совместной встречаемости

При создании матрицы совместной встречаемости (co-occurrence matrix) можно смотреть не только на непосредственных соседей, но и на слова на некотором расстоянии. Например:

Традиционные счётные модели: матрицы совместной встречаемости

При создании матрицы совместной встречаемости (co-occurrence matrix) можно смотреть не только на непосредственных соседей, но и на слова на некотором расстоянии. Например:
*«**Кора** головного мозга — структура головного мозга, слой **серого вещества** толщиной 1,3—4,5 мм, расположенный по периферии **полушарий** большого мозга, и покрывающий их.»*

Традиционные счётные модели: матрицы совместной встречаемости

При создании матрицы совместной встречаемости (co-occurrence matrix) можно смотреть не только на непосредственных соседей, но и на слова на некотором расстоянии. Например:
*«**Кора** головного мозга — структура головного мозга, слой **серого вещества** толщиной 1,3—4,5 мм, расположенный по периферии **полушарий** большого мозга, и покрывающий их.»*
Размер контекста: 2-3 слова. Можно изменять веса в матрице в зависимости от расстояния, на котором находится «сосед»...

Традиционные счётные модели: матрицы совместной встречаемости

При создании матрицы совместной встречаемости (co-occurrence matrix) можно смотреть не только на непосредственных соседей, но и на слова на некотором расстоянии. Например:
*«**Кора** головного мозга — структура головного мозга, слой **серого вещества** толщиной 1,3—4,5 мм, расположенный по периферии **полушарий** большого мозга, и покрывающий их.»*
Размер контекста: 2-3 слова. Можно изменять веса в матрице в зависимости от расстояния, на котором находится «сосед»...или от того, слева он или справа...

Традиционные счётные модели: матрицы совместной встречаемости

При создании матрицы совместной встречаемости (co-occurrence matrix) можно смотреть не только на непосредственных соседей, но и на слова на некотором расстоянии. Например:
*«**Кора** головного мозга — структура головного мозга, слой **серого вещества** толщиной 1,3—4,5 мм, расположенный по периферии **полушарий** большого мозга, и покрывающий их.»*
Размер контекста: 2-3 слова. Можно изменять веса в матрице в зависимости от расстояния, на котором находится «сосед»...или от того, слева он или справа...или в зависимости от типа синтаксической связи между словом и соседом...

Традиционные счётные модели: матрицы совместной встречаемости

При создании матрицы совместной встречаемости (co-occurrence matrix) можно смотреть не только на непосредственных соседей, но и на слова на некотором расстоянии. Например:
*«**Кора** головного мозга — структура головного мозга, слой **серого вещества** толщиной 1,3—4,5 мм, расположенный по периферии **полушарий** большого мозга, и покрывающий их.»*
Размер контекста: 2-3 слова. Можно изменять веса в матрице в зависимости от расстояния, на котором находится «сосед»...или от того, слева он или справа...или в зависимости от типа синтаксической связи между словом и соседом...Много можно придумать разных методов взвешивания.

Традиционные счётные модели: матрицы совместной встречаемости

Итак, слова – это векторы, зависящие от соседей этих слов в том корпусе, который мы выбрали в качестве исходного.

Традиционные счётные модели: матрицы совместной встречаемости

Итак, слова – это векторы, зависящие от соседей этих слов в том корпусе, который мы выбрали в качестве исходного.
NB: когнитивная информация (слова и образы) в мозге хранится в виде паттернов возбуждений нейронов. Очень похоже на векторные репрезентации!

Традиционные счётные модели: матрицы совместной встречаемости

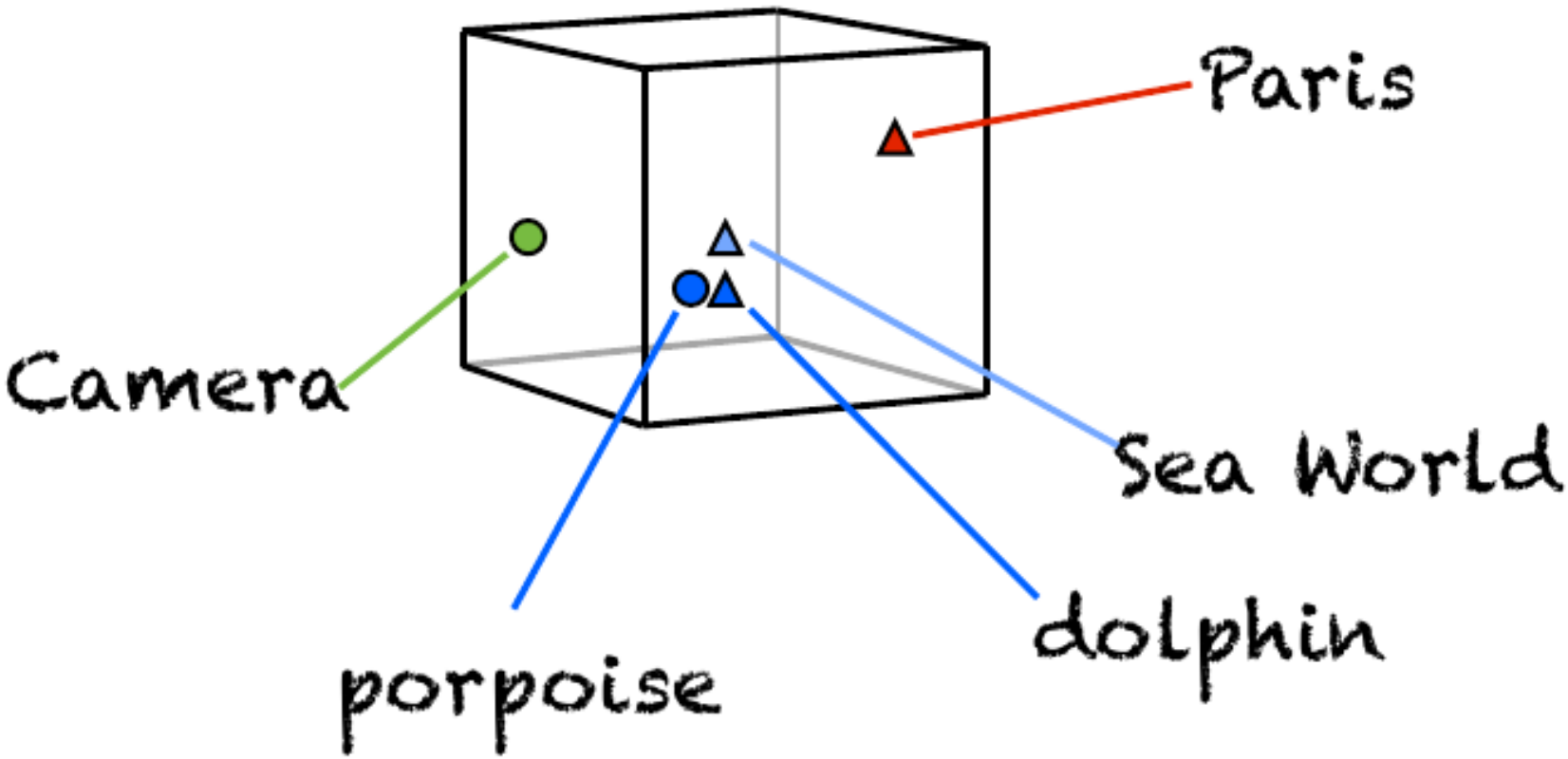
Итак, слова – это векторы, зависящие от соседей этих слов в том корпусе, который мы выбрали в качестве исходного.
NB: когнитивная информация (слова и образы) в мозге хранится в виде паттернов возбуждений нейронов. Очень похоже на векторные репрезентации!

Но что теперь с этими распределенными статистическими репрезентациями делать?
Если слова – это векторы, то теперь мы можем сравнивать их друг с другом чисто математическими методами.

Содержание

Как вычислить семантическую близость?

Ещё раз: похожие векторы в семантическом пространстве оказываются рядом



Как вычислить семантическую близость?

Классический способ определения семантической близости слов в векторном пространстве – через косинусную близость векторов.

Как вычислить семантическую близость?

Классический способ определения семантической близости слов в векторном пространстве – через косинусную близость векторов.

- Схожесть ниже по мере увеличения угла между векторами слов.

Как вычислить семантическую близость?

Классический способ определения семантической близости слов в векторном пространстве – через косинусную близость векторов.

- Схожесть ниже по мере увеличения угла между векторами слов.
- Схожесть выше по мере увеличения косинуса угла между векторами слов.

Как вычислить семантическую близость?

Классический способ определения семантической близости слов в векторном пространстве – через косинусную близость векторов.

- Схожесть ниже по мере увеличения угла между векторами слов.
- Схожесть выше по мере увеличения косинуса угла между векторами слов.

NB: Косинус – это монотонно убывающая функция в интервале от 0 до 180 градусов.

Как вычислить семантическую близость?

Собственно формула косинусной близости

Как вычислить семантическую близость?

Собственно формула косинусной близости

$$cos(w1,w2) = \frac{\vec{V}(w1) \times \vec{V}(w2)}{|\vec{V}(w1)| \times |\vec{V}(w2)|}$$

Как вычислить семантическую близость?

Собственно формула косинусной близости

$$cos(w1,w2) = \frac{\vec{V}(w1) \times \vec{V}(w2)}{|\vec{V}(w1)| \times |\vec{V}(w2)|}$$

- Числитель: скалярное произведение векторов слов *w1* и *w2*

Как вычислить семантическую близость?

Вернёмся к нашим хомякам:

Как вычислить семантическую близость?

Вернёмся к нашим хомякам:

	вектор	**значение**	**хомяк**	**семантика**	**суслик**
вектор	0	10	0	8	0
значение	10	0	1	15	0
хомяк	0	1	0	0	20
семантика	8	15	0	0	0
суслик	0	0	20	0	0

Как вычислить семантическую близость?

Вернёмся к нашим хомякам:

	вектор	**значение**	**хомяк**	**семантика**	**суслик**
вектор	0	10	0	8	0
значение	10	0	1	15	0
хомяк	0	1	0	0	20
семантика	8	15	0	0	0
суслик	0	0	20	0	0

Сначала нормируем векторы по длине.

Как вычислить семантическую близость?

Нормированные векторы:

	вектор	**значение**	**хомяк**	**семантика**	**суслик**
вектор	0	0.78	0	0.625	0
значение	0.55	0	0.055	0.83	0
хомяк	0	0.05	0	0	0.99
семантика	0.47	0.88	0	0	0
суслик	0	0	1	0	0

Как вычислить семантическую близость?

Нормированные векторы:

	вектор	**значение**	**хомяк**	**семантика**	**суслик**
вектор	0	0.78	0	0.625	0
значение	0.55	0	0.055	0.83	0
хомяк	0	0.05	0	0	0.99
семантика	0.47	0.88	0	0	0
суслик	0	0	1	0	0

Вычисляем косинусную близость:

$cos(\text{вектор}, \text{значение}) =$

Как вычислить семантическую близость?

Нормированные векторы:

	вектор	**значение**	**хомяк**	**семантика**	**суслик**
вектор	0	0.78	0	0.625	0
значение	0.55	0	0.055	0.83	0
хомяк	0	0.05	0	0	0.99
семантика	0.47	0.88	0	0	0
суслик	0	0	1	0	0

Вычисляем косинусную близость:

$cos(\text{вектор}, \text{значение}) =$

$0 \times 0.55 + 0.78 \times 0 + 0 \times 0.055 + 0.625 \times 0.83 + 0 \times 0 =$

Как вычислить семантическую близость?

Нормированные векторы:

	вектор	**значение**	**хомяк**	**семантика**	**суслик**
вектор	0	0.78	0	0.625	0
значение	0.55	0	0.055	0.83	0
хомяк	0	0.05	0	0	0.99
семантика	0.47	0.88	0	0	0
суслик	0	0	1	0	0

Вычисляем косинусную близость:

$cos(\text{вектор}, \text{значение}) =$

$0 \times 0.55 + 0.78 \times 0 + 0 \times 0.055 + 0.625 \times 0.83 + 0 \times 0 = 0.519$

Как вычислить семантическую близость?

Нормированные векторы:

	вектор	**значение**	**хомяк**	**семантика**	**суслик**
вектор	0	0.78	0	0.625	0
значение	0.55	0	0.055	0.83	0
хомяк	0	0.05	0	0	0.99
семантика	0.47	0.88	0	0	0
суслик	0	0	1	0	0

Вычисляем косинусную близость:

$cos(\text{вектор}, \text{значение}) =$

$0 \times 0.55 + 0.78 \times 0 + 0 \times 0.055 + 0.625 \times 0.83 + 0 \times 0 = 0.519$

$cos(\text{вектор}, \text{хомяк}) =$

Как вычислить семантическую близость?

Нормированные векторы:

	вектор	**значение**	**хомяк**	**семантика**	**суслик**
вектор	0	0.78	0	0.625	0
значение	0.55	0	0.055	0.83	0
хомяк	0	0.05	0	0	0.99
семантика	0.47	0.88	0	0	0
суслик	0	0	1	0	0

Вычисляем косинусную близость:

$cos(\text{вектор}, \text{значение}) =$

$0 \times 0.55 + 0.78 \times 0 + 0 \times 0.055 + 0.625 \times 0.83 + 0 \times 0 = 0.519$

$cos(\text{вектор}, \text{хомяк}) =$

$0 \times 0 + 0.78 \times 0.05 + 0 \times 0 + 0.625 \times 0 + 0 \times 0.99 =$

Как вычислить семантическую близость?

Нормированные векторы:

	вектор	**значение**	**хомяк**	**семантика**	**суслик**
вектор	0	0.78	0	0.625	0
значение	0.55	0	0.055	0.83	0
хомяк	0	0.05	0	0	0.99
семантика	0.47	0.88	0	0	0
суслик	0	0	1	0	0

Вычисляем косинусную близость:

$cos(\text{вектор}, \text{значение}) =$

$0 \times 0.55 + 0.78 \times 0 + 0 \times 0.055 + 0.625 \times 0.83 + 0 \times 0 = 0.519$

$cos(\text{вектор}, \text{хомяк}) =$

$0 \times 0 + 0.78 \times 0.05 + 0 \times 0 + 0.625 \times 0 + 0 \times 0.99 = 0.039$

Как вычислить семантическую близость?

Нормированные векторы:

	вектор	**значение**	**хомяк**	**семантика**	**суслик**
вектор	0	0.78	0	0.625	0
значение	0.55	0	0.055	0.83	0
хомяк	0	0.05	0	0	0.99
семантика	0.47	0.88	0	0	0
суслик	0	0	1	0	0

Вычисляем косинусную близость:

$cos(\text{вектор}, \text{значение}) =$

$0 \times 0.55 + 0.78 \times 0 + 0 \times 0.055 + 0.625 \times 0.83 + 0 \times 0 = 0.519$

$cos(\text{вектор}, \text{хомяк}) =$

$0 \times 0 + 0.78 \times 0.05 + 0 \times 0 + 0.625 \times 0 + 0 \times 0.99 = 0.039$

Ура, вектор ближе к значению, чем к хомяку. Теперь мы умеем моделировать семантику при помощи векторных пространств и отличать похожие слова от непохожих. В дистрибутивной семантике такая модель называется «счётная» (count model).

Содержание

Недостатки счётных моделей

Счётные модели хороши, но у них есть глобальные недостатки:

Недостатки счётных моделей

Счётные модели хороши, но у них есть глобальные недостатки:

- Размер векторов получается огромным (в общем случае равен объёму лексикона).

Недостатки счётных моделей

Счётные модели хороши, но у них есть глобальные недостатки:

- Размер векторов получается огромным (в общем случае равен объёму лексикона).
- Это очень замедляет операции сравнения векторов.

Недостатки счётных моделей

Счётные модели хороши, но у них есть глобальные недостатки:

- Размер векторов получается огромным (в общем случае равен объёму лексикона).
- Это очень замедляет операции сравнения векторов.
- Можно применять различные методы снижения размерности (PCA, SVD и т.п.), но часто страдает качество.

Недостатки счётных моделей

Счётные модели хороши, но у них есть глобальные недостатки:

- Размер векторов получается огромным (в общем случае равен объёму лексикона).
- Это очень замедляет операции сравнения векторов.
- Можно применять различные методы снижения размерности (PCA, SVD и т.п.), но часто страдает качество.
- Мы не знаем точно, что в наших векторах нужная информация, а что мусор. Они просто взяты из корпуса «как есть».

Недостатки счётных моделей

Счётные модели хороши, но у них есть глобальные недостатки:

- Размер векторов получается огромным (в общем случае равен объёму лексикона).
- Это очень замедляет операции сравнения векторов.
- Можно применять различные методы снижения размерности (PCA, SVD и т.п.), но часто страдает качество.
- Мы не знаем точно, что в наших векторах нужная информация, а что мусор. Они просто взяты из корпуса «как есть».

Поэтому следующий шаг – применение для получения хороших векторов машинного обучения.

Содержание

Предсказательные модели и нейронные сети: выучиваем вектора

Машинное обучение

Предсказательные модели и нейронные сети: выучиваем вектора

Машинное обучение

Некоторые задачи настолько сложны, что мы не можем сформулировать алгоритм для программы. Мы сами точно не знаем, как наш мозг это делает.

Предсказательные модели и нейронные сети: выучиваем вектора

Машинное обучение

Некоторые задачи настолько сложны, что мы не можем сформулировать алгоритм для программы. Мы сами точно не знаем, как наш мозг это делает.

Для решения таких задач применяется машинное обучение: попытка построить программу, которая будет самостоятельно обучаться правильным решениям на каком-то тренировочном материале, который мы ей дадим.

Предсказательные модели и нейронные сети: выучиваем вектора

Машинное обучение

Некоторые задачи настолько сложны, что мы не можем сформулировать алгоритм для программы. Мы сами точно не знаем, как наш мозг это делает.

Для решения таких задач применяется машинное обучение: попытка построить программу, которая будет самостоятельно обучаться правильным решениям на каком-то тренировочном материале, который мы ей дадим.

Один из популярнейших методов машинного обучения для задач языкового моделирования – искусственные нейронные сети.

Предсказательные модели и нейронные сети: выучиваем вектора

Как работает мозг

Предсказательные модели и нейронные сети: выучиваем вектора

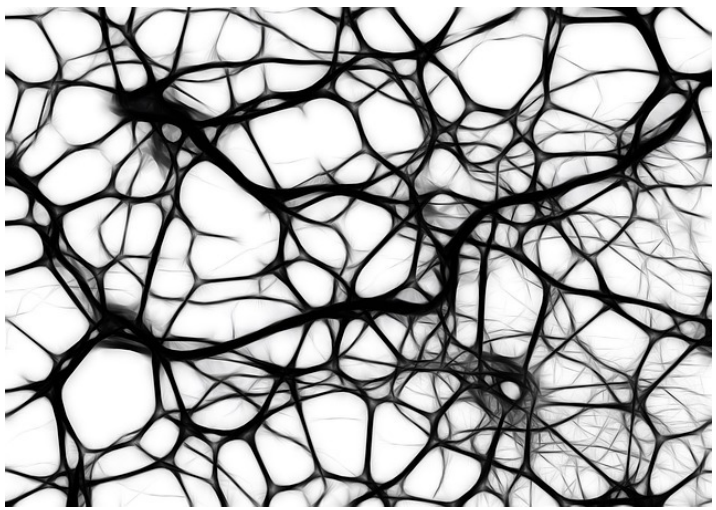
Как работает мозг

В мозгу 10^{11} нейронов, и 10^4 связей у каждого из них.

Предсказательные модели и нейронные сети: выучиваем вектора

Как работает мозг

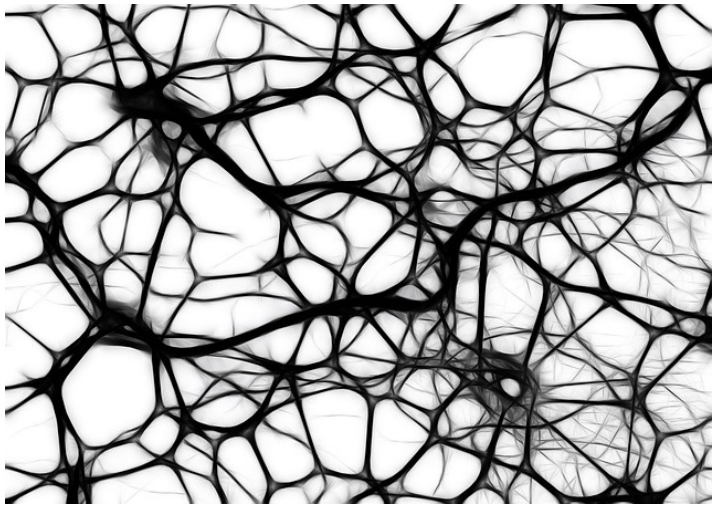
В мозгу 10^{11} нейронов, и 10^4 связей у каждого из них. На нейроны поступают сигналы разного веса с других нейронов. Нейрон реагирует в зависимости от того, что поступило на вход.



Предсказательные модели и нейронные сети: выучиваем вектора

Как работает мозг

В мозгу 10^{11} нейронов, и 10^4 связей у каждого из них.
На нейроны поступают сигналы разного веса с других нейронов. Нейрон реагирует в зависимости от того, что поступило на вход.



Искусственные нейронные сети пытаются имитировать этот процесс.

Предсказательные модели и нейронные сети: выучиваем вектора

В дистрибутивной семантике модели на основе машинного обучения называются предсказательными (predict models). Если в счётных моделях мы подсчитываем частоту встречаемости и воспринимаем её как вектор, то в предсказательных моделях всё наоборот:

Предсказательные модели и нейронные сети: выучиваем вектора

В дистрибутивной семантике модели на основе машинного обучения называются предсказательными (predict models).
Если в счётных моделях мы подсчитываем частоту встречаемости и воспринимаем её как вектор, то в предсказательных моделях всё наоборот:
Мы пытаемся для каждого слова найти такой вектор, чтобы он был максимально схож с векторами типичных соседей и максимально отличался от векторов слов, которые соседями данному слову не являются.

Предсказательные модели и нейронные сети: выучиваем вектора

В дистрибутивной семантике модели на основе машинного обучения называются предсказательными (predict models).
Если в счётных моделях мы подсчитываем частоту встречаемости и воспринимаем её как вектор, то в предсказательных моделях всё наоборот:
Мы пытаемся для каждого слова найти такой вектор, чтобы он был максимально схож с векторами типичных соседей и максимально отличался от векторов слов, которые соседями данному слову не являются.
В такой модели вектор обычно небольшой размерности (порядка сотен компонентов), по-английски он называется embedding.

Предсказательные модели и нейронные сети: выучиваем вектора

В счётных моделях вектор для слова «хомяк» будет выглядеть так:

Предсказательные модели и нейронные сети: выучиваем вектора

В счётных моделях вектор для слова «хомяк» будет выглядеть так:
$\vec{\text{хомяк}} = [w_1, w_2, w_3 ... w_n]$, где n – это количество слов в лексиконе (например, миллион).

Предсказательные модели и нейронные сети: выучиваем вектора

В счётных моделях вектор для слова «хомяк» будет выглядеть так:
$\vec{\text{хомяк}} = [w_1, w_2, w_3 ... w_n]$, где n – это количество слов в лексиконе (например, миллион).
А в предсказательных моделях так:

Предсказательные модели и нейронные сети: выучиваем вектора

В счётных моделях вектор для слова «хомяк» будет выглядеть так:
$\vec{\text{хомяк}} = [w_1, w_2, w_3 ... w_n]$, где n – это количество слов в лексиконе (например, миллион).
А в предсказательных моделях так:
$\vec{\text{хомяк}} = [w_1, w_2, w_3 ... w_n]$, где n – это заданный при обучении размер (например, около 500).

Предсказательные модели и нейронные сети: выучиваем вектора

В счётных моделях вектор для слова «хомяк» будет выглядеть так:
$\vec{\text{хомяк}} = [w_1, w_2, w_3 ... w_n]$, где n – это количество слов в лексиконе (например, миллион).
А в предсказательных моделях так:
$\vec{\text{хомяк}} = [w_1, w_2, w_3 ... w_n]$, где n – это заданный при обучении размер (например, около 500).
В счётных моделях компоненты векторов постепенно наполняются значениями частоты совместной встречаемости.

Предсказательные модели и нейронные сети: выучиваем вектора

В счётных моделях вектор для слова «хомяк» будет выглядеть так:
$\vec{\text{хомяк}} = [w_1, w_2, w_3 ... w_n]$, где n – это количество слов в лексиконе (например, миллион).
А в предсказательных моделях так:
$\vec{\text{хомяк}} = [w_1, w_2, w_3 ... w_n]$, где n – это заданный при обучении размер (например, около 500).
В счётных моделях компоненты векторов постепенно наполняются значениями частоты совместной встречаемости.
А в предсказательных моделях вектора сначала инициализируются случайным образом. Но постепенно модель обучается и сходится. Тогда вектора у семантически близких слов становятся похожими.

Предсказательные модели и нейронные сети: выучиваем вектора

Важно: в предсказательных моделях конкретные компоненты векторов (например, w_5) никак не связаны с конкретными словами, как в счётных моделях. Это некие обобщённые «свойства семантического пространства».

Предсказательные модели и нейронные сети: выучиваем вектора

Важно: в предсказательных моделях конкретные компоненты векторов (например, w_5) никак не связаны с конкретными словами, как в счётных моделях. Это некие обобщённые «свойства семантического пространства».
Сначала это кажется парадоксальным.

Предсказательные модели и нейронные сети: выучиваем вектора

Важно: в предсказательных моделях конкретные компоненты векторов (например, w_5) никак не связаны с конкретными словами, как в счётных моделях. Это некие обобщённые «свойства семантического пространства».
Сначала это кажется парадоксальным.
Но, если подумать: наша цель – это создать такие репрезентации для слов, которые потом можно будет использовать в практических приложениях. Неважно, что «значат» конкретные компоненты, важно, что в целом модель хорошо представляет семантическое пространство языка.

Предсказательные модели и нейронные сети: выучиваем вектора

Важно: в предсказательных моделях конкретные компоненты векторов (например, w_5) никак не связаны с конкретными словами, как в счётных моделях. Это некие обобщённые «свойства семантического пространства».
Сначала это кажется парадоксальным.
Но, если подумать: наша цель – это создать такие репрезентации для слов, которые потом можно будет использовать в практических приложениях. Неважно, что «значат» конкретные компоненты, важно, что в целом модель хорошо представляет семантическое пространство языка.
Если мы можем посмотреть на репрезентацию слова и сказать, какие слова к нему близки по смыслу – цель достигнута.
Нейронная сеть обучилась семантике.

Содержание

Революция word2vec

В 2013 году исследователь Tomas Mikolov из Google с соавторами опубликовал статью *Efficient Estimation of Word Representations in Vector Space*, а чуть позже выложил код утилиты *word2vec*, которая позволяет тренировать нейронные языковые модели на больших корпусах.

Революция word2vec

В 2013 году исследователь Tomas Mikolov из Google с соавторами опубликовал статью *Efficient Estimation of Word Representations in Vector Space*, а чуть позже выложил код утилиты *word2vec*, которая позволяет тренировать нейронные языковые модели на больших корпусах.

- `http://arxiv.org/abs/1301.3781`
- `https://code.google.com/p/word2vec/`

Революция word2vec

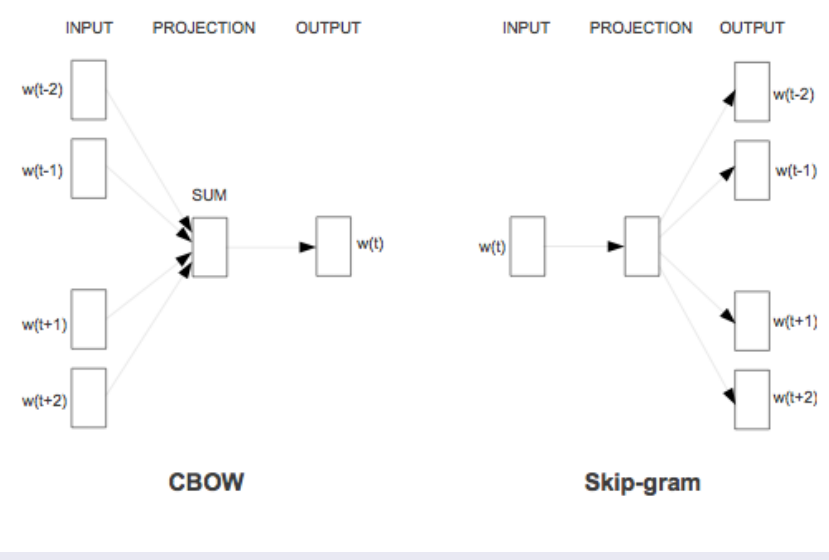
В 2013 году исследователь Tomas Mikolov из Google с соавторами опубликовал статью *Efficient Estimation of Word Representations in Vector Space*, а чуть позже выложил код утилиты *word2vec*, которая позволяет тренировать нейронные языковые модели на больших корпусах.

- `http://arxiv.org/abs/1301.3781`
- `https://code.google.com/p/word2vec/`

Миколов модифицировал существовавшие до этого алгоритмы: удалил из сети скрытый слой, использовал при обучении иерархический софтмакс и ещё несколько вещей. Важно, что за счет этого *word2vec* обучается на порядки быстрее, чем нейронные языковые модели до него.

Революция word2vec

Continuous Bag-of-Words и Continuous Skip-Gram: два основных алгоритма word2vec



При обучении CBOW тренируется предсказывать слово на основании его окружения, а Skip-Gram – предсказывать окружение на основании слова.

Революция word2vec

В итоге порог вхождения в предсказательную дистрибутивную семантику значительно снизился. Теперь практически любой может взять большой корпус, взять *word2vec* и получать хорошие вектора (вы тоже можете). Легко получать списки синонимов/ассоциатов:

Революция word2vec

В итоге порог вхождения в предсказательную дистрибутивную семантику значительно снизился. Теперь практически любой может взять большой корпус, взять *word2vec* и получать хорошие вектора (вы тоже можете). Легко получать списки синонимов/ассоциатов:

динозавр

1. *мамонт* 0.397899210453
2. *рептилия* 0.360172241926
3. *млекопитающее* 0.328677803278
4. *ящерица* 0.326320767403
5. *птеродактиль* 0.320571988821

Революция word2vec

В итоге порог вхождения в предсказательную дистрибутивную семантику значительно снизился. Теперь практически любой может взять большой корпус, взять *word2vec* и получать хорошие вектора (вы тоже можете). Легко получать списки синонимов/ассоциатов:

динозавр

1. *мамонт* 0.397899210453
2. *рептилия* 0.360172241926
3. *млекопитающее* 0.328677803278
4. *ящерица* 0.326320767403
5. *птеродактиль* 0.320571988821

Вопрос: что за цифры после слов?

Революция word2vec

В итоге порог вхождения в предсказательную дистрибутивную семантику значительно снизился. Теперь практически любой может взять большой корпус, взять *word2vec* и получать хорошие вектора (вы тоже можете). Легко получать списки синонимов/ассоциатов:

динозавр

1. *мамонт* 0.397899210453
2. *рептилия* 0.360172241926
3. *млекопитающее* 0.328677803278
4. *ящерица* 0.326320767403
5. *птеродактиль* 0.320571988821

Вопрос: что за цифры после слов?

Ответ: конечно, косинусная близость с вектором слова *«динозавр»*.

Содержание

Что знает модель: отношения между словами

Модели, натренированные на огромных корпусах (миллиарды слов), демонстрируют удивительные свойства: алгебраические операции на векторах отражают операции семантические.

Что знает модель: отношения между словами

Модели, натренированные на огромных корпусах (миллиарды слов), демонстрируют удивительные свойства: алгебраические операции на векторах отражают операции семантические.

Операции

Если мы вычтем из вектора слова *Париж* вектор слова *Франция* и прибавим вектор слова *Германия*, получится вектор слова *Берлин* (точнее, он будет ближайшим к получившемуся вектору).

Что знает модель: отношения между словами

Модели, натренированные на огромных корпусах (миллиарды слов), демонстрируют удивительные свойства: алгебраические операции на векторах отражают операции семантические.

Операции

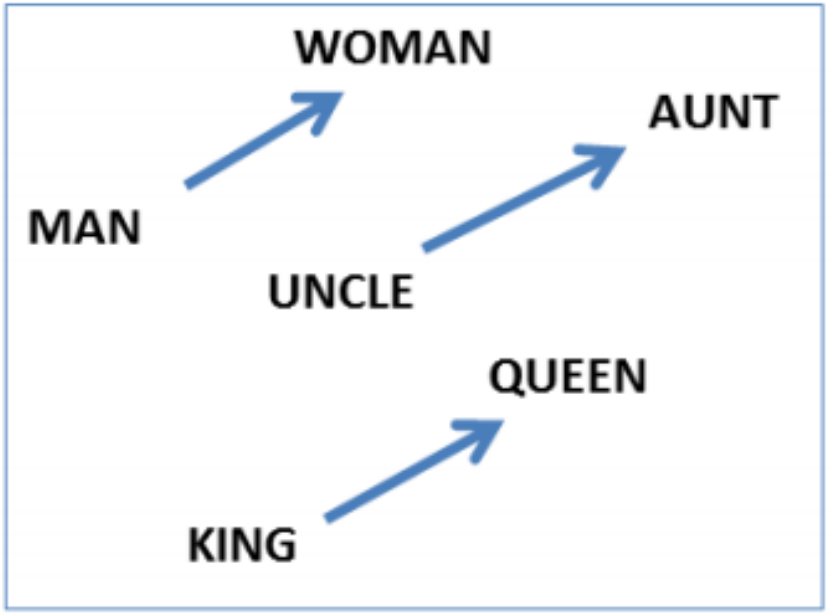
Если мы вычтем из вектора слова *Париж* вектор слова *Франция* и прибавим вектор слова *Германия*, получится вектор слова *Берлин* (точнее, он будет ближайшим к получившемуся вектору).
Из любого набора слов можно легко убирать «лишние», простым сравнением векторов: удаляем наиболее далекие от среднего. Например, модель легко определяет лишнее слово в наборе «*джихад, мечеть, мусульманин, пастор*».

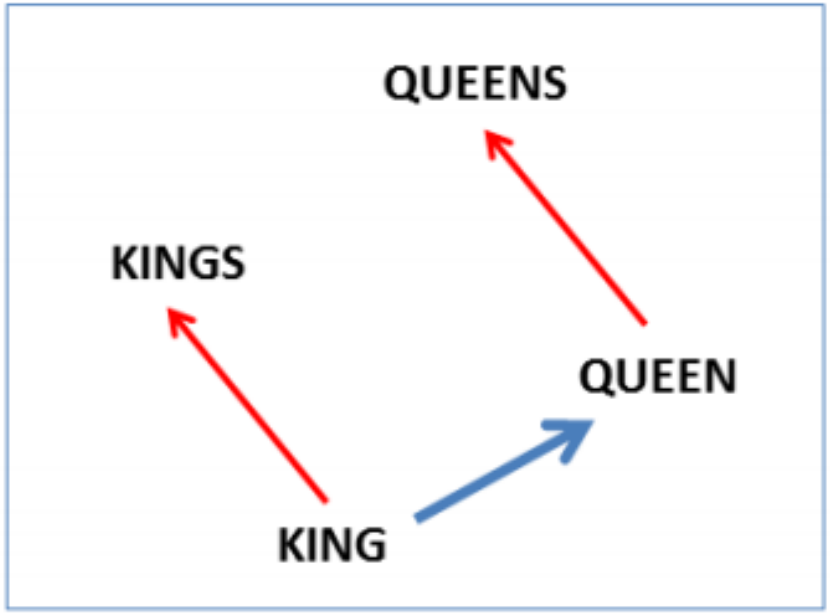
Что знает модель: отношения между словами

Это открывает потрясающие перспективы для любых приложений, связанных со смыслом. Фактически, мы видим семантические отношения в системе языка, можем их «потрогать». *«Женщина»* относится к *«мужчине»* как *«королева»* к *«королю»* и так далее.

Что знает модель: отношения между словами

Это открывает потрясающие перспективы для любых приложений, связанных со смыслом. Фактически, мы видим семантические отношения в системе языка, можем их «потрогать». *«Женщина»* относится к *«мужчине»* как *«королева»* к *«королю»* и так далее.

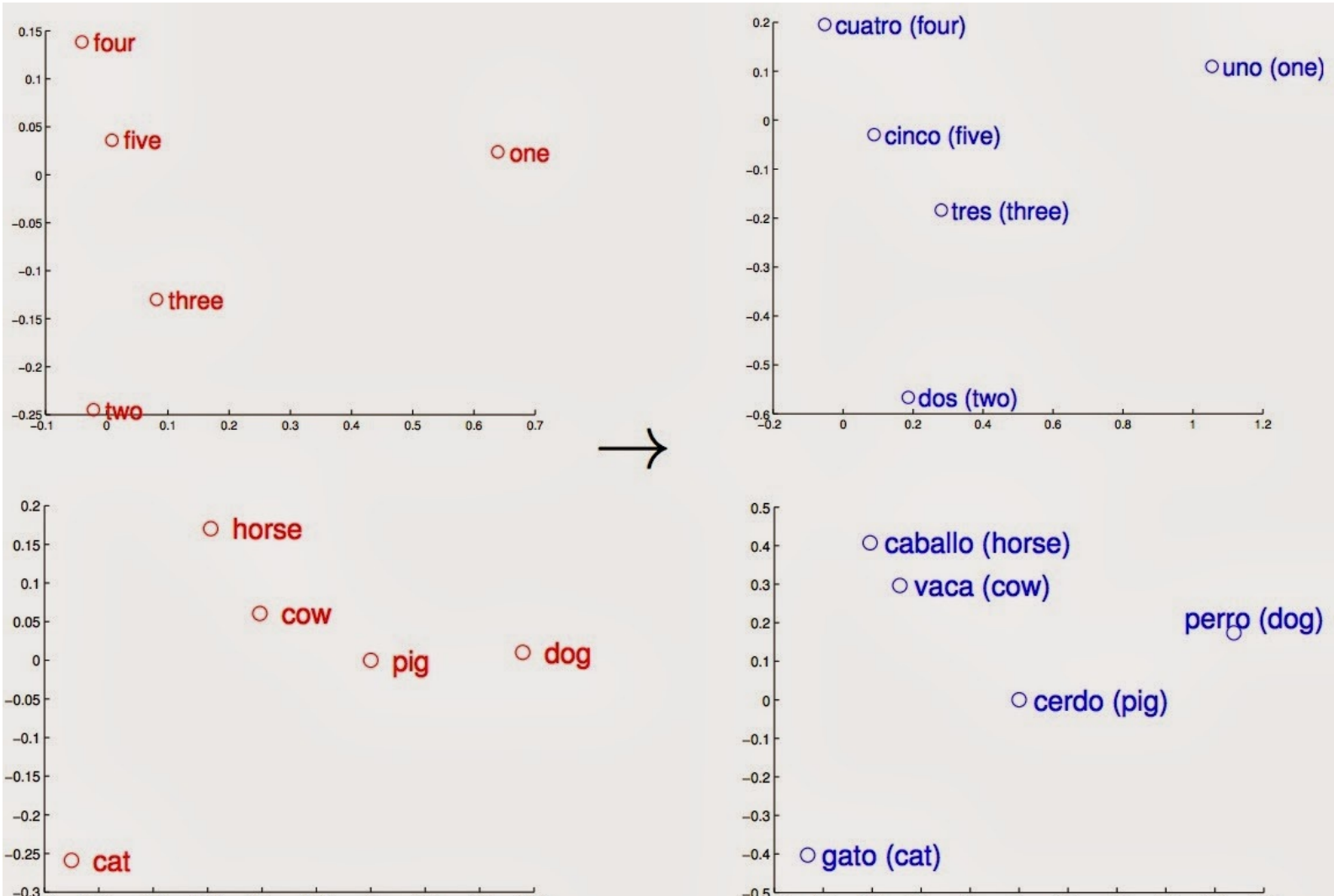




(Mikolov et al., NAACL HLT, 2013)

Что знает модель: отношения между словами

Семантические структуры воспроизводятся даже в разных языках:



Содержание

Для чего применяются векторные репрезентации

- вычисление семантической близости (некая общая задача, строительный кирпичик); недавно прошло первое соревнование по вычислению семантической близости для русского языка (`http://russe.nlpub.ru`)

Для чего применяются векторные репрезентации

- вычисление семантической близости (некая общая задача, строительный кирпичик); недавно прошло первое соревнование по вычислению семантической близости для русского языка (`http://russe.nlpub.ru`)
- машинный перевод (ищем похожие слова в разных языках);

Для чего применяются векторные репрезентации

- вычисление семантической близости (некая общая задача, строительный кирпичик); недавно прошло первое соревнование по вычислению семантической близости для русского языка (`http://russe.nlpub.ru`)
- машинный перевод (ищем похожие слова в разных языках);
- расширение поисковых запросов (ищем похожие слова или фразы);

Для чего применяются векторные репрезентации

- вычисление семантической близости (некая общая задача, строительный кирпичик); недавно прошло первое соревнование по вычислению семантической близости для русского языка (`http://russe.nlpub.ru`)
- машинный перевод (ищем похожие слова в разных языках);
- расширение поисковых запросов (ищем похожие слова или фразы);
- классификация текстов на заранее заданные категории;

Для чего применяются векторные репрезентации

- вычисление семантической близости (некая общая задача, строительный кирпичик); недавно прошло первое соревнование по вычислению семантической близости для русского языка (`http://russe.nlpub.ru`)
- машинный перевод (ищем похожие слова в разных языках);
- расширение поисковых запросов (ищем похожие слова или фразы);
- классификация текстов на заранее заданные категории;
- кластеризация текстов на заранее неизвестные категории;

Для чего применяются векторные репрезентации

- вычисление семантической близости (некая общая задача, строительный кирпичик); недавно прошло первое соревнование по вычислению семантической близости для русского языка (`http://russe.nlpub.ru`)
- машинный перевод (ищем похожие слова в разных языках);
- расширение поисковых запросов (ищем похожие слова или фразы);
- классификация текстов на заранее заданные категории;
- кластеризация текстов на заранее неизвестные категории;
- определение тональности высказывания (положительный отзыв или отрицательный);

Для чего применяются векторные репрезентации

- вычисление семантической близости (некая общая задача, строительный кирпичик); недавно прошло первое соревнование по вычислению семантической близости для русского языка (`http://russe.nlpub.ru`)
- машинный перевод (ищем похожие слова в разных языках);
- расширение поисковых запросов (ищем похожие слова или фразы);
- классификация текстов на заранее заданные категории;
- кластеризация текстов на заранее неизвестные категории;
- определение тональности высказывания (положительный отзыв или отрицательный);
- ...и многое другое.

Содержание

Ближайшее будущее

Над чем сейчас работают в этой области

1. Извлечение информации о том, какие соседи всё-таки были наиболее значимыми для обучения семантическим свойствам того или иного слова.

Ближайшее будущее

Над чем сейчас работают в этой области

1. Извлечение информации о том, какие соседи всё-таки были наиболее значимыми для обучения семантическим свойствам того или иного слова.
2. Композициональная дистрибутивная семантика: как в терминах предсказательных моделей описывать словосочетания? А предложения? А целые тексты? Всегда ли вектор словосочетания равен сумме или скалярному произведению векторов его частей?

Ближайшее будущее

Над чем сейчас работают в этой области

1. Извлечение информации о том, какие соседи всё-таки были наиболее значимыми для обучения семантическим свойствам того или иного слова.
2. Композициональная дистрибутивная семантика: как в терминах предсказательных моделей описывать словосочетания? А предложения? А целые тексты? Всегда ли вектор словосочетания равен сумме или скалярному произведению векторов его частей?
3. Соединение векторных представлений слов и изображений.