

Introduction to word embeddings

Agenda

- language modeling
- limitations of traditional n-gram language models
- Bengio et al. (2003)'s NNLM
- Google's word2vec (Mikolov et al. 2013)

Antoine Tixier, [DaSciM team](#), LIX
November 2015

Language model

- Goal: determine $P(s = w_1 \dots w_k)$ in some domain of interest

$$P(s) = \prod_{i=1}^k P(w_i \mid w_1 \dots w_{i-1})$$

e.g., $P(w_1 w_2 w_3) = P(w_1) P(w_2 \mid w_1) P(w_3 \mid w_1 w_2)$

- Traditional n-gram language model assumption:
“the probability of a word depends only on **context** of $n - 1$ previous words”

$$\Rightarrow \hat{P}(s) = \prod_{i=1}^k P(w_i \mid w_{i-n+1} \dots w_{i-1})$$

- Typical ML-smoothing learning process (e.g., Katz 1987):
 1. compute $\hat{P}(w_i \mid w_{i-n+1} \dots w_{i-1}) = \frac{\#w_{i-n+1} \dots w_{i-1} w_i}{\#w_{i-n+1} \dots w_{i-1}}$ on training corpus
 2. smooth to avoid zero probabilities

Traditional n-gram language model

Limitation 1): curse of dimensionality

- Example
 - train a 10-gram LM on a corpus of 100.000 unique words
 - space: 10-dimensional hypercube where each dimension has 100.000 slots
 - model training \leftrightarrow assigning a probability to each of the 100.000^{10} slots
 - **probability mass vanishes** \rightarrow more data is needed to fill the huge space
 - the more data, the more unique words! \rightarrow vicious circle
 - what about corpuses of 10^6 unique words?
- \rightarrow in practice, contexts are typically limited to size 2 (trigram model)
e.g., famous Katz (1987) smoothed trigram model
- \rightarrow such short context length is a limitation: a lot of information is not captured

Traditional n-gram language model

Limitation 2): word similarity ignorance

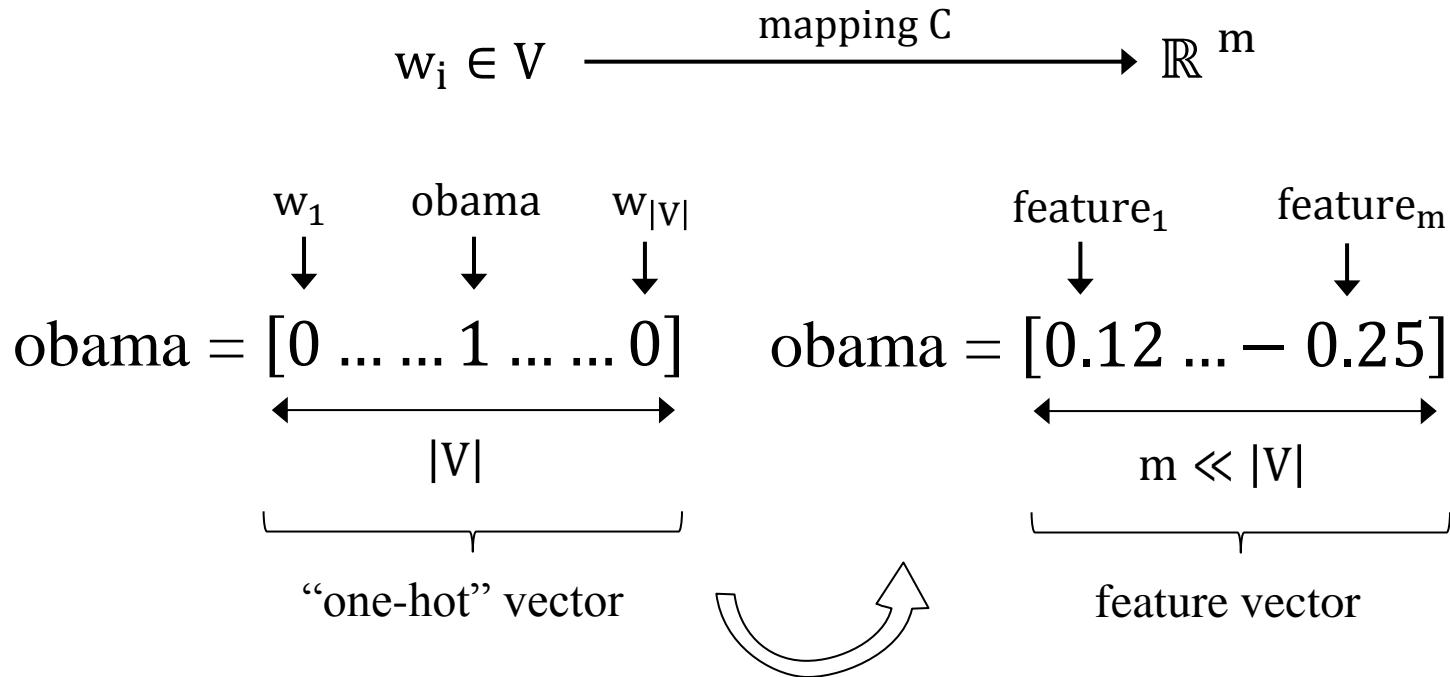
- We should assign similar probabilities to Obama speaks to the media in Illinois **and** the President addresses the press in Chicago
- This does not happen because of the “one-hot” vector space representation:

$$\begin{array}{l} \text{obama} = [0 \ 0 \ 0 \ 0 \ \dots \ 0 \ 1 \ 0 \ 0] \\ \text{president} = [0 \ 0 \ 0 \ 1 \ \dots \ 0 \ 0 \ 0 \ 0] \end{array} \left. \vphantom{\begin{array}{l} \text{obama} \\ \text{president} \end{array}} \right\} \overrightarrow{\text{obama}} \cdot \overrightarrow{\text{president}} = \vec{0}$$
$$\begin{array}{l} \text{speaks} = [0 \ 0 \ 1 \ 0 \ \dots \ 0 \ 0 \ 0 \ 0] \\ \text{addresses} = [0 \ 0 \ 0 \ 0 \ \dots \ 0 \ 0 \ 1 \ 0] \end{array} \left. \vphantom{\begin{array}{l} \text{speaks} \\ \text{addresses} \end{array}} \right\} \overrightarrow{\text{speaks}} \cdot \overrightarrow{\text{addresses}} = \vec{0}$$
$$\begin{array}{l} \text{illinois} = [1 \ 0 \ 0 \ 0 \ \dots \ 0 \ 0 \ 0 \ 0] \\ \text{chicago} = [0 \ 1 \ 0 \ 0 \ \dots \ 0 \ 0 \ 0 \ 0] \end{array} \left. \vphantom{\begin{array}{l} \text{illinois} \\ \text{chicago} \end{array}} \right\} \overrightarrow{\text{illinois}} \cdot \overrightarrow{\text{chicago}} = \vec{0}$$

- In each case, word pairs share no similarity
- This is obviously wrong
- We need to encode **word similarity** to be able to **generalize**

Word embeddings: distributed representation of words

- Each unique word is mapped to a point in a real continuous m -dimensional space
- Typically, $|V| > 10^6$, $100 < m < 500$

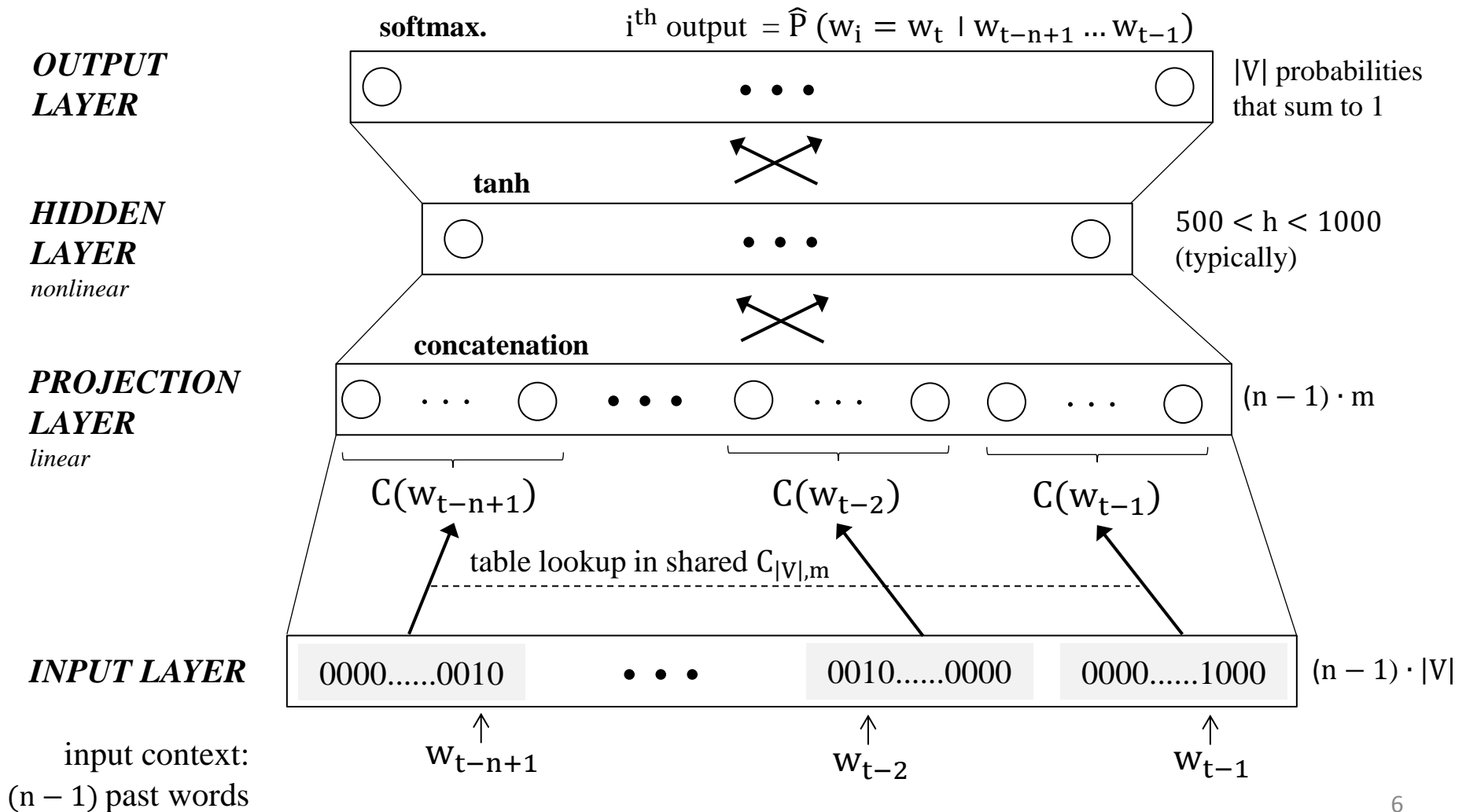


- Fighting the curse of dimensionality with:
 - **compression** (*dimensionality reduction*)
 - **smoothing** (*discrete to continuous*)
 - **densification** (*sparse to dense*)
- Similar words end up close to each other in the feature space

Neural Net Language Model (Bengio et al. 2003)

For each training sequence: input = (context, target) pair: $(w_{t-n+1} \dots w_{t-1}, w_t)$

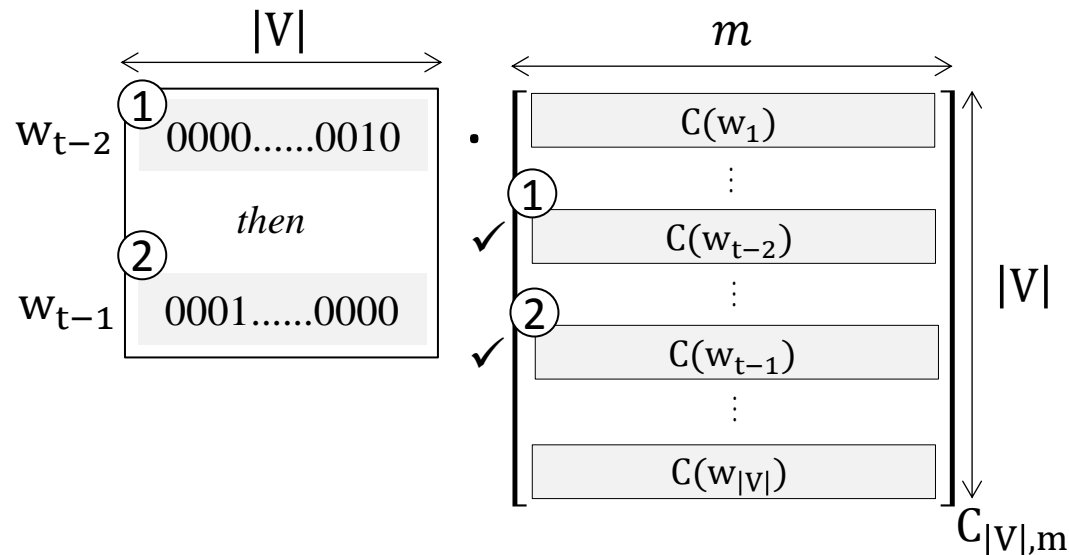
objective: minimize $E = -\log \hat{P}(w_t | w_{t-n+1} \dots w_{t-1})$



NNLM Projection layer

- Performs a simple table lookup in $C_{|V|,m}$: concatenate the rows of the shared mapping matrix $C_{|V|,m}$ corresponding to the context words

Example for a two-word context $w_{t-2}w_{t-1}$:



Concatenate ① and ② \rightarrow $C(w_{t-2})$ $C(w_{t-1})$

- $C_{|V|,m}$ is **critical**: it contains the weights that are tuned at each step. After training, it contains what we're interested in: the **word vectors**

NNLM hidden/output layers and training

- Softmax (log-linear classification model) is used to output positive numbers that sum to one (a multinomial probability distribution):

$$\text{for the } i^{\text{th}} \text{ unit in the output layer: } \hat{P}(w_i = w_t \mid w_{t-n+1} \dots w_{t-1}) = \frac{e^{y w_i}}{\sum_{i'=1}^{|V|} e^{y w_{i'}}$$

Where:

- $y = b + U \cdot \tanh(d + H \cdot x)$
 - \tanh : nonlinear squashing (link) function
 - x : concatenation $C(w)$ of the context weight vectors seen previously
 - b : output layer biases ($|V|$ elements)
 - d : hidden layer biases (h elements). Typically $500 < h < 1000$
 - U : $|V| * h$ matrix storing the *hidden-to-output* weights
 - H : $(h * (n - 1)m)$ matrix storing the *projection-to-hidden* weights
 - $\theta = (\mathbf{b}, \mathbf{d}, \mathbf{U}, \mathbf{H}, \mathbf{C})$
- Complexity per training sequence: $n * m + n * m * h + \mathbf{h} * |\mathbf{V}|$
computational bottleneck: **nonlinear hidden layer** ($h * |V|$ term)
 - **Training** is performed via stochastic gradient descent (learning rate ε):

$$\theta \leftarrow \theta + \varepsilon \cdot \frac{\partial E}{\partial \theta} = \theta + \varepsilon \cdot \frac{\partial \log \hat{P}(w_t \mid w_{t-n+1} \dots w_{t-1})}{\partial \theta}$$

(weights are initialized randomly, then updated via backpropagation)

NNLM facts

- - tested on Brown (1.2M words, $|V| \cong 16\text{K}$, 200K test set) and AP News (14M words, $|V| \cong 150\text{K}$ reduced to 18K, 1M test set) corpuses
- - Brown: $h = 100$, $n = 5$, $m = 30$
- - AP News: $h = 60$, $n = 6$, $m = 100$, **3 week** training using **40 cores**
- - 24% and 8% relative improvement (resp.) over traditional smoothed n-gram LMs in terms of test set perplexity: geometric average of $1/\hat{P}(w_t \mid w_{t-n+1} \dots w_{t-1})$
- Due to **complexity**, NNLM can't be applied to large data sets → poor performance on rare words
- Bengio et al. (2003) initially thought their main contribution was a more accurate LM. They left the interpretation and use of the word vectors as **future work**
- On the opposite, Mikolov et al. (2013) focus on the **word vectors**

Google's word2vec (Mikolov et al. 2013a)

- Key idea of word2vec: achieve better performance not by using a more complex model (i.e., with more layers), but by allowing a **simpler (shallower) model** to be trained on **much larger amounts of data**
- Two algorithms for learning words vectors:
 - **CBOW**: from context predict target (focus of what follows)
 - **Skip-gram**: from target predict context
- Compared to Bengio et al.'s (2003) NNLM:
 - no hidden layer (leads to 1000X speedup)
 - projection layer is shared (not just the weight matrix)
 - context: words from both **history & future**:
“You shall know a word by the company it keeps” (John R. Firth 1957:11):

...Pelé has called **Neymar** an excellent player...
...At the age of just 22 years, **Neymar** had scored 40 goals in 58 internationals...
...occasionally as an attacking midfielder, **Neymar** was called a true phenomenon...

← These words will represent **Neymar** →

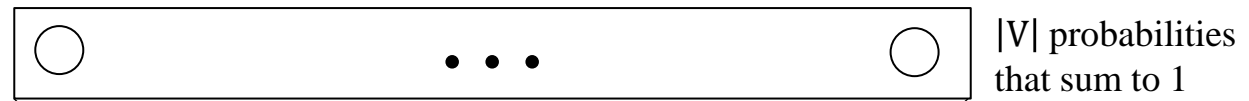
word2vec's Continuous Bag-of-Words (CBOW)

For each training sequence: input = (context, target) pair: $(w_{t-\frac{n}{2}} \dots w_{t-1} w_{t+1} \dots w_{t+\frac{n}{2}}, w_t)$

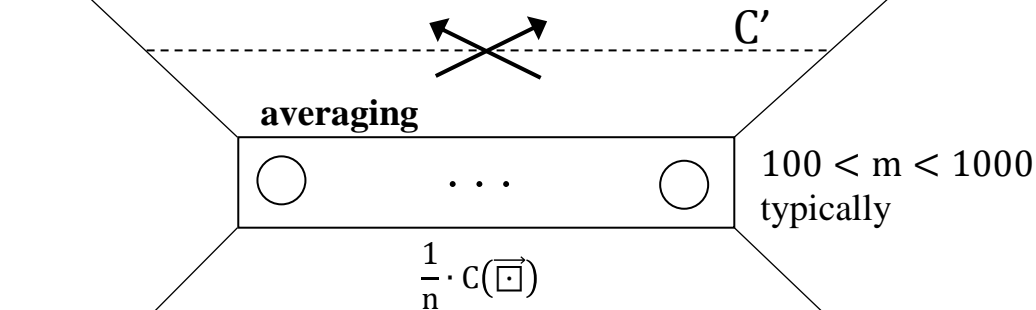
objective: minimize $E = -\log \hat{P}(w_t | w_{t-\frac{n}{2}} \dots w_{t-1} w_{t+1} \dots w_{t+\frac{n}{2}})$

hierarchical softmax. t^{th} output = $P(w_i = w_t | w_{t-\frac{n}{2}} \dots w_{t-1} w_{t+1} \dots w_{t+\frac{n}{2}})$

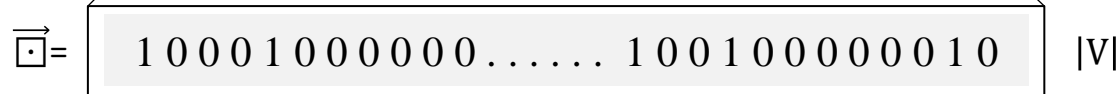
OUTPUT LAYER



PROJECTION LAYER
linear



INPUT LAYER



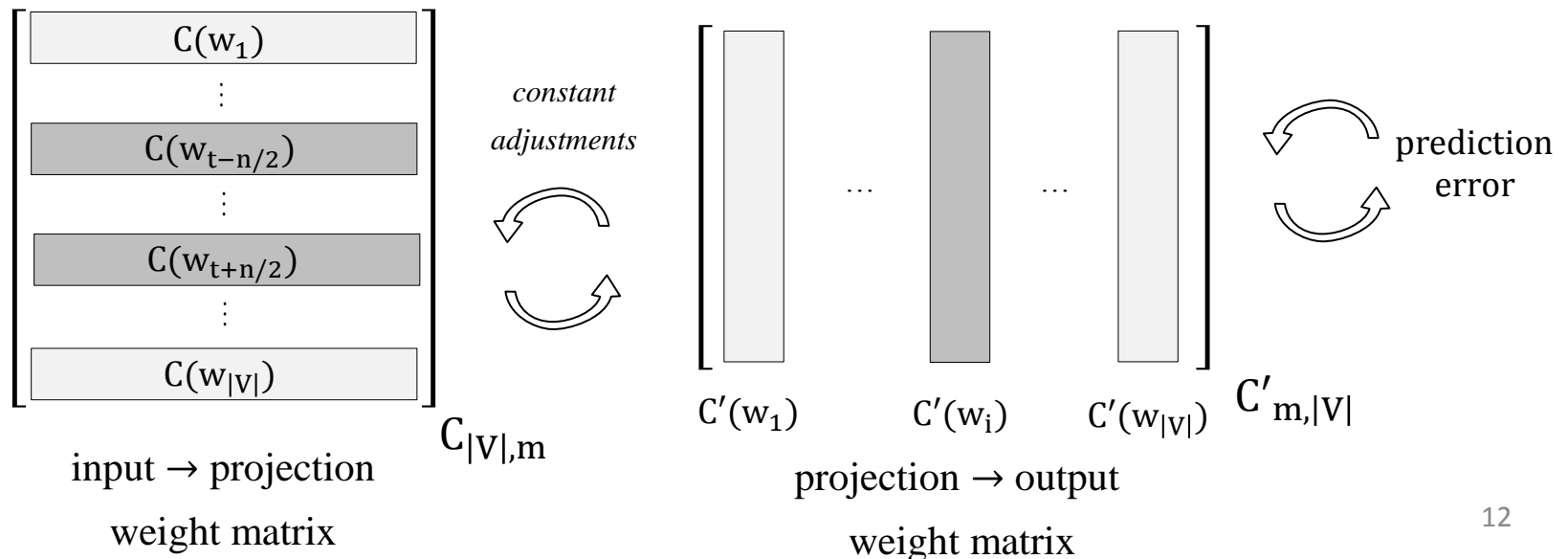
input context:

$n/2$ history words: $w_{t-\frac{n}{2}} \dots w_{t-1}$ $n/2$ future words: $w_{t+1} + \dots + w_{t+\frac{n}{2}}$

$n \cong 8$ typically

Weight updating intuition

- For each (context, target= w_t) pair, only the word vectors from matrix C corresponding to the context words are updated
 - Recall that we compute $P(w_i = w_t \mid \text{context}) \forall w_i \in V$. We compare this distribution to the true probability distribution (1 for w_t , 0 elsewhere)
 - If $P(w_i = w_t \mid \text{context})$ is **overestimated** (i.e., > 0 , happens in potentially $|V| - 1$ cases), some portion of $C'(w_i)$ is **subtracted** from the context word vectors in C , proportionally to the magnitude of the error
 - Reversely, if $P(w_i = w_t \mid \text{context})$ is **underestimated** (< 1 , happens in potentially 1 case), some portion of $C'(w_i)$ is **added** to the context word vectors in C
- at each step the words move away or get closer to each other in the feature space → clustering
 → analogy with a **spring force** layout. See online [demo](#) with Chrome



word2vec facts

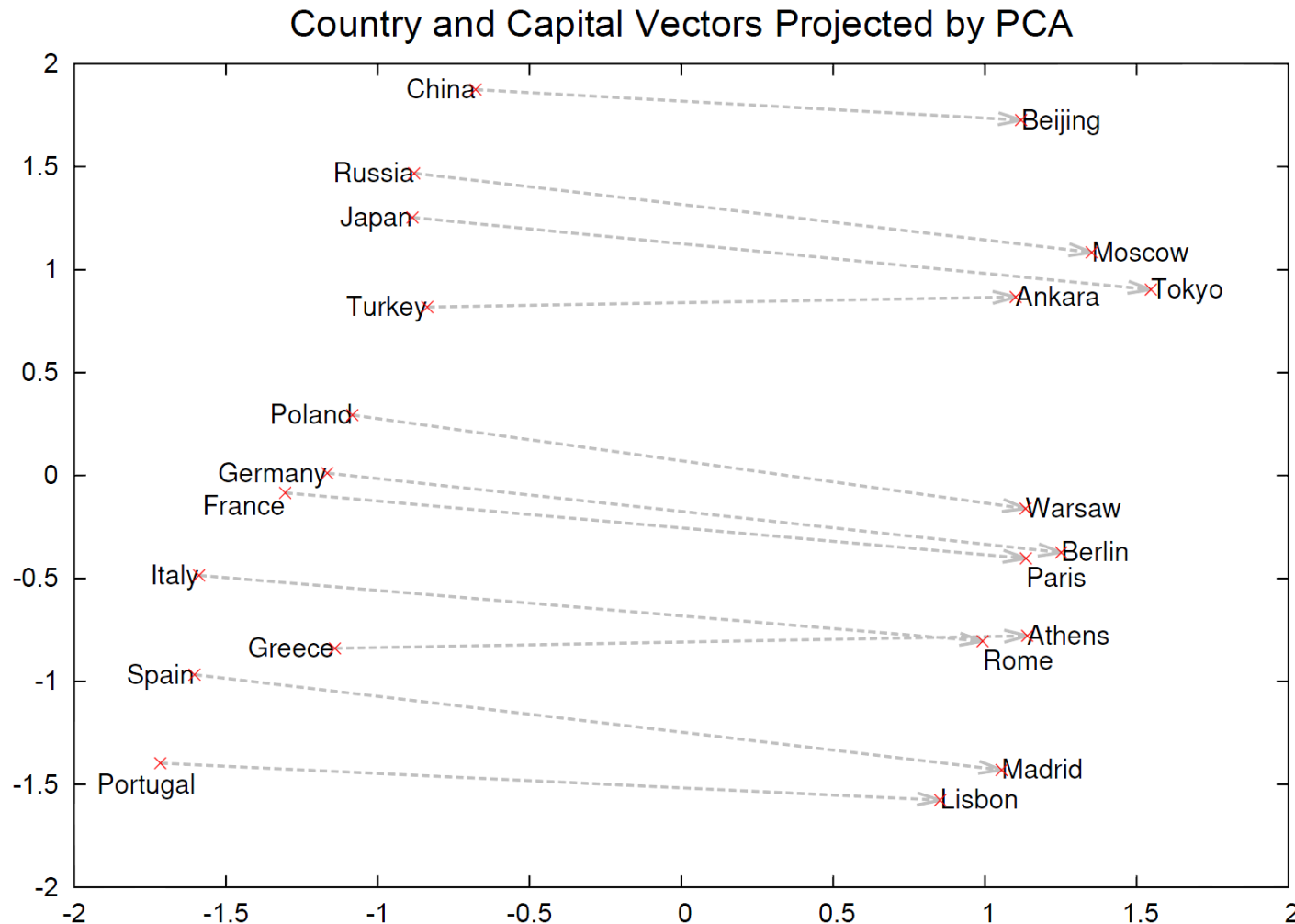
- Complexity is $n * m + m * \log|V|$ (Mikolov et al. 2013a)
- On Google news 6B words training corpus, with $|V| \sim 10^6$:
 - CBOW with $m = 1000$ took **2 days** to train on **140 cores**
 - Skip-gram with $m = 1000$ took **2.5 days** on **125 cores**
 - NNLM (Bengio et al. 2003) took **14 days** on **180 cores**, for $m = 100$ only!
(note that $m = 1000$ was not reasonably feasible on such a large training set)
- word2vec training speed $\cong 100K-5M$ words/s
- Quality of the word vectors:
 - \nearrow significantly with **amount of training data** and **dimension of the word vectors** (m),
with diminishing relative improvements
 - measured in terms of accuracy on 20K semantic and syntactic association tasks.
e.g., words in **bold** have to be returned:

Capital-Country	Past tense	Superlative	Male-Female	Opposite
Athens: Greece	walking: walked	easy: easiest	brother: sister	ethical: unethical

Adapted from Mikolov et al. (2013a)

- Best NNLM: 12.3% overall accuracy. Word2vec (with Skip-gram): 53.3%

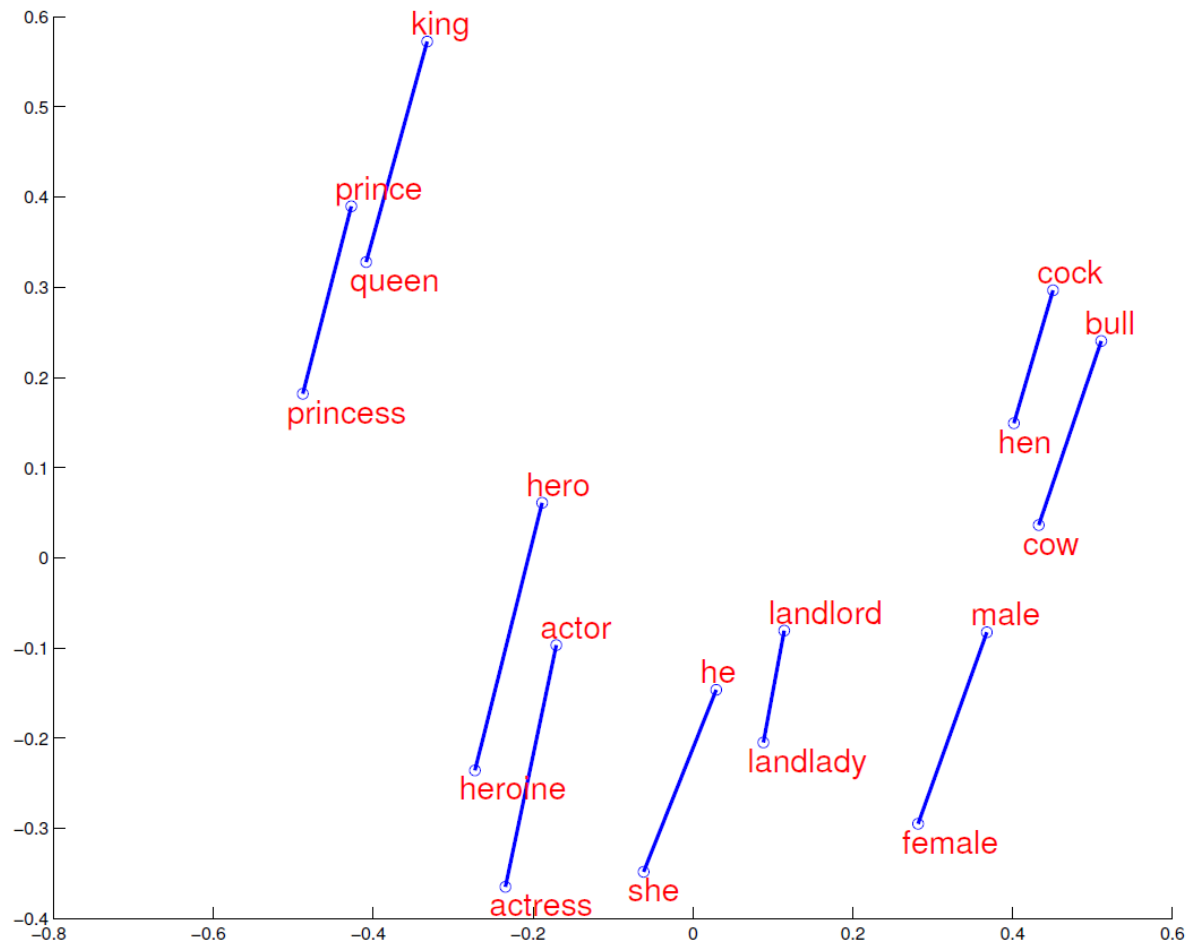
Remarkable properties of word2vec's word vectors



Mikolov et al. (2013b)

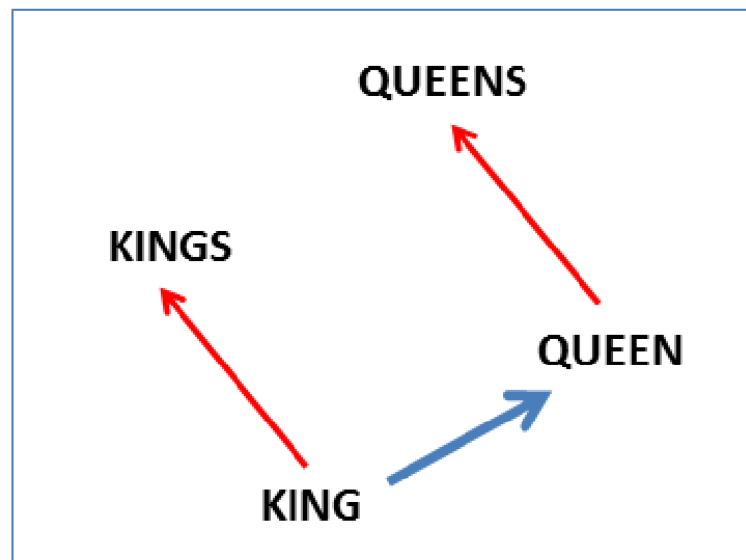
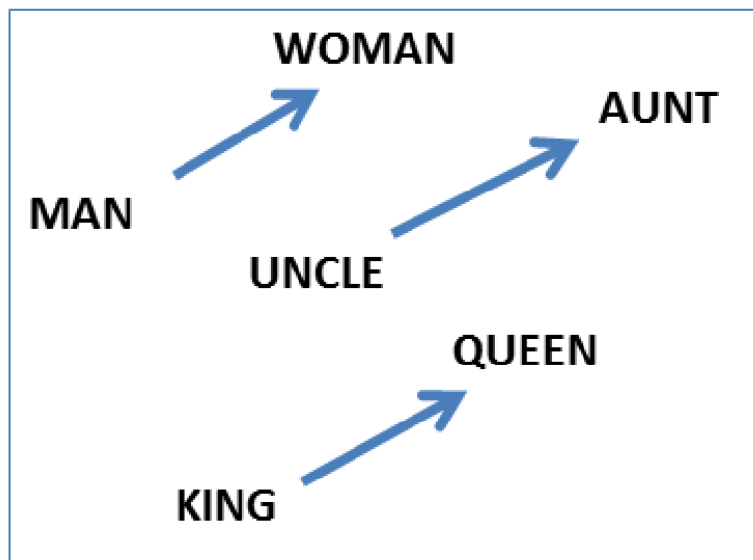
regularities between words are encoded in the difference vectors
e.g., there is a constant **country-capital** difference vector

Remarkable properties of word2vec's word vectors



constant **female-male** difference vector

Remarkable properties of word2vec's word vectors



constant **male-female** difference vector

constant **singular-plural** difference vector

- Vector operations are supported and make intuitive sense:

$$w_{king} - w_{man} + w_{woman} \cong w_{queen}$$

$$w_{einstein} - w_{scientist} + w_{painter} \cong w_{picasso}$$

$$w_{paris} - w_{france} + w_{italy} \cong w_{rome}$$

$$w_{his} - w_{he} + w_{she} \cong w_{her}$$

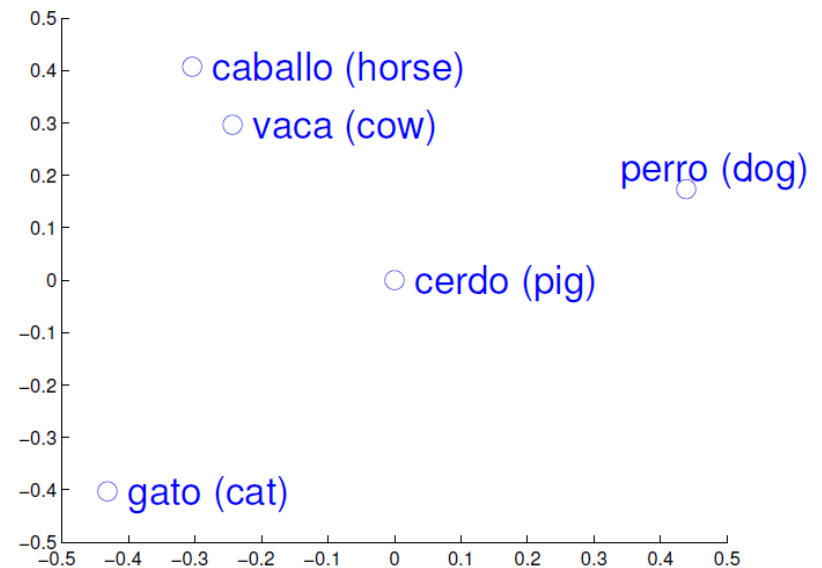
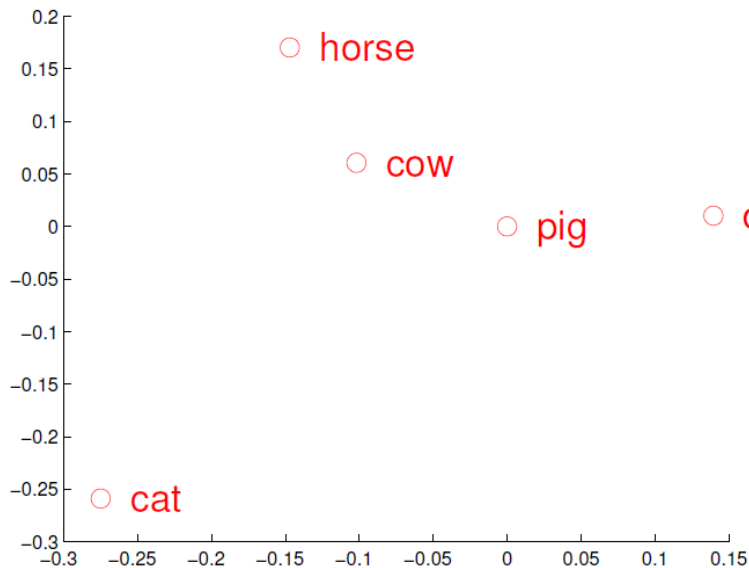
$$w_{windows} - w_{microsoft} + w_{google} \cong w_{android}$$

$$w_{cu} - w_{copper} + w_{gold} \cong w_{au}$$

- Online [demo](#) (scroll down to end of tutorial)

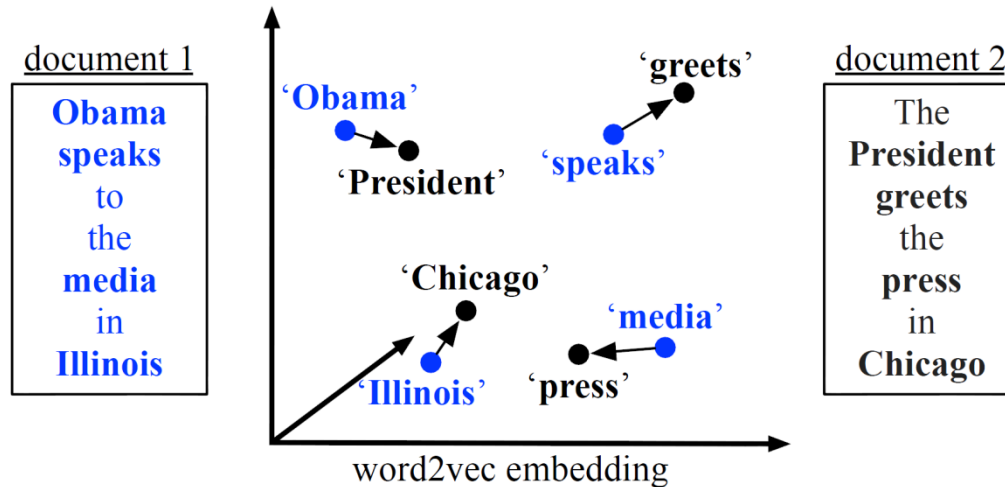
Applications

- High quality word vectors boost performance of all NLP tasks, including document classification, machine translation, information retrieval...
- Example for English to Spanish machine translation:

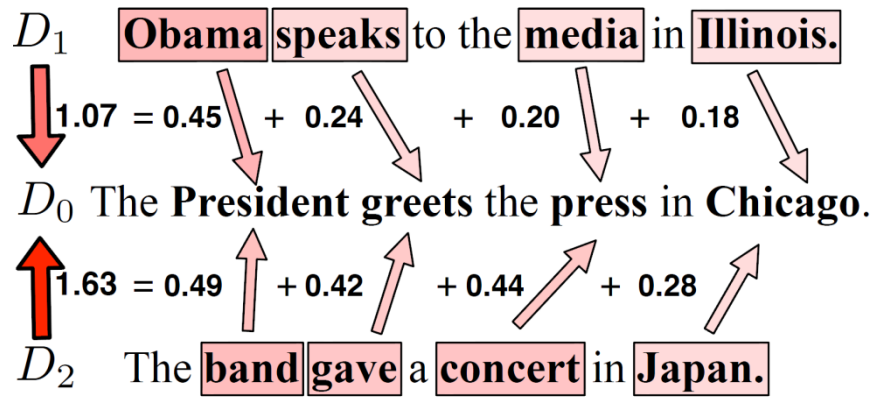


About 90% reported accuracy (Mikolov et al. 2013c)

Application to document classification



With the BOW representation D_1 and D_2 are at equal distance from D_0 . Word embeddings allow to capture the fact that D_1 is closer.



Resources

Papers:

[Chen, S. F., & Goodman, J. \(1999\). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13\(4\), 359-393.](#)

[Katz, S. M. \(1987\). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35\(3\), 400-401.](#)

[Bengio, Yoshua, et al. "A neural probabilistic language model." *The Journal of Machine Learning Research* 3 \(2003\): 1137-1155.](#)

[Mikolov, T., Chen, K., Corrado, G., & Dean, J. \(2013a\). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.](#)

[Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. \(2013b\). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* \(pp. 3111-3119\).](#)

[Mikolov, T., Le, Q. V., & Sutskever, I. \(2013c\). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.](#)

[Rong, X. \(2014\). word2vec Parameter Learning Explained. *arXiv preprint arXiv:1411.2738*.](#)

Google word2vec webpage (with link to C code):

<https://code.google.com/p/word2vec/>

Python implementation:

<https://radimrehurek.com/gensim/models/word2vec.html>

Kaggle tutorial on movie review classification with word2vec:

<https://www.kaggle.com/c/word2vec-nlp-tutorial/details/part-2-word-vectors>

Insightful blogpost: <http://colah.github.io/posts/2014-07-NLP-RNNs-Representations/>