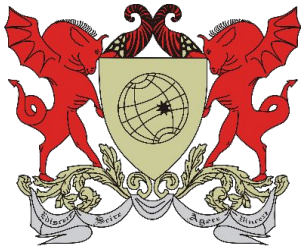


Word Embedding

Alcione de Paiva Oliveira - DPI/UFV

Baseado em CS224N/Ling284 - Christopher Manning and Richard Socher

Word Embedding

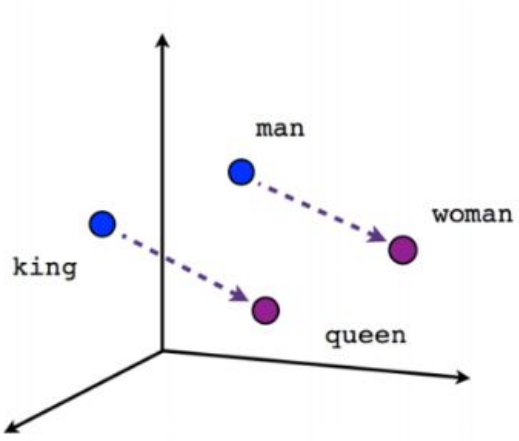
Introdução

- A Word Embedding é o nome de um conjunto de técnicas de modelagem de linguagem e de aprendizado de recursos no processamento de linguagem natural (PLN), em que palavras ou frases do vocabulário são mapeadas para vetores de números reais.

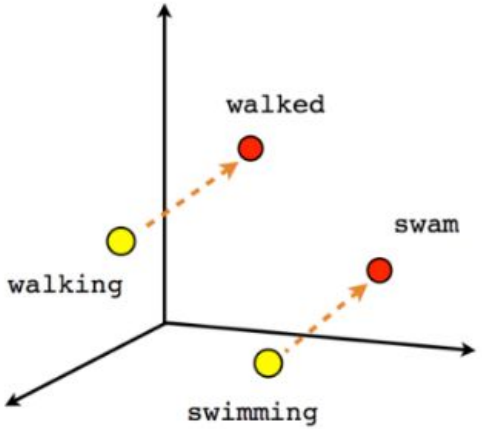
- Conceitualmente, envolve uma incorporação matemática de um espaço com uma dimensão por palavra para um espaço vetorial contínuo com uma dimensão muito maior.

Word Embedding

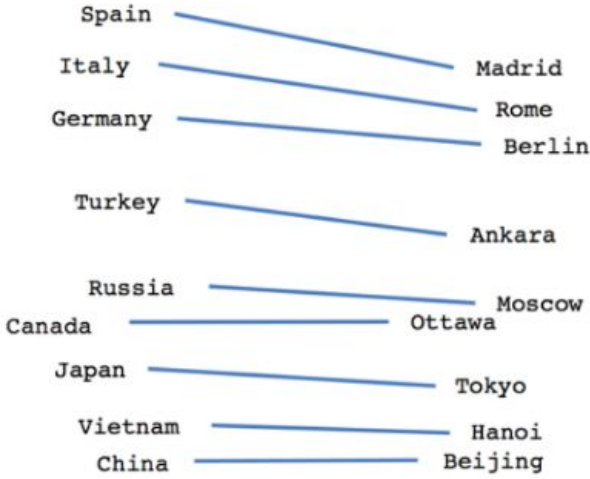
Introdução



Male-Female



Verb tense



Country-Capital

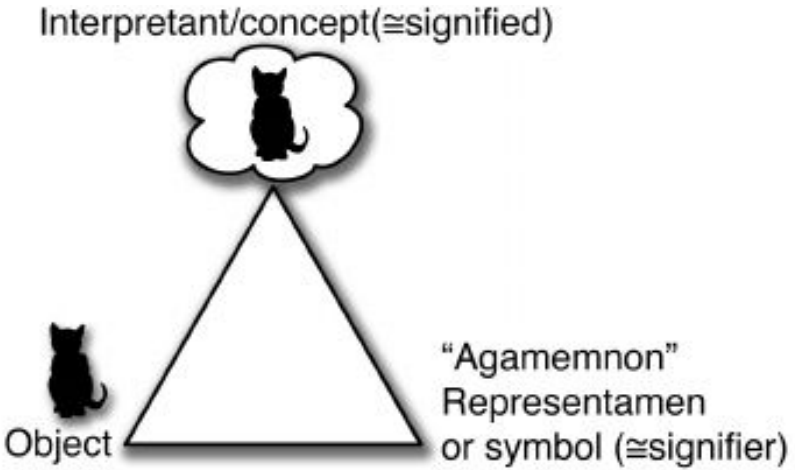
vetor[Queen] = vetor[king] - vetor[Man] + vetor[Woman]

Word Embedding

Introdução

- Definição de **significado**:
  - A ideia que é representada por uma palavra ou frase (Webster)
  - forma representativa e mental que se relaciona com a forma linguística; o que o signo quer significar; a parte do signo linguístico definida pelo conceito.

- Definição comum na linguística:



Word Embedding

Introdução

Como podemos ter um significado utilizável em um computador?

Resposta comum: use um recurso taxonômico, como o Wordnet que possui relações hierárquicas e de sinonímia.

```
from nltk.corpus import wordnet as wn
panda = wn.synset('panda.n.01')
hyper = lambda s: s.hypernyms()
list(panda.closure(hyper))
```

```
[Synset('procyonid.n.01'),
Synset('carnivore.n.01'),
Synset('placental.n.01'),
Synset('mammal.n.01'),
Synset('vertebrate.n.01'),
Synset('chordate.n.01'),
Synset('animal.n.01'),
Synset('organism.n.01'),
Synset('living_thing.n.01'),
Synset('whole.n.02'),
Synset('object.n.01'),
Synset('physical_entity.n.01'),
Synset('entity.n.01')]
```

(here, for *good*):

```
S: (adj) full, good
S: (adj) estimable, good, honorable, respectable
S: (adj) beneficial, good
S: (adj) good, just, upright
S: (adj) adept, expert, good, practiced,
proficient, skillful
S: (adj) dear, good, near
S: (adj) good, right, ripe
...
S: (adv) well, good
S: (adv) thoroughly, soundly, good
S: (n) good, goodness
S: (n) commodity, trade good, good
```

Word Embedding

Introdução

Problemas com a representação discreta.

- Não captura nuances, exemplo, **sinônimos**: rico, milionário, abastado?
- Faltam palavras e expressões novas: deletou, na moral, velho
- Subjetivo
- Requer trabalho humano
- Difícil computar similaridade

Word Embedding

Introdução

Problemas com a representação discreta.

Sistemas baseados em regras e estatísticos tratam às palavras como símbolos atômicos: **andar**, **amor**, **casa**

Em termos de espaço vetorial, isto implica em um vetor com um "1" e muitos zeros (500k para um vocabulário grande)

[0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0… 0 0 0]

Essa representação é chamada de "**one-hot**"

É uma representação localista

Word Embedding

Introdução

Problemas com a representação discreta.

Its problem, e.g., for web search

- If user searches for [Dell notebook battery size], we would like to match documents with "Dell laptop battery capacity"
- If user searches for [Seattle motel], we would like to match documents containing "Seattle hotel"

But

motel [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]^T
hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] = 0

Our query and document vectors are **orthogonal**

There is no natural notion of similarity in a set of one-hot vectors

Could deal with similarity separately;

instead we explore a direct approach where vectors encode it

Word Embedding

Introdução

É possível obter bastante informação representando o uma palavra em termos de seus "vizinhos".

"You shall know a word by the company it keeps"

(J. R. Firth 1957: 11)

One of the most successful ideas of modern statistical NLP

government debt problems turning into	banking	crises as has happened in
saying that Europe needs unified	banking	regulation to replace the hodgepodge

↖ These words will represent *banking* ↗

Word Embedding

Introdução

O significado das palavras é definido em termo de vetores.

É construído um vetor denso para cada palavra, feito de tal forma a ser útil em prever as outras palavras que aparecem em seu contexto.

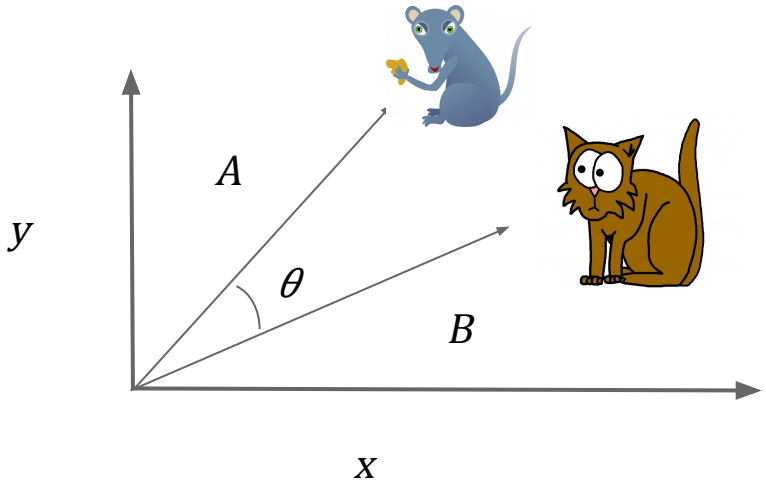
$$linguistics = \begin{pmatrix} 0.286 \\ 0.792 \\ -0.177 \\ -0.107 \\ 0.109 \\ -0.542 \\ 0.349 \\ 0.271 \end{pmatrix}$$

Word Embedding

Introdução

O significado das palavras é definido em termo de vetores.

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$



Word Embedding

Ideias Básicas

Definir um modelo com o objetivo de prever a coocorrência de palavras por meio de representação vetorial.

$\mathrm{P}(contexto|w_t)$=...

O aprendizado é feito por meio do exame de muitas posições t em um *corpus* grande.

Word Embedding

Ideias Básicas

Existem várias técnicas para a criação da representação vetorial (embedding)

Word2vec

MIKOLOV, Tomas et al. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301**.3781, 2013.

Glove

PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher D. Glove: Global vectors for word representation. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. 2014. p. 1532-1543.

Fasttext

BOJANOWSKI, Piotr et al. Enriching word vectors with subword information. **Transactions of the Association for Computational Linguistics**, v. 5, p. 135-146, 2017.

Word Embedding

WORD2VEC

Predição entre uma palavra e as palavras em seu contexto.

Dois Algoritmos:

Skip-grams (SG) - Prediz as palavras do contexto dado uma palavra:

w_{i-1} servidor w_{i+1} [w_{i-1} =o, w_{i+1}=caiu]

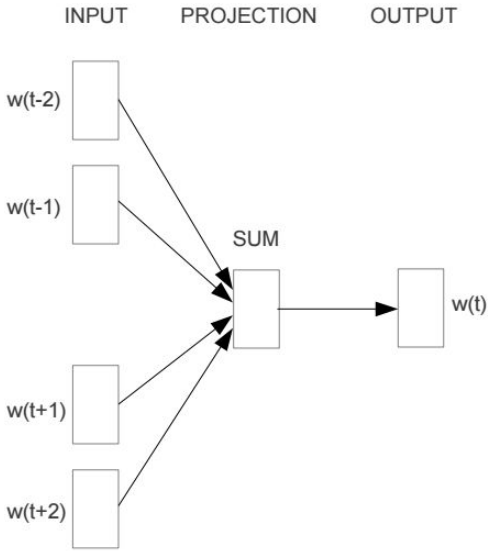
Continuous Bag of Words (CBOW) - Prediz a palavra dado o contexto:

Gosto de w_i pizza [w_i = comer]

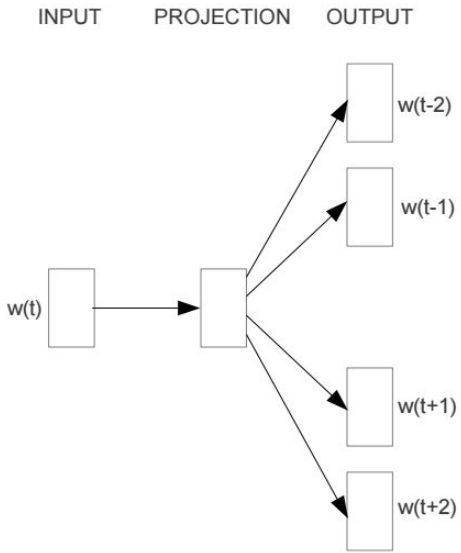
Word Embedding

WORD2VEC

Predição entre uma palavra e as palavras em seu contexto.



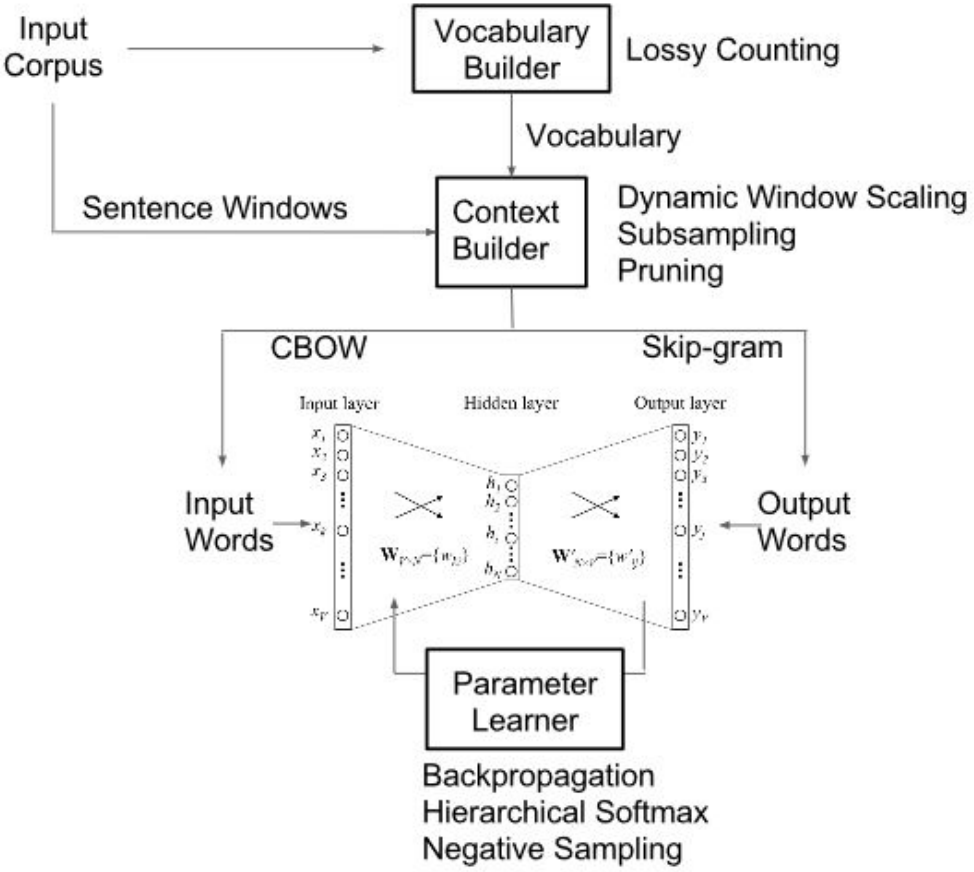
CBOW



Skip-gram

Word Embedding

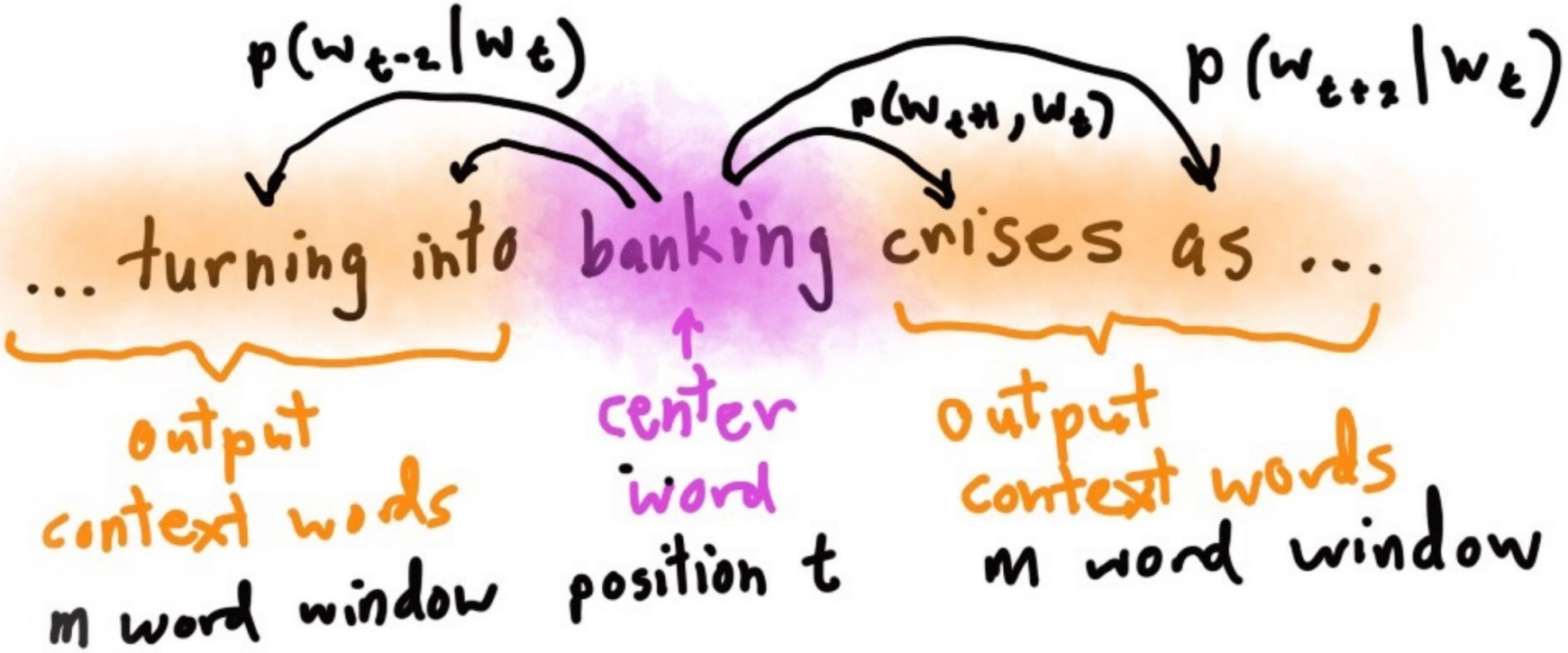
WORD2VEC



Word Embedding

WORD2VEC

Predição Skip-gram



Word Embedding

Skip-Gram

Para cada palavra t=1,…,T prediga as palavras na vizinhança com um "raio" ou janela de tamanho m.

Objective function: Maximize the probability of any context word given the current center word:

$$J'(\theta) = \prod_{t=1}^{T} \prod_{\substack{-m \leq j \leq m \\ j \neq 0}} p(w_{t+j} | w_t ; \theta)$$

Negative Log Likelihood

$$J(\theta) = -\frac{1}{T} \sum_{t=1}^{T} \sum_{\substack{-m \leq j \leq m \\ j \neq 0}} \log p(w_{t+j} | w_t)$$

Where θ represents all variables we will optimize

Word Embedding

Skip-Gram

Detalhes da função de perda

- Terminologia: função de perda = função de custo = função objetivo

- Função usualmente utilizada: Entropia cruzada

Word Embedding

Skip-Gram

Como fazer a predição?

Para $p(w_{t+j}|w_t)$ a formulação mais simples é SOFTMAX

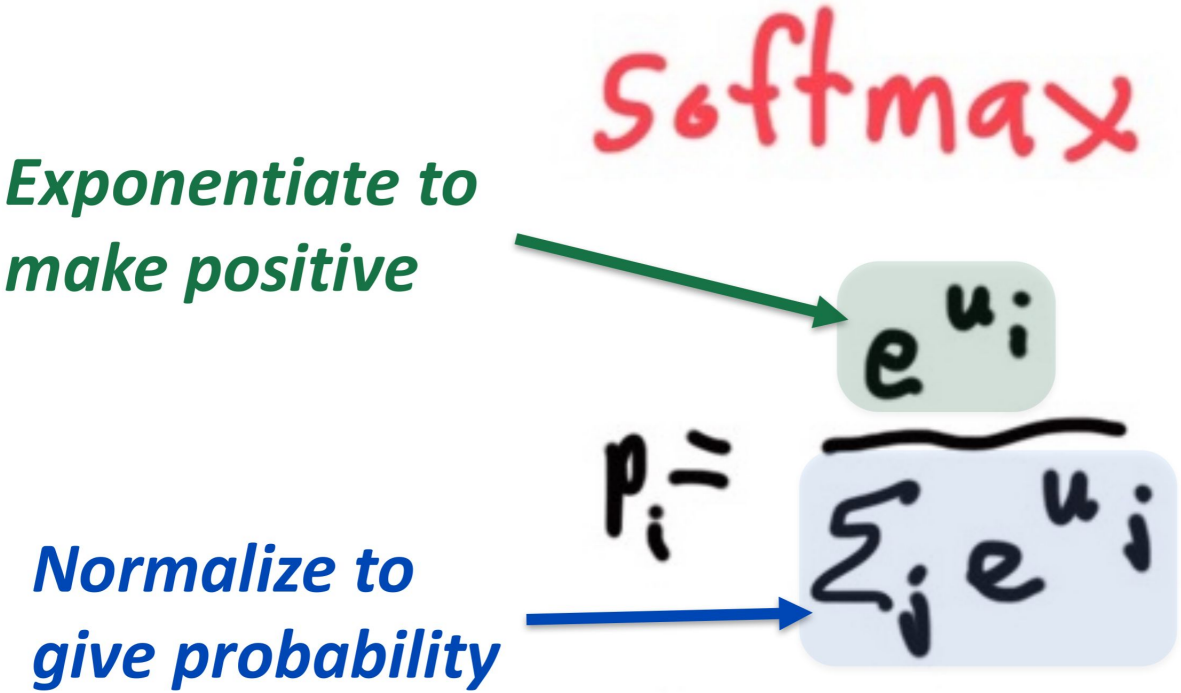
$$p(o|c) = \frac{\exp(u_o^T v_c)}{\sum_{w=1}^{V} \exp(u_w^T v_c)}$$

Onde o é a plavra de fora (outside), c é a palavra central

Word Embedding

Skip-Gram

Softmax: mapeamento padrão de $\mathbb{R}^V$ para uma distribuição de probabilidade



Word Embedding

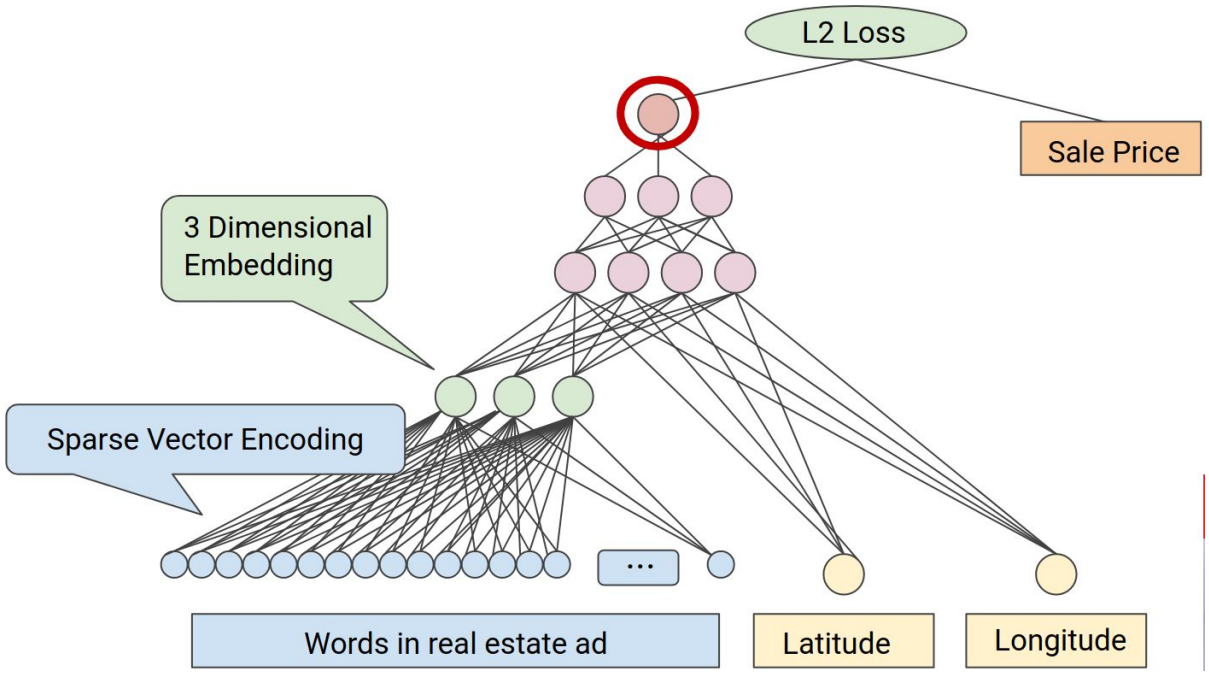
Skip-Gram

- Para treinar o modelo é necessário calcular o gradiente de todos os vetores.
- Toda palavra possui dois vetores: um como palavra central e uma como contexto.
- Torna o modelo mais simples.
- O conjunto de todos os parâmetros do modelo é definido em termos de um longo vetor θ.

Word Embedding

Aplicação

As camadas de embedding podem ser associadas a qualquer aplicação



Word Embedding

Glove

Word Embedding

Introdução

- Word2vec é subótimo, uma vez que não explora totalmente as informações estatísticas globais sobre co-ocorrências de palavras.
- O **GloVe** é um algoritmo de aprendizado não supervisionado para obter representações vetoriais para palavras.
- O treinamento é realizado em estatísticas agregadas de **co-ocorrência globais** de palavras e palavras de um corpus.
- As representações resultantes mostram subestruturas lineares interessantes no espaço vetorial.

Word Embedding

Introdução

2 opções de contagem: documento inteiro X Janela

- Co-ocorrência no documento inteiro permite trabalhar por tópicos (exemplo: esportes). Latent semantic analysis (**LSA**)
- A janela permite capturar informação sintática (POS) e semântica.

Word Embedding

Introdução

- Tamanho da janela: 5 a 10.
- Exemplo de corpus:
  - I like deep learning
  - I like NLP
  - I enjoy flying

Word Embedding

Introdução

Janela 1

- I like deep learning
- I like NLP
- I enjoy flying

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

Word Embedding

Glove

Resultados do Glove

- Palavras próximas a Frog

1. frogs
2. toad
3. litoria
4. leptodactylidae
5. rana
6. lizard
7. eleutherodactylus

litoria

leptodactylidae

rana

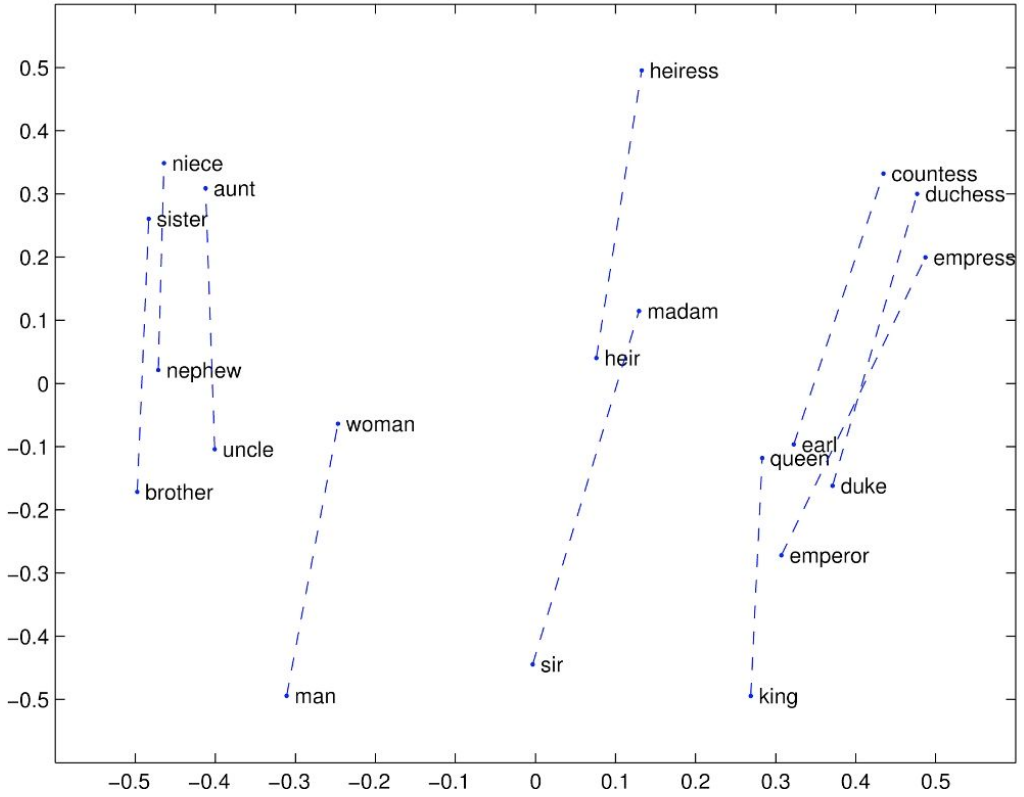
eleutherodactylus

Word Embedding

Glove

Resultados do Glove

- Visualizações



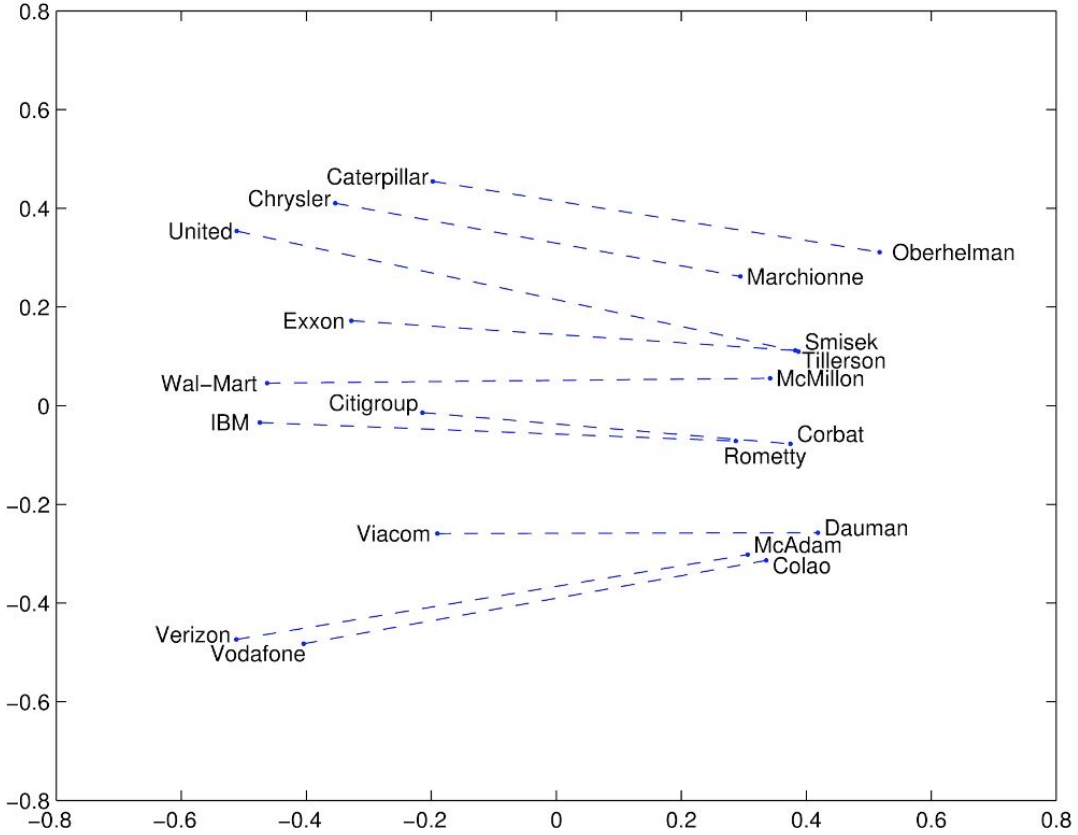
Word Embedding

Glove

Resultados do Glove

- Visualizações

Presidentes de empresas

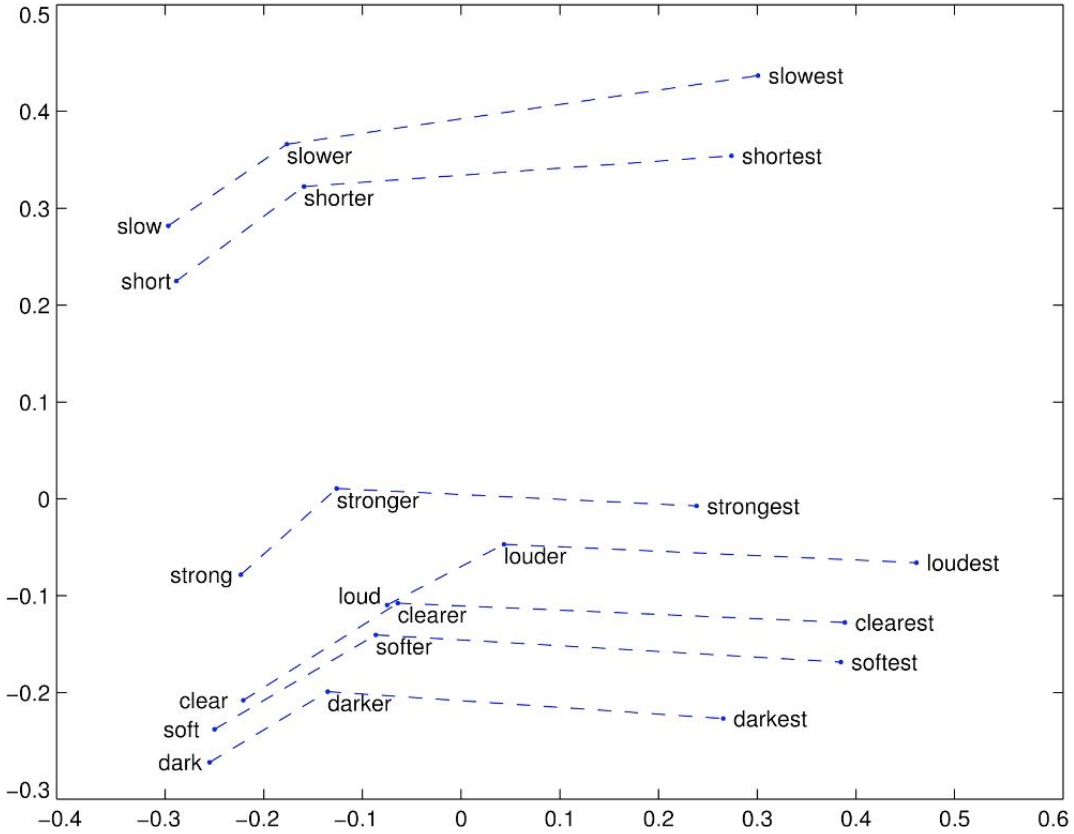


Word Embedding

Glove

Resultados do Glove

- Visualizações Superlativos



Word Embedding

Glove

Avaliação e hiperparâmetros

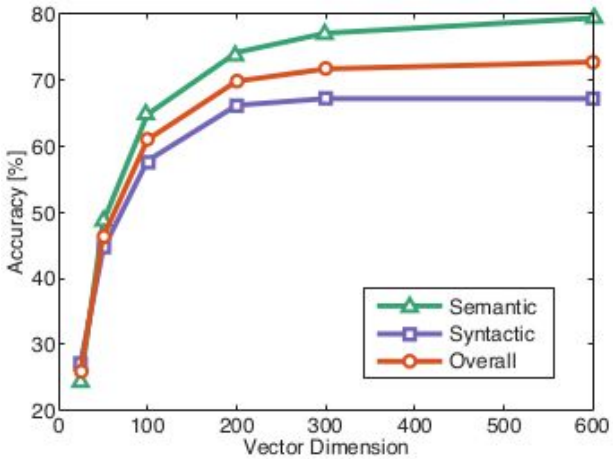
Model	Dim.	Size	Sem.	Syn.	Tot.
ivLBL	100	1.5B	55.9	50.1	53.2
HPCA	100	1.6B	4.2	16.4	10.8
GloVe	100	1.6B	67.5	54.3	60.3
SG	300	1B	61	61	61
CBOW	300	1.6B	16.1	52.6	36.1
vLBL	300	1.5B	54.2	64.8	60.0
ivLBL	300	1.5B	65.2	63.0	64.0
GloVe	300	1.6B	80.8	61.5	70.3
SVD	300	6B	6.3	8.1	7.3
SVD-S	300	6B	36.7	46.6	42.1
SVD-L	300	6B	56.6	63.0	60.1
CBOW†	300	6B	63.6	67.4	65.7
SG†	300	6B	73.0	66.0	69.1
GloVe	300	6B	77.4	67.0	71.7
CBOW	1000	6B	57.3	68.9	63.7
SG	1000	6B	66.1	65.1	65.6
SVD-L	300	42B	38.4	58.2	49.2
GloVe	300	42B	**81.9**	**69.3**	**75.0**

Word Embedding

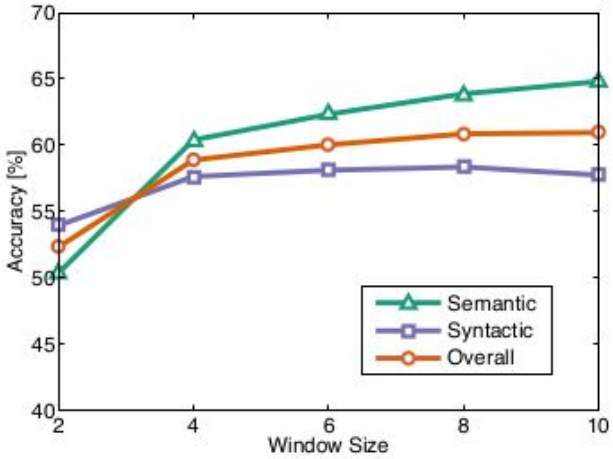
Glove

Avaliação e hiperparâmetros

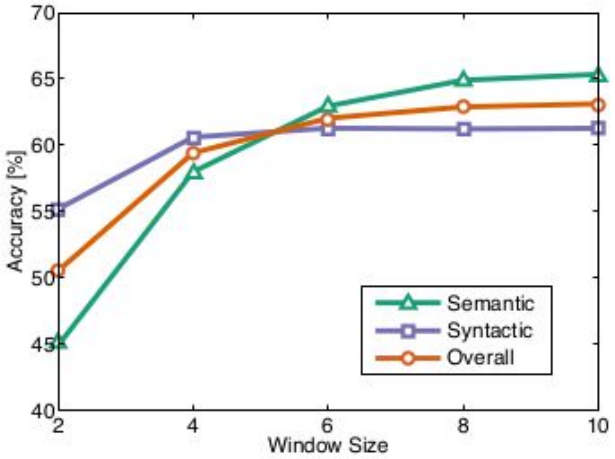
- Contextos assimétricos não são bons (somente as palavras à esquerda)
- Melhor dimensão em torno de 300
- Janela em torno de 8



(a) Symmetric context



(b) Symmetric context



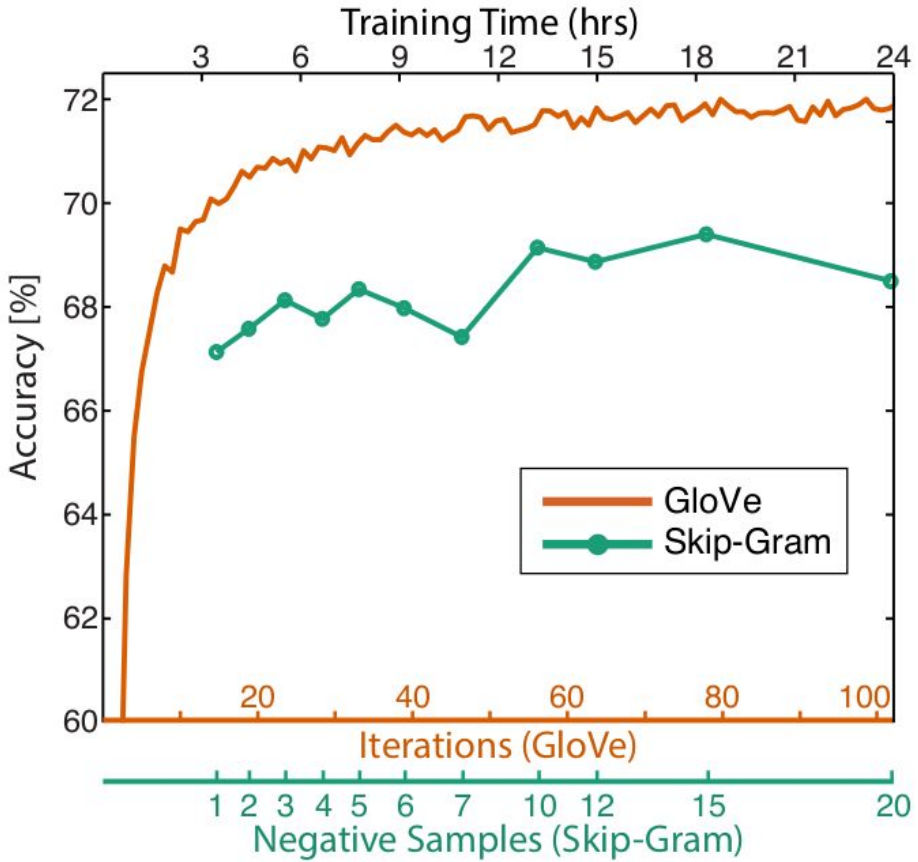
(c) Asymmetric context

Word Embedding

Glove

Avaliação e hiperparâmetros

- Maior tempo de treinamento ajuda

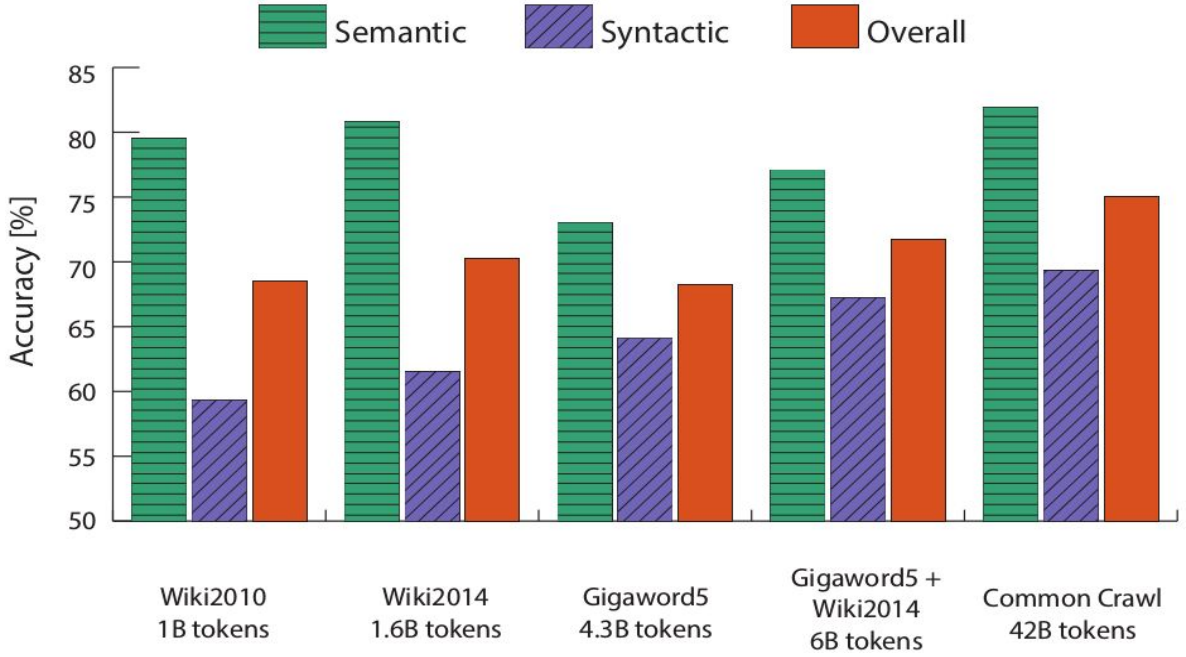


Word Embedding

Glove

Avaliação e hiperparâmetros

- Maior quantidade e variedade de dados melhor: wikipedia é melhor que notícias



Word Embedding

FastText

Word Embedding

FastText

- FastText (Facebook) é outro método de embedding que é uma extensão do modelo word2vec.
- Em vez de aprender vetores para palavras diretamente, fastText representa cada palavra como um n-grama de caracteres.
- Exemplo: palavra "Viçosa" com n = 3, a representação fastText dessa palavra é <Vi, Viç, içosa, çosa, osa, sa>, onde os colchetes angulares indicam o início e fim da palavra.

Word Embedding

FastText

- Isso ajuda a capturar o significado de palavras mais curtas e permite que os embeddings entendam sufixos e prefixos.
- fastText funciona bem com palavras raras. Portanto, mesmo que uma palavra não tenha sido vista durante o treinamento, ela pode ser dividida em n-gramas para obter seus embeddings.

Word Embedding

Links interessantes

http://web.stanford.edu/class/cs224n (curso de Stanford)