Word2vec, синтаксические парсеры и автокодировщики

М. В. Кузнецова

Московский физико-технический институт

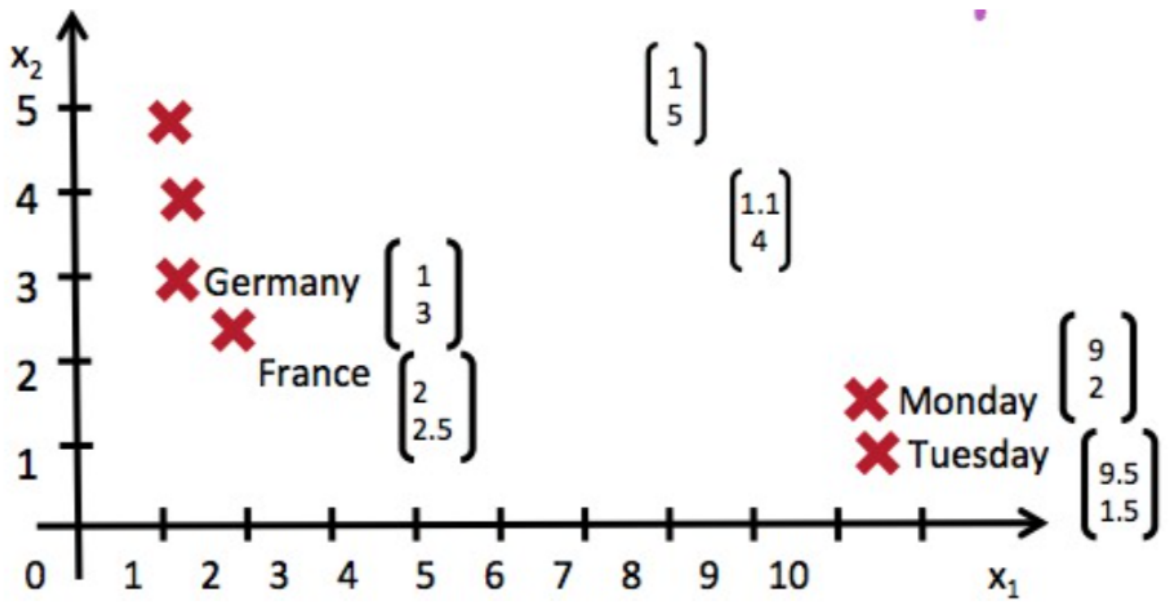
15.10.2015

План презентации

1. Word2vec
2. Синтаксические парсеры
3. Автокодировщики

Дистрибутивная гипотеза

- Лингвистические единицы, встречающиеся в схожих контекстах, имеют близкие значения.
- На основе анализа больших объемов текста можно попытаться каждому слову из словаря сопоставить вектор в пространстве $\mathbb{R}^n$.



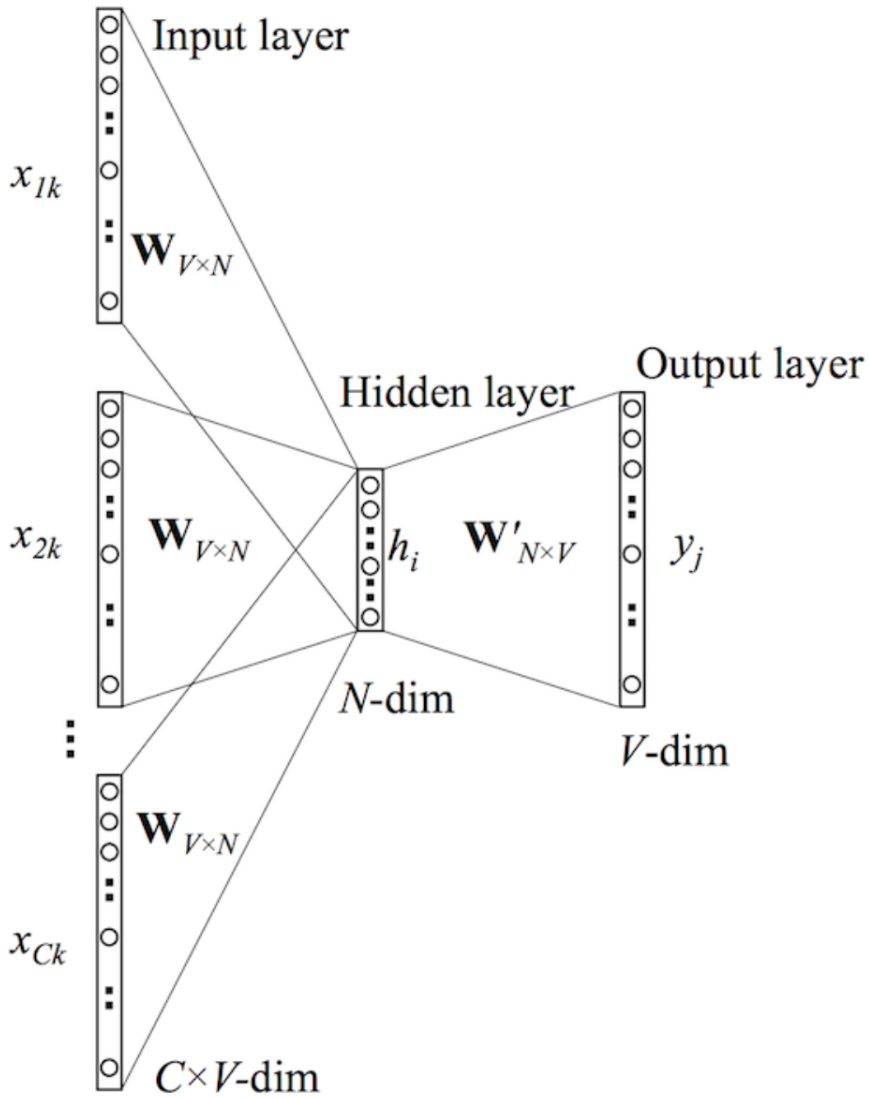
Введение

Word2vec — инструмент для вычисления векторных репрезентаций слов. Для этого использует две модели — Continuous Bag-of-Words и Skip-Gram.

Обозначения:

- Каждое слово в корпусе кодируется V -мерным бинарным вектором длины 1, где V —– размер словаря корпуса.
- $\mathbf{W}_{V \times N}$ — матрица весов между входным и скрытым слоем. i -я строка $\mathbf{W}$ соответствует i -му слову в словаре.
- $\mathbf{h}_{1 \times V}$ — вектор скрытого слоя.
- $\mathbf{W}'_{V \times N}$ — матрица весов между скрытым и выходным слоем.

Continuous Bag-of-Words Model, Mikolov et al., 2013

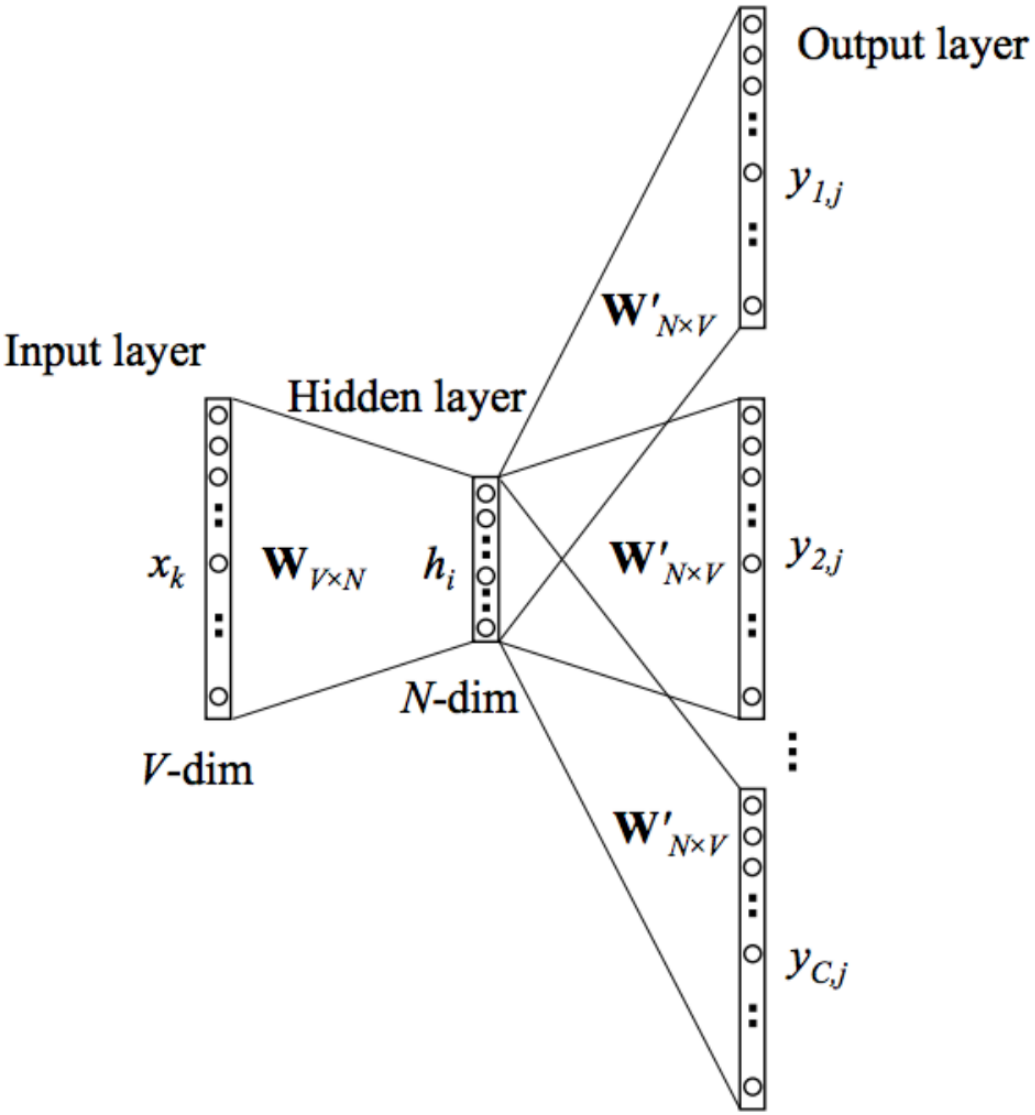


Continuous Bag-of-Words Model

- Просматриваем обучающий текстовый корпус окном ширины C.
- По входному слову w_t предсказыватся его контекст $w_{t-C}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+C}$.
- Максимизируется функция вида:

$$L = \sum_{t,C} \log P(w_t | w_{t-C}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+C}).$$

The Skip-Gram Model



The Skip-Gram Model

- Просматриваем обучающий текстовый корпус окном ширины C.
- По входному слову w_t предсказываем его контекст $w_{t-C}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+C}$.
- Максимизируется функция вида:

$$L = \sum_{t,C} \sum_{j \in context_C(t)} \log P(w_j | w_t).$$

Models

CBOW Model

- $\mathbf{h} = \frac{1}{C}\mathbf{W}\sum_{i=1}^{C}\mathbf{x}_i = \mathbf{W}_{(k,\cdot)}$

The Skip-Gram Model

- $\mathbf{h} = \mathbf{x}^{\intercal}\mathbf{W} = \mathbf{W}_{(k,\cdot)}$

Из скрытого слоя в выходной : $\mathbf{u} = \mathbf{h}^{\intercal}\mathbf{W}^{'}$.

Каждая компонента u_j вектора $\mathbf{u}$ — "вклад"каждого слова в словаре.

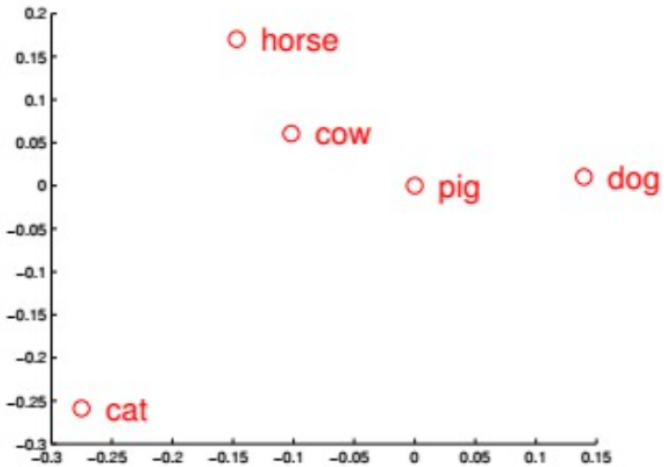
Для **CBOW Model**: $p(w_I|w_{C,j} = w_{O,C}) = \frac{exp(u_{C,j})}{\sum_{j'=1}^{V} exp(u_j^{'})}$

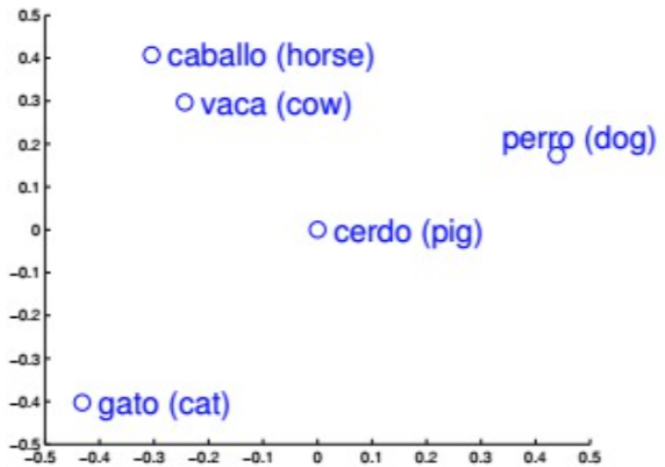
Для **The Skip-Gram Model**: $p(w_{C,j} = w_{O,C}|w_I) = \frac{exp(u_{C,j})}{\sum_{j'=1}^{V} exp(u_j^{'})}$

Примеры ближайших слов

FRANCE	JESUS	XBOX	REDDISH	SCRATCHED	MEGABITS
AUSTRIA	GOD	AMIGA	GREENISH	NAILED	OCTETS
BELGIUM	SATI	PLAYSTATION	BLUISH	SMASHED	MB/S
GERMANY	CHRIST	MSX	PINKISH	PUNCHED	BIT/S
ITALY	SATAN	IPOD	PURPLISH	POPPED	BAUD
GREECE	KALI	SEGA	BROWNISH	CRIMPED	CARATS
SWEDEN	INDRA	PSNUMBER	GREYISH	SCRAPED	KBIT/S
NORWAY	VISHNU	HD	GRAYISH	SCREWED	MEGAHERTZ
EUROPE	ANANDA	DREAMCAST	WHITISH	SECTIONED	MEGAPIXELS
HUNGARY	PARVATI	GEFORCE	SILVERY	SLASHED	GBIT/S
SWITZERLAND	GRACE	CAPCOM	YELLOWISH	RIPPED	AMPERES

Машинный перевод





Решается задача оптимизации $\min \sum_{i=1}^{n} \| \mathbf{W}\mathbf{x}_i - \mathbf{z}_i \|^2$,

где $\{\mathbf{x}_i, \mathbf{z}_i\}_{i=1}^{n}$ — векторные репрезентации слов на разных языках.

Введение

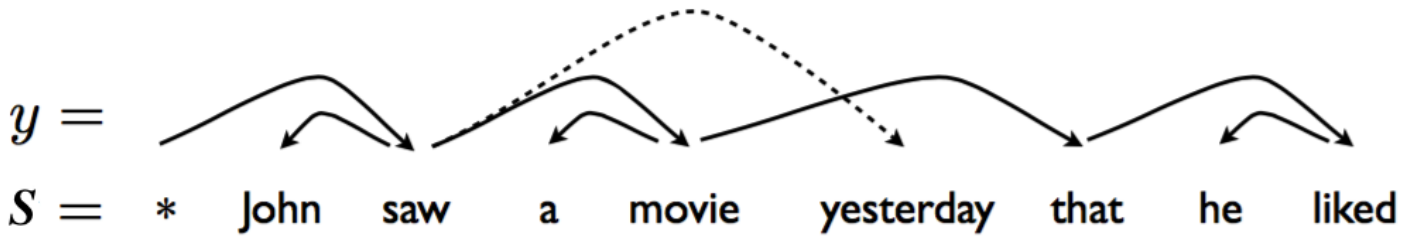
Синтаксический анализ

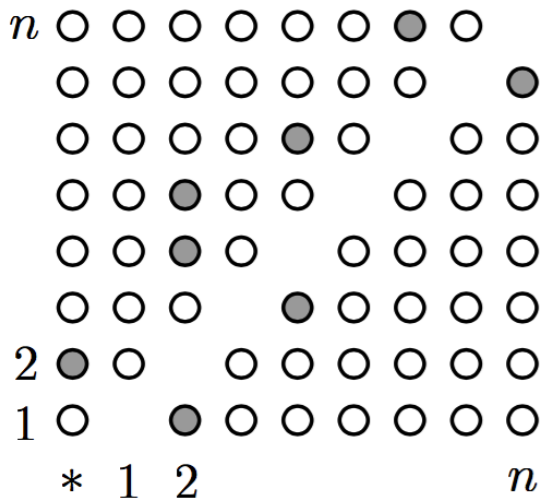
Процесс сопоставления линейной последовательности лексем (слов, токенов) естественного или формального языка с его формальной грамматикой. Результат — синтаксическое дерево разбора.

Синтаксическое дерево разбора

Направленный ациклический граф, представляющий синтаксическую структуру предложения в соответствии с некоторой контекстно-свободной грамматикой. Дерево состоит из листьев — слов предложения, внутренних узлов (узлов ветвления) — словосочетаний и терминальной вершины.

Постановка задачи





- $y(i,j) = \begin{cases} 1, & \text{если } i \to j; \\ 0, & \text{иначе.} \end{cases}$
- $\mathcal{Y}$ — набор направленных остовных деревьев,
- $f : \mathcal{Y} \to \mathbf{R}$.

$$y^* = \arg\max_{y \in \mathcal{Y}} f(y) = \arg\max_{y \in \mathcal{Y}} \sum_{(i,j)} y(i,j)\Theta(i,j).$$

Автокодировщик - определение

Автокодировщик — суперпозиция блоков

$$\boldsymbol{\mu} = \boldsymbol{\varphi}(\mathbf{g}(\mathbf{x})),$$

где $\mathbf{g} = \mathbf{f}(\mathbf{W}_e\mathbf{x} + \mathbf{b}_e)$ — кодирующий блок или encoder,

$\boldsymbol{\varphi}(\mathbf{g}(\mathbf{x})) = \mathbf{f}(\mathbf{W}_d\mathbf{g}(\mathbf{x}) + \mathbf{b}_d)$ — декодирующий блок или decoder,

$\mathbf{f}$ — нелинейная функция активации (tahn, σ).

Вектор параметров: $\boldsymbol{\Theta} = (\mathbf{W}_e, \mathbf{W}_d, \mathbf{b}_e, \mathbf{b}_d), \mathbf{W}_e = \mathbf{W}_d^{\mathsf{T}}$.

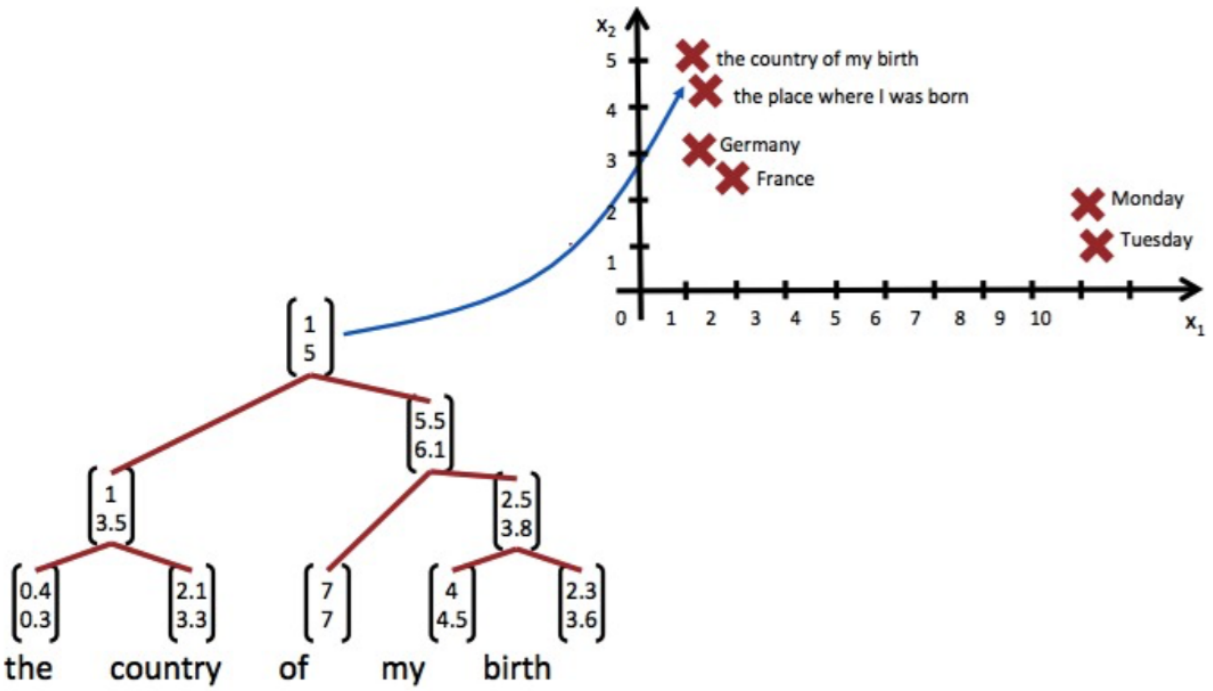
Функция ошибки: $\mathbf{S}(\boldsymbol{\Theta}, \mathbf{x}) = \| \mathbf{f}(\mathbf{x}|\boldsymbol{\Theta}) - \mathbf{x} \|_2^2$

Векторные репрезентации для предложений

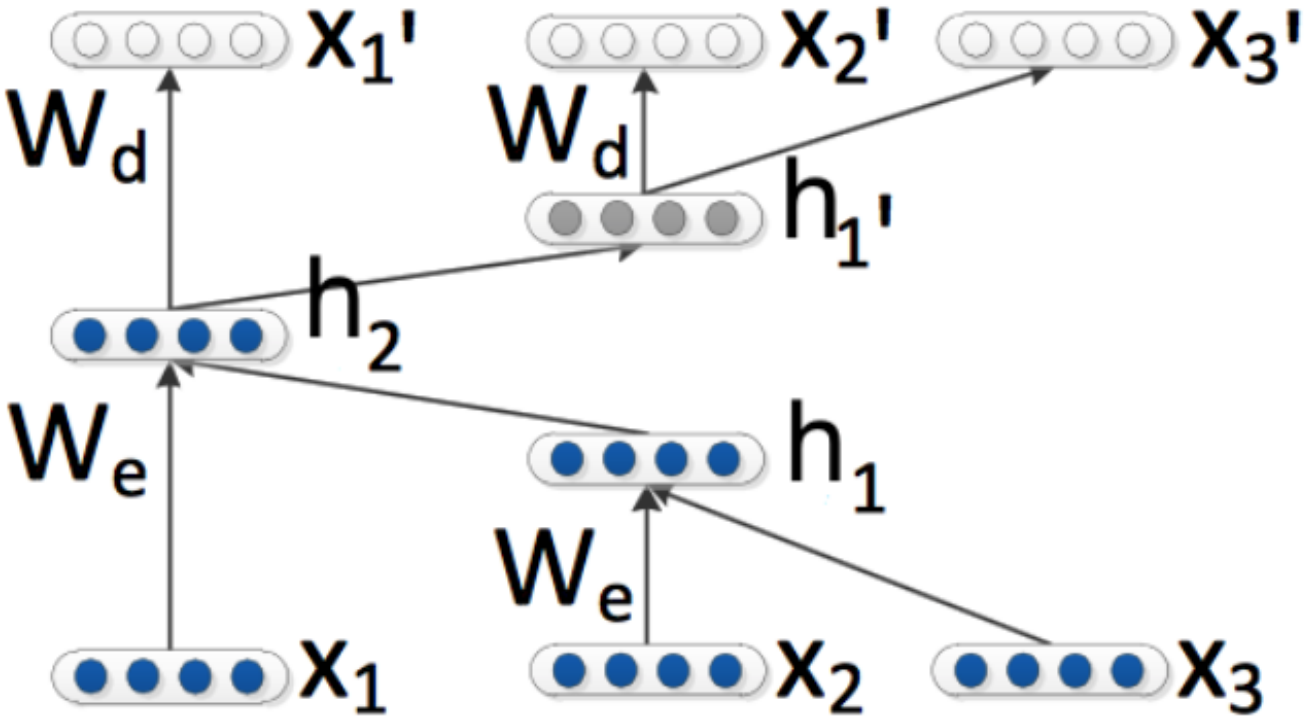
Что есть:

- Векторные репрезентации слов из Word2vec.
- Синтаксический парсер.
- Автокодировщик.

Хотелось бы получить:



Рекурсивный автокодировщик



Развертывающийся рекурсивный автокодировщик, Socher et al., 2011

Обозначим $\mathcal{H} = \{h\}_{\eta=1}^{(k-1)}$, k количество слов — множество внутренних узлов синтаксического дерева разбора.

$$\boldsymbol{\mu} = \boldsymbol{\varphi}(\mathbf{g}(\mathbf{p}_i^{(k-1)}, \mathbf{p}_j^{(k-1)})), (k-1) \ - \text{номер шага}$$

где

$$\mathbf{g}(\mathbf{p}_i, \mathbf{p}_j) = \mathbf{f}(\mathbf{W}_e[\mathbf{p}_i, \mathbf{p}_j] + \mathbf{b}_e) \text{ — encoder,}$$

$$\boldsymbol{\varphi}(\mathbf{g}(\mathbf{p}_i, \mathbf{p}_j)) = \mathbf{f}(\mathbf{W}_d\mathbf{g}(\mathbf{p}_i, \mathbf{p}_j)) + \mathbf{b}_d) \text{ — decoder,}$$

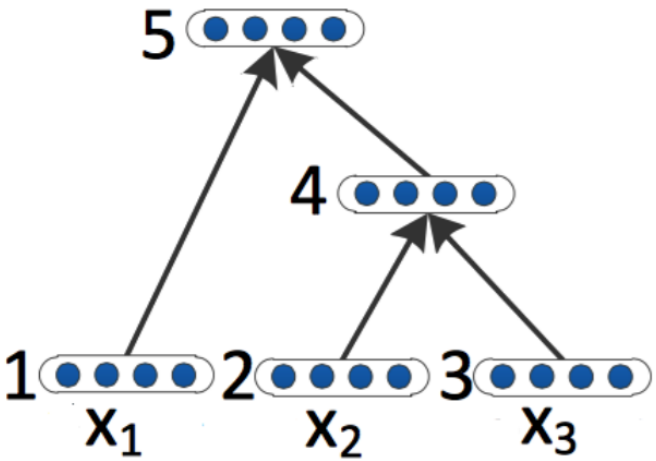
$$[\mathbf{p}_i, \mathbf{p}_j] \text{ — операция конкатенации,}$$

$$\mathbf{p}_i = \begin{cases} \mathbf{x}_i, & \text{если } \mathbf{p}_i \in X; \\ \mathbf{h}_i, & \text{если } \mathbf{p}_i \in \mathcal{H}. \end{cases},$$

Функция ошибки: $\mathbf{S}(\boldsymbol{\Theta}, [\mathbf{x^1}, \ldots, \mathbf{x^k}]) = \|\mathbf{z}([\mathbf{x^1}, \ldots, \mathbf{x^k}]|\boldsymbol{\Theta}) - [\mathbf{x^1}, \ldots, \mathbf{x^k}])$

Разберем на примере

Пусть имеется дерево разбора:



Узел $\mathbf{h}_1$ — родитель для $\mathbf{x}_2$, $\mathbf{x}_3$:

$$\mathbf{h}_1 \to \mathbf{x}_2\mathbf{x}_3.$$

Узел $\mathbf{h}_2$ — родителем для $\mathbf{x}_1$ и $\mathbf{h}_1$:

$$\mathbf{h}_2 \to \mathbf{x}_1\mathbf{h}_1.$$

Разберем на примере

Кодирующий блок:

$$\mathbf{h}_1 = \mathbf{f}(\mathbf{W}_e[\mathbf{x}_2, \mathbf{x}_3] + \mathbf{b}_e), \mathbf{h}_2 = \mathbf{f}(\mathbf{W}_e[\mathbf{x}_1, \mathbf{h}_1] + \mathbf{b}_e).$$

Декодирующий блок:

$$[\mathbf{x}'_1, \mathbf{h}'_1] = \mathbf{f}(\mathbf{W}_d\mathbf{h}_2 + \mathbf{b}_d), [\mathbf{x}'_2, \mathbf{x}'_3] = \mathbf{f}(\mathbf{W}_d\mathbf{h}_1 + \mathbf{b}_d).$$

Примеры

the U.S.	the former U.S.
suffering low morale	suffering heavy casualties
to watch hockey	to watch a video
enforcement of the economic	enforcement of the national
embargo	embargo

СПАСИБО ЗА ВНИМАНИЕ!