Linguistic Regularities in Sparse and Explicit Word Representations

CONLL 2014

Omer Levy and Yoav Goldberg.

Astronomers have been able to capture and record the moment when a massive star begins to blow itself apart.

International Politics

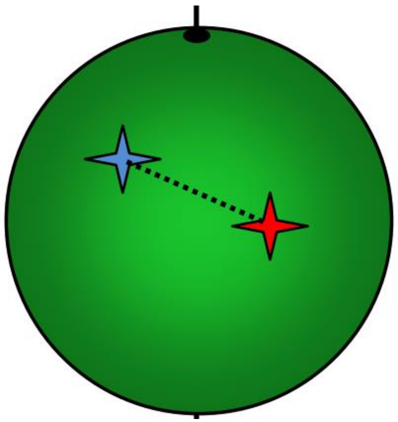
Space Science

A distance measure between words?

$$d(w_i, w_j)$$

- Documen
- Machine
- Informat
- Questior
- ....

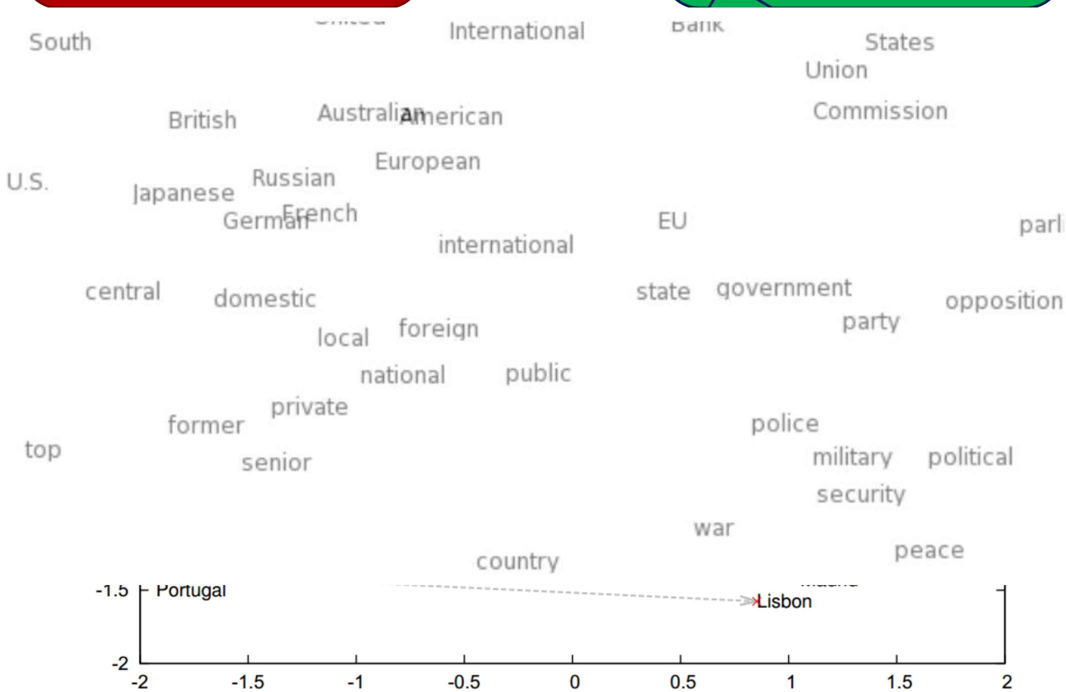
Vector representation?



Vector Representation

Explicit

Continuous ~ Embeddings ~ Neural

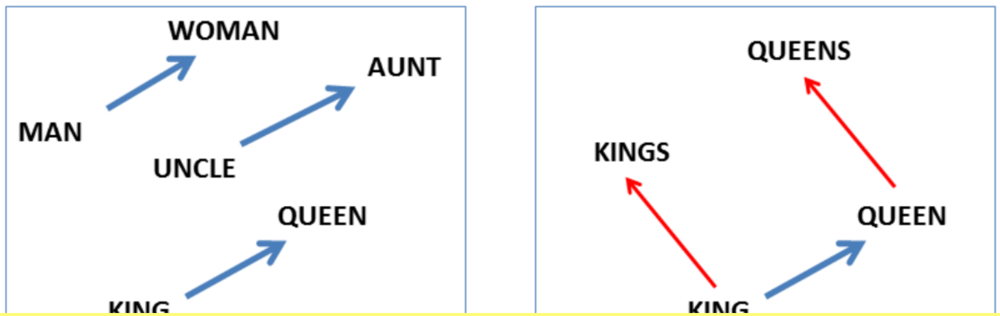


Russia is to *Moscow* as *Japan* is to *Tokyo*

Figures from (Mikolov et al.,NAACL, 2013) and (Turian et al., ACL, 2010)

Continuous vectors representation

- Interesting properties: directional similarity



- Analogy b

Older continuous vector representations:

- Latent Semantic Analysis
- Latent Dirichlet Allocation (Blei, Ng, Jordan, 2003)
- etc.

woman − man ≈ queen − king

A recent result (Levy and Golberg, NIPS, 2014) mathematically showed that the two representation are "almost" equivalent.

Figure from (Mikolov et al.,NAACL, 2013)

Goals of the paper

- Analogy:

$$a \text{ is to } a^* \text{ as } b \text{ is to } b^*$$

- With simple vector arithmetic:

$$a - a^* = b - b^*$$

- Given 3 words:

$$a \text{ is to } a^* \text{ as } b \text{ is to } ?$$

- Can we solve analogy problems with explicit representation?
- Compare the performance of the
  - Continuous (neural) representation
  - Explicit representation

Baroni et al. (ACL, 2014) showed that embeddings outperform explicit

Explicit representation

- Term-Context vector

	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

- $Apricot\colon \{Computer\colon 0, data\colon 0,\ pinch\colon 1, \ldots\} \in \mathbb{R}^{|Vocab| \approx 100{,}000}$
- Sparse!
- **Pointwise Mutual Information:**

$$\mathrm{PMI}(word, context) = \log_2 \frac{P(word, context)}{P(word)P(context)}$$

- **Positive PMI:**
  - Replace the negative PMI values with zero.

Continuous vectors representation

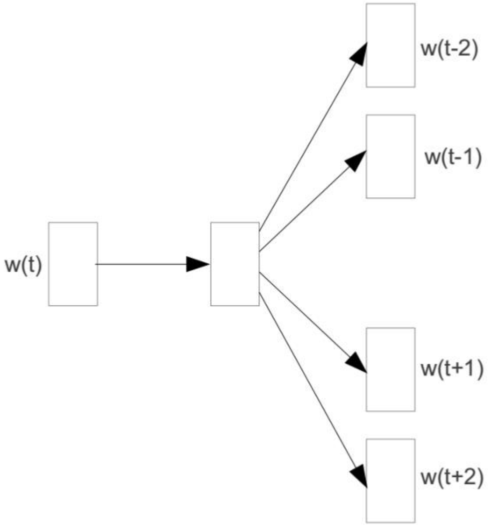
- Example:

$$Italy\colon\ (-7.35, 9.42, 0.88, \dots) \in \mathbb{R}^{100}$$

- Continuous values and dense
- How to create?
- Example: Skip-gram model (Mikolov et al., arXiv preprint arXiv:1301.3781)

$$L = \frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-c\le j\le c\\ j\ne 0}} \log p(w_{j+t}|w_t)$$

$$p(w_i|w_j) = \frac{\exp\,(v_i . v_j)}{\sum_k \exp\,(v_i . v_k)}$$



- Problem: The denominator is of size k.
- Hard to calculate the gradient

Formulating analogy objective

- Given 3 words:

$$a \text{ is to } a^* \text{ as } b \text{ is to } ?$$

- Cosine similarity:

$$\cos(a, b) = \frac{a.b}{\|a\|\|b\|}$$

- Prediction:

$$\arg\max_{b^*}(\cos(b^*, b - a + a^*))$$

- If using normalized vectors:

$$\arg\max_{b^*}(\cos(b^*, b) - \cos(b^*, a) + \cos(b^*, a^*))$$

- Alternative?
- Instead of **adding** similarities, **multiply** them!

$$\arg\max_{b^*}\left(\frac{\cos(b^*, b)\cos(b^*, a^*)}{\cos(b^*, a)}\right)$$

Corpora

- **MSR:** ~8000 syntactic analogies
- **Google:** ~19,000 syntactic and semantic analogies

a	b	a*	b*
Good	Better	Rough	?
better	good	rougher	?
good	best	rough	?
…	…	…	…

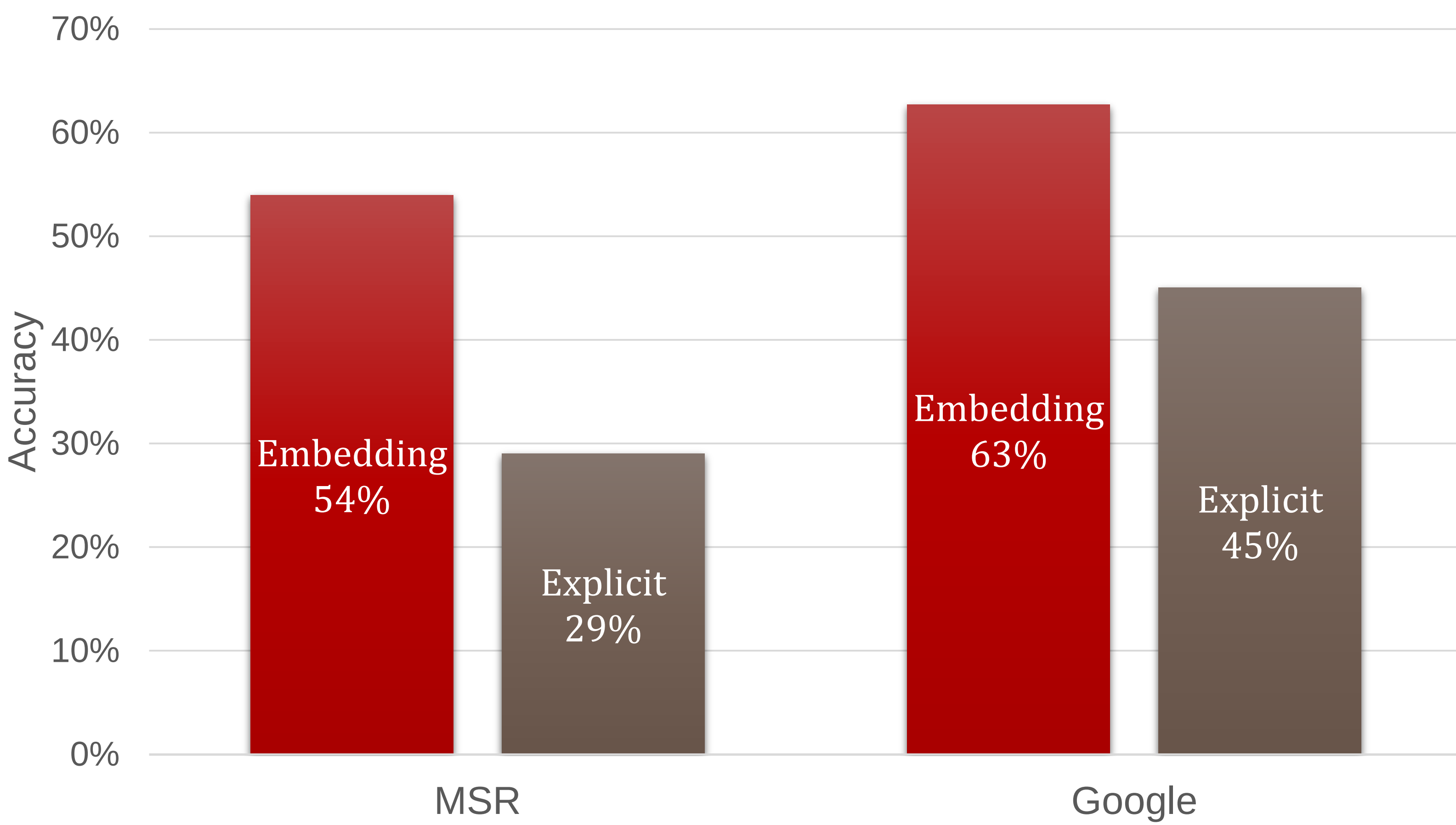
- Learn **different** representations from the same corpus:

WIKIPEDIA
The Free Encyclopedia

Very important!!

Recent controversies over inaccurate evaluations for “GloVe word-representation” (Pennington et al, EMNLP)

Embedding vs Explicit (Round 1)



Many analogies recovered by **explicit**, but many more by **embedding**.

Using multiplicative form

- Given 3 words:

a is to a^* as b is to ?

- Cosine similarity:

$$\cos(a, b) = \frac{a.b}{\|a\|\|b\|}$$

- Additive objective:

$$\arg\max_{b^*}(\cos(b^*, b) - \cos(b^*, a) + \cos(b^*, a^*))$$

- Multiplicative objective:

$$\arg\max_{b^*}\left(\frac{\cos(b^*, b)\cos(b^*, a^*)}{\cos(b^*, a)}\right)$$

A relatively weak justification

England − London + Baghdad = Iraq

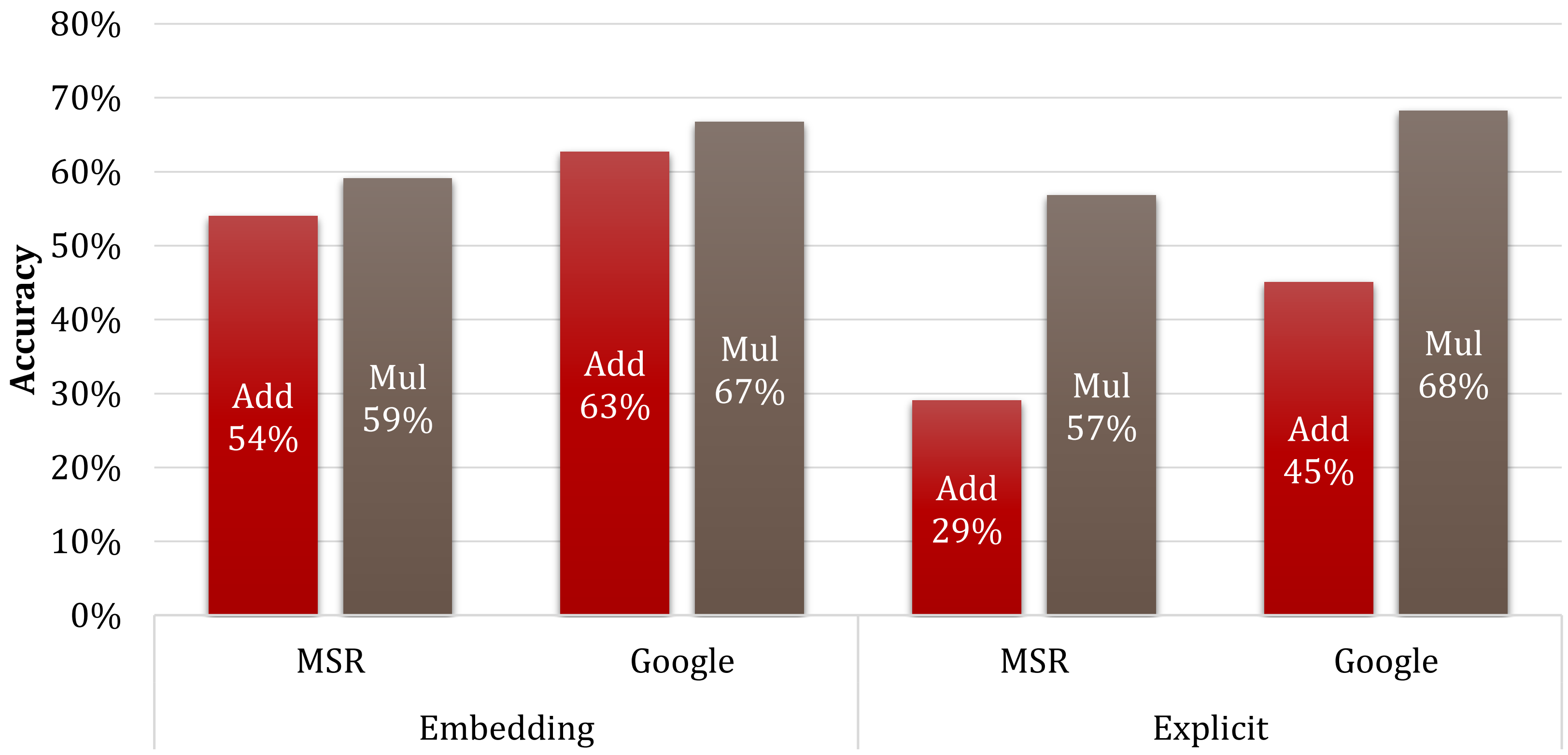
$$\cos(Iraq, England) - \cos(Iraq, London) + \cos(Iraq, Baghdad)$$

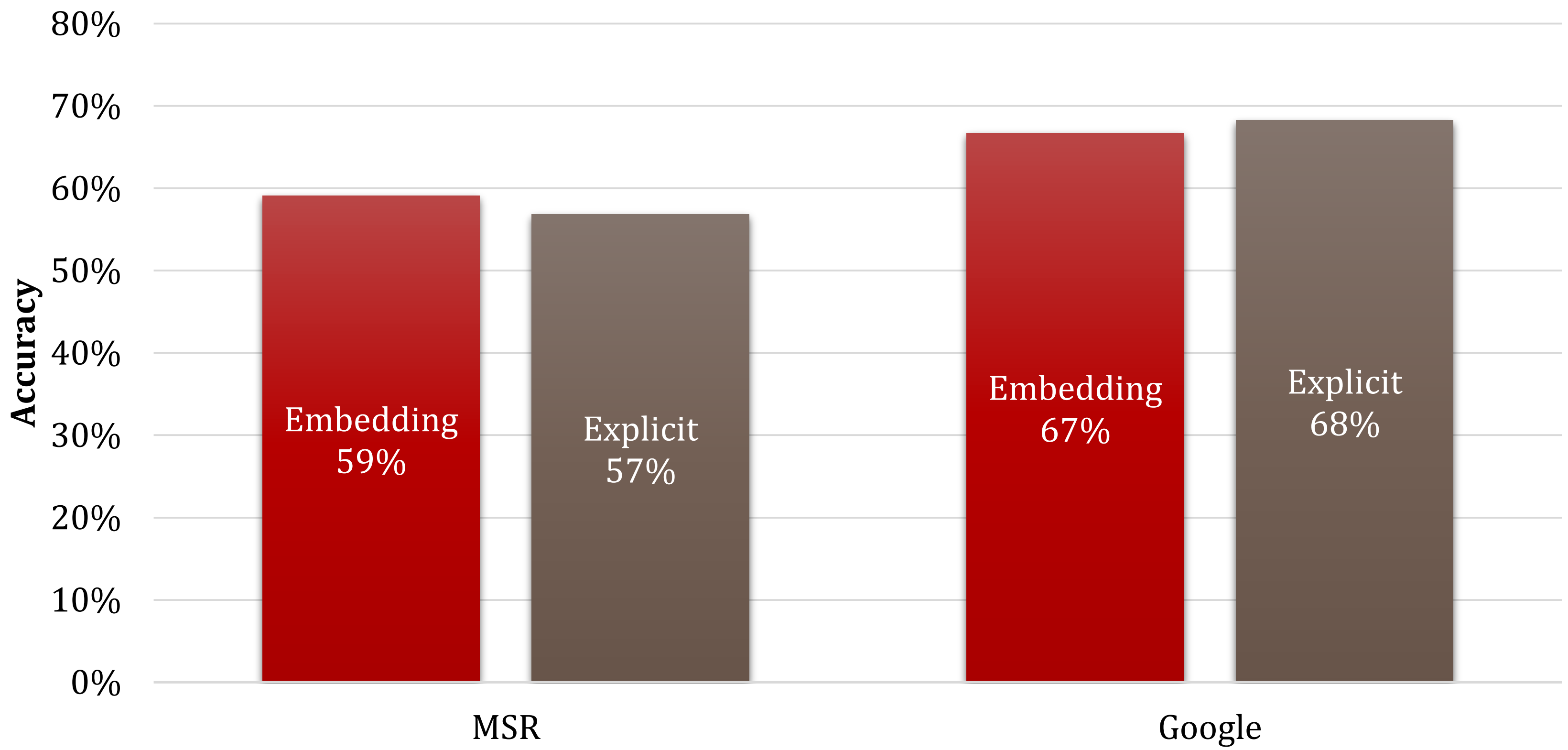
0.15 0.13 0.63

0.13 0.14 0.75

$$\cos(Mosul, England) - \cos(Mosul, London) + \cos(Mosul, Baghdad)$$

Embedding vs Explicit (Round 2)




80%
70%
60%
50%
40%
30%
20%
10%
0%
Accuracy
Embedding
59%
Explicit
57%
Embedding
67%
Explicit
68%
MSR
Google

Summary

- On **analogies**, continuous (neural) representation **is not magical**
- Analogies **are possible** in the explicit representation
- On analogies explicit representation can be as good as continuous (neural) representation.
- Objective (function) matters!

Agreement between representations

Objective	Both Correct	Both Wrong	Embedding Correct	Explicit Correct
MSR	43.97%	28.06%	15.12%	12.85%
Google	57.12%	22.17%	9.59%	11.12%

Comparison of objectives

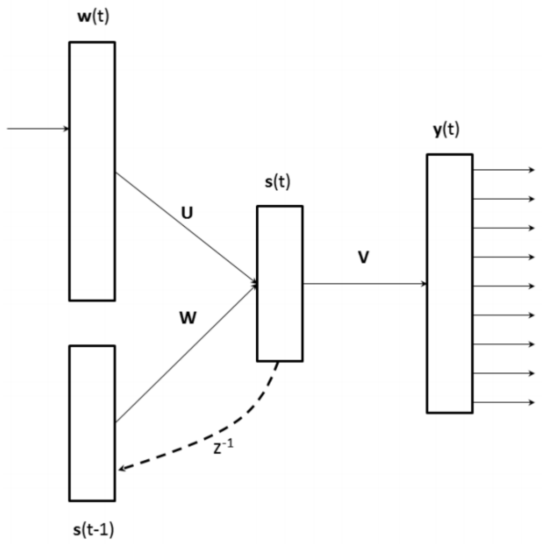
Objective	Representation	MSR	Google
Additive	Embedding	53.98%	62.70%
	Explicit	29.04%	45.05%
Multiplicative	Embedding	59.09%	66.72%
	Explicit	56.83%	68.24%

Recurrent Neural Network

- Recurrent Neural Networks (Mikolov et al.,NAACL, 2013)
- Input-output relations:

$$\begin{cases} y(t) = \boldsymbol{g}(\boldsymbol{V}s(t)) \\ s(t) = f(\boldsymbol{W}s(t-1) + \boldsymbol{U}w(t)) \end{cases}$$

- $y \in \mathbb{R}^{|Vocab|}$
- $w(t) \in \mathbb{R}^{|Vocab|}$
- $s(t) \in \mathbb{R}^{d}$
- $\boldsymbol{U} \in \mathbb{R}^{|Vocab| \times d}$
- $\boldsymbol{W} \in \mathbb{R}^{d \times d}$
- $\boldsymbol{V} \in \mathbb{R}^{d \times |Vocab|}$



Continuous vectors representation

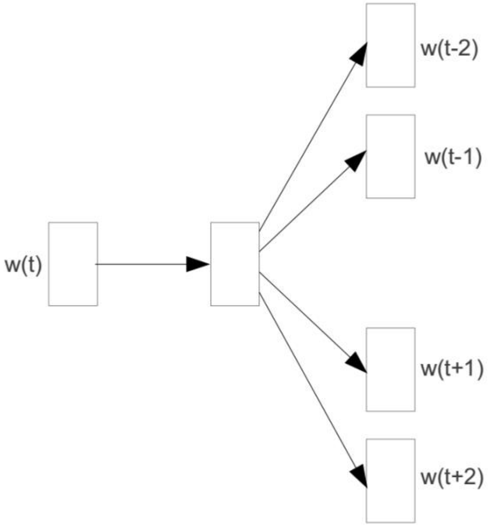
- Example:

$$Italy: \ (-7.35, 9.42, 0.88, \dots) \in \mathbb{R}^{100}$$

- Continuous values and dense
- How to create?
- Example: Skip-gram model (Mikolov et al., arXiv preprint arXiv:1301.3781)

$$L = \frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-c\le j\le c\\ j\neq 0}} \log p(w_{j+t}|w_t)$$

$$p(w_i|w_j) = \frac{\exp\,(v_i.v_j)}{\sum_k \exp\,(v_i.v_k)}$$



- Problem: The denominator is of size k.
- Hard to calculate the gradient

Continuous vectors representation

$$L = \frac{1}{T}\sum_{t=1}^{T}\sum_{\substack{-c\le j\le c\\ j\neq 0}} \log p(w_{j+t}|w_t)$$

$$p(w_i|w_j) = \frac{\exp(v_i.v_j)}{\sum_k \exp(v_i.v_k)}$$

- Problem: The denominator is of size k.
- Hard to calculate the gradient
- Change the objective:

$$p(w_i|w_j) = \frac{1}{1+\exp(v_i.v_j)}$$

- Has trivial solution!
- Introduce artificial negative instances!

$$L_{ij} = \log\sigma(v_i.v_j) + \sum_{l=1}^{k} E_{w_l\sim P_n(w)}[\log\sigma(-v_l.v_j)]$$

References

Slides from: http://levyomer.wordpress.com/2014/04/25/linguistic-regularities-in-sparse-and-explicit-word-representations/

Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." *Advances in Neural Information Processing Systems*. 2013.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. "Linguistic Regularities in Continuous Space Word Representations." *HLT-NAACL*. 2013.