Analogies Explained: Towards Understanding Word Embeddings

Carl Allen, Timothy Hospedales

{carl.allen, t.hospedales}@ed.ac.uk

School of Informatics, University of Edinburgh

The Problem: linking semantics to geometry

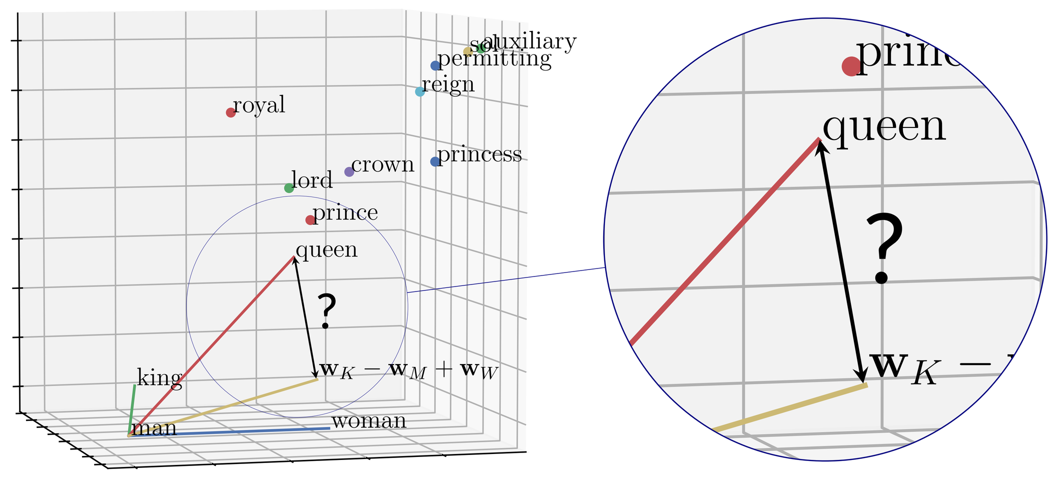
from:

"**man** is to **king** as **woman** is to **queen**"

explain:

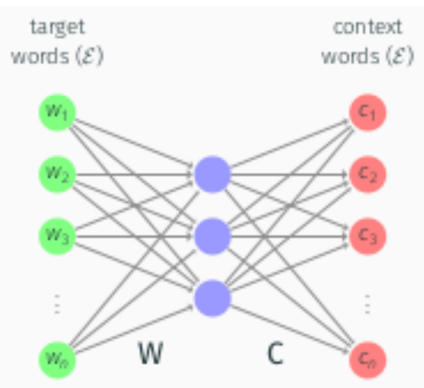
$$\mathbf{w}_{king} - \mathbf{w}_{man} + \mathbf{w}_{woman} \approx \mathbf{w}_{queen}$$

or rather:



Word2Vec: SkipGram with Negative Sampling

Mikolov et al. (2013a,b)



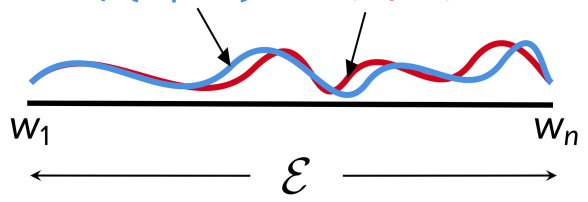
- $p(c_j|w_i)$ by **softmax** expensive
- use **sigmoid** with negative sampling (k)
- Levy and Goldberg (2014)

$$\mathbf{w}_i^\top \mathbf{c}_j \approx \log \frac{p(w_i,c_j)}{p(w_i)p(c_j)} - \log k$$

$$\boxed{\mathbf{W}^\top\mathbf{C} \approx \mathbf{PMI} - \log k}$$

Paraphrase† of $\mathcal{W}$ by w_*

Word $w_* \in \mathcal{E}$ **paraphrases** words $\mathcal{W} = \{w_1, \ldots, w_m\} \subseteq \mathcal{E}$, if w_* and $\mathcal{W}$ are *semantically interchangeable*.



Definition (D1): $w_* \in \mathcal{E}$ paraphrases $\mathcal{W} \subseteq \mathcal{E}$, $|\mathcal{W}| < l$, if paraphrase error $\rho^{\mathcal{W},w_*} \in \mathbb{R}^n$ is (element-wise) small:

$$\rho_j^{\mathcal{W},w_*} = \log \frac{p(c_j|w_*)}{p(c_j|\mathcal{W})}, \quad c_j \in \mathcal{E}$$

†Inspired by Gittens et al. (2017)

Paraphrase: $\mathbf{PMI}_1 + \mathbf{PMI}_2 \approx \mathbf{PMI}_*$?

$$\mathrm{PMI}(w_*, c_j) - \big(\mathrm{PMI}(w_1,c_j) + \mathrm{PMI}(w_2,c_j)\big)$$

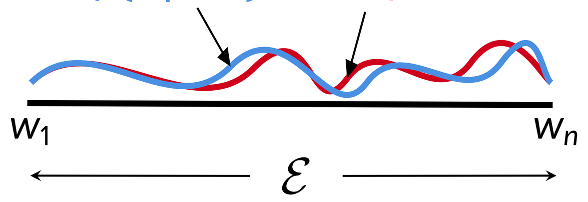
$$= \underbrace{\log \frac{p(c_j|w_*)}{p(c_j|\mathcal{W})}}_{\substack{\rho_j^{\mathcal{W},w_*} \\ \text{paraphrase error}}} + \underbrace{\log \frac{p(\mathcal{W}|c_j)}{p(w_1|c_j)p(w_2|c_j)}}_{\substack{\sigma_j^{\mathcal{W}} \\ \text{conditional independence error}}} - \underbrace{\log \frac{p(\mathcal{W})}{p(w_1)p(w_2)}}_{\substack{\tau^{\mathcal{W}} \\ \text{independence error}}}$$

Lemma 1: For any word $w_* \in \mathcal{E}$, words $\mathcal{W} \subseteq \mathcal{E}$, $|\mathcal{W}| < l$:

$$\mathbf{PMI}_* = \sum_{w \in \mathcal{W}} \mathbf{PMI}_i + \rho^{\mathcal{W},w_*} + \sigma^{\mathcal{W}} - \tau^{\mathcal{W}}\mathbf{1}$$

Generalised Paraphrase

Replace word w_* with word set $\mathcal{W}_* \subseteq \mathcal{E}$:



Lemma 2: For word sets $\mathcal{W}, \mathcal{W}_* \subseteq \mathcal{E}$, $|\mathcal{W}|, |\mathcal{W}_*| < l$:

$$\sum_{w \in \mathcal{W}_*} \mathbf{PMI}_i = \sum_{w \in \mathcal{W}} \mathbf{PMI}_i + \rho^{\mathcal{W},\mathcal{W}_*} + \sigma^{\mathcal{W}} - \sigma^{\mathcal{W}_*} - (\tau^{\mathcal{W}} - \tau^{\mathcal{W}_*})\mathbf{1}$$

Linking semantics to geometry

So, if:

$\mathcal{W} = \{woman, king\}$

paraphrases $\mathcal{W}_* = \{man, queen\}$,

then:

$$\mathbf{PMI}_{queen} \approx \mathbf{PMI}_{king} - \mathbf{PMI}_{man} + \mathbf{PMI}_{woman} + \underbrace{\sigma^{\mathcal{W}} - \sigma^{\mathcal{W}_*} - (\tau^{\mathcal{W}} - \tau^{\mathcal{W}_*})\mathbf{1}}_{\text{net dependence error}}$$

Geometry links to semantics, but to the wrong relationship.

Word Transformation: changing perspective

A paraphrase w_* of $\mathcal{W}$ can be thought of as a **word transformation** from some $w \in \mathcal{W}$ to w_* by *adding*:

$$\{man, royal\} \approx_p king \quad \Longrightarrow \quad man \xrightarrow{+royal} king$$

Adding ***context*** $\Longrightarrow$ induced distributions better align.

Similarly, consider paraphrase $\mathcal{W}$ of $\mathcal{W}_*$ as word transformation from a $w \in \mathcal{W}$ to $w_* \in \mathcal{W}_*$ by adding:

$$w \xrightarrow{+\mathcal{W}^+} \mathcal{W} \approx_p \mathcal{W}_* \xleftarrow{+\mathcal{W}^-} w_*$$

or adding to one side and *subtracting* from the other:

$$w \xrightarrow{+\mathcal{W}^+} \mathcal{W} \approx_p \mathcal{W}_* \xrightarrow{-\mathcal{W}^-} w_*$$

word transformation

A generalised paraphrase ***is*** a word transformation from $w \in \mathcal{W}$ to $w_* \in \mathcal{W}_*$:

- added words *narrow* context
- subtracted words *broaden* context

Providing a "richer dictionary" to explain the difference between w and w_*, or, how **"w is to w_*"**.

Definition (D4): We say *"w_a is to w_{a^*} as w_b is to w_{b^*}" iff* there exist $\mathcal{W}^+, \mathcal{W}^- \subseteq \mathcal{E}$ that simultaneously transform w_a to w_{a^*} and w_b to w_{b^*}.

We say: "**man** is to **king** as **woman** is to **queen**"

iff $\exists\ \mathcal{W}^+, \mathcal{W}^- \subseteq \mathcal{E}$ that transform

man to king ***and*** woman to queen.

w.l.o.g. choose $\mathcal{W}^+ = \{king\}$, $\mathcal{W}^- = \{man\}$.

$$man \xrightarrow{+king} \mathcal{W} \approx_p \mathcal{W}_* \xrightarrow{-man} king$$

word transformation

$$woman \xrightarrow{+king} \mathcal{W} \approx_p \mathcal{W}_* \xrightarrow{-man} queen$$

word transformation

Routemap

"**man** is to **king** as **woman** is to **queen**"

$\Updownarrow$

man *transforms to* **king** as **woman** *transforms to* **queen**

$\Updownarrow$

{**woman**, **king**} *paraphrases* {**man**, **queen**}

$\Downarrow$

$$\mathbf{PMI}_{king} - \mathbf{PMI}_{man} + \mathbf{PMI}_{woman} \approx \mathbf{PMI}_{queen}$$

$\Downarrow$

$$\mathbf{w}_{king} - \mathbf{w}_{man} + \mathbf{w}_{woman} \approx \mathbf{w}_{queen}$$

semantic

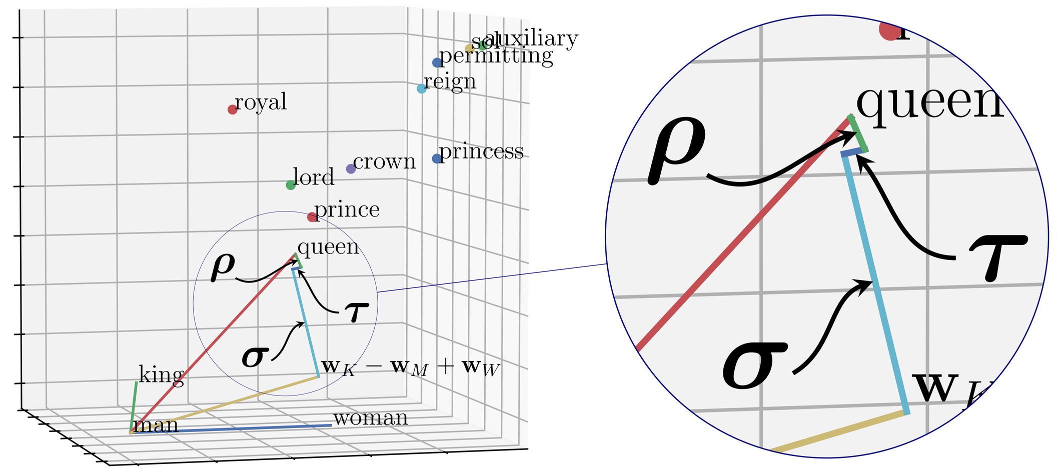
geometric

The Solution: linking semantics to geometry

"**man** is to **king** as **woman** is to **queen**"

implies:

$$\mathbf{w}_{king} - \mathbf{w}_{man} + \mathbf{w}_{woman} \overset{\rho,\sigma,\tau}{\approx} \mathbf{w}_{queen}$$



References

Alex Gittens, Dimitris Achlioptas, and Michael W Mahoney. Skip-Gram - Zipf + Uniform = Vector Additivity. In *Association for Computational Linguistics*, 2017.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, 2014.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *Workshop of the International Conference on Learning Representations*, 2013a.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013b.