Text factorisation — II: Distributional semantics

Course "Text-as-data analysis of international trade"

Dmitriy Skougarevskiy

The Institute for the Rule of Law, European University at Saint Petersburg

Saint Petersburg
4 May 2018

Course outline

1. Preferential Trade Agreements and International Economic Order
2. Gravity and Gravitas
3. Text factorisation — I: Bag-of-words methods
4. **Text factorisation — II: Distributive semantics**
5. Welfare effects of Preferential Trade Agreements

Outline

1. Understanding the meaning of the words
   - Context matters
   - Is TCM able to capture meaning?
2. Using the TCM
   - Computing semantic distances
   - Factorising TCM to arrive at document vectors
   - Word2vec model to learn word emebeddings

Outline

1. Understanding the meaning of the words
   - Context matters
   - Is TCM able to capture meaning?
2. Using the TCM
   - Computing semantic distances
   - Factorising TCM to arrive at document vectors
   - Word2vec model to learn word emebeddings

tesgüino

tesgüino

Tesgüino

tesgüino

A bottle of tesgüino is on the table.
Everybody likes tesgüino.
Tesgüino makes you drunk.
We make tesgüino out of corn.

(Jurafsky and Martin, 2017, Ch. 15.1)

Firth (1957): "You shall know a word by the company it keeps!"

The Distributional hypothesis

- Linguistic items with similar distributions have similar meanings (Harris, 1954)

Figure 1: Syntagmatic vs. Paradigmatic relations (Sahlgren, 2008, p. 6)

	Paradigmatic			
Syntagmatic	she	adores	green	paint
	he	likes	blue	dye
	they	love	red	colour

- "Paradigmatic relations hold between linguistic entities that occur in the same context but not at the same time, like the words "hungry" and "thirsty" in a sentence "the wolf is [hungry|thirsty]"" (ibid.)
- "Syntagmatic relations concern positioning, and relate entities that co-occur in the text, as in a normal sentence like "the wolf is hungry."" (ibid.)
- We can feed a distributional model with syntagmatic relations "if we collect information about word co-occurence, and with paradigmatic relations if we collect information about which words tend to share neighbors." (ibid.)

The Distributional hypothesis

- Linguistic items with similar distributions have similar meanings (Harris, 1954)

Figure 1: Syntagmatic vs. Paradigmatic relations (Sahlgren, 2008, p. 6)

	Paradigmatic			
	she	adores	green	paint
Syntagmatic	he	likes	blue	dye
	they	love	red	colour

- "Paradigmatic relations hold between linguistic entities that occur in the same context but not at the same time, like the words "hungry" and "thirsty" in a sentence "the wolf is [hungry|thirsty]"" (ibid.)
- "Syntagmatic relations concern positioning, and relate entities that co-occur in the text, as in a normal sentence like "the wolf is hungry."" (ibid.)
- We can feed a distributional model with syntagmatic relations "if we collect information about word co-occurence, and with paradigmatic relations if we collect information about which words tend to share neighbors." (ibid.)

The Distributional hypothesis

- Linguistic items with similar distributions have similar meanings (Harris, 1954)

Figure 1: Syntagmatic vs. Paradigmatic relations (Sahlgren, 2008, p. 6)

	Paradigmatic			
	she	adores	green	paint
Syntagmatic	he	likes	blue	dye
	they	love	red	colour

- "Paradigmatic relations hold between linguistic entities that occur in the same context but not at the same time, like the words "hungry" and "thirsty" in a sentence "the wolf is [hungry|thirsty]"" (ibid.)
- "Syntagmatic relations concern positioning, and relate entities that co-occur in the text, as in a normal sentence like "the wolf is hungry."" (ibid.)
- We can feed a distributional model with syntagmatic relations "if we collect information about word co-occurence, and with paradigmatic relations if we collect information about which words tend to share neighbors." (ibid.)

The Distributional hypothesis

- Linguistic items with similar distributions have similar meanings (Harris, 1954)

Figure 1: Syntagmatic vs. Paradigmatic relations (Sahlgren, 2008, p. 6)

	Paradigmatic			
	she	adores	green	paint
Syntagmatic	he	likes	blue	dye
	they	love	red	colour

- "Paradigmatic relations hold between linguistic entities that occur in the same context but not at the same time, like the words "hungry" and "thirsty" in a sentence "the wolf is [hungry|thirsty]"" (ibid.)
- "Syntagmatic relations concern positioning, and relate entities that co-occur in the text, as in a normal sentence like "the wolf is hungry."" (ibid.)
- We can feed a distributional model with syntagmatic relations "if we collect information about word co-occurence, and with paradigmatic relations if we collect information about which words tend to share neighbors." (ibid.)

Context window

- We can look at words in context. Phrases like
  - `to have a splendid time in Rome`
  - `to have a wonderful time in Rome`
- become with a symmetric context window of size 2 (Sahlgren, 2008):
  - `splendid: (have a) + (time in)`
  - `wonderful: (have a) + (time in)`
  - `time: (a splendid) + (in Rome)`
  - `time: (a wonderful) + (in Rome)`
- We slide the context window one word at a time until we reach the end of the phrase

Term co-occurrence matrix

- So far we've worked with document-term matrices (DTMs)

	John	likes	to	watch	movies	Mary	too	also	football	games
doc1	1	2	1	1	2	1	1	0	0	0
doc2	1	1	1	1	0	0	0	1	1	1

- Counts of words that occur together within the context window are stored in a term co-occurrence matrix (TCM).
- TCM for "whereof one cannot speak thereof one must be silent."

	whereof	one	cannot	speak	thereof	must	be	silent
whereof	0	1	0	0	0	0	0	0
one	0	0	1	0	0	1	0	0
cannot	0	0	0	1	0	0	0	0
speak	0	0	0	0	1	0	0	0
thereof	0	1	0	0	0	0	0	0
must	0	0	0	0	0	0	1	0
be	0	0	0	0	0	0	0	1
silent	0	0	0	0	0	0	0	0

- Context window size? Is it symmetric (hint: word order)?
- DTM vs TCM and syntagmatic vs paradigmatic relations

Outline

1. Understanding the meaning of the words
   - Context matters
   - Is TCM able to capture meaning?
2. Using the TCM
   - Computing semantic distances
   - Factorising TCM to arrive at document vectors
   - Word2vec model to learn word emebeddings

Benefits and challenges of the model

- We are abstracting away from almost all linguistic knowledge when we apply the Distributional hypothesis and build TCMs
  - word order may not be captured in successful way
  - we do not store information on part-of-speech of elements
  - we fail to capture information on dependencies between words
- Strikingly, in practice all this barely matters, TCMs are enough to do well empirically

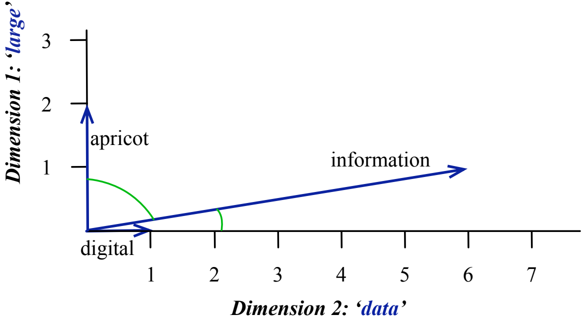
Outline

1. Understanding the meaning of the words
   - Context matters
   - Is TCM able to capture meaning?
2. Using the TCM
   - Computing semantic distances
   - Factorising TCM to arrive at document vectors
   - Word2vec model to learn word emebeddings

- Consider this "apricot" TCM (Jurafsky and Martin, 2017, Ch. 15.3):

	apricot	digital	information
apricot	2	0	0
digital	0	1	2
information	1	6	1

Figure 2: "Apricot" TCM representation (Jurafsky and Martin, 2017, Fig. 15.10)



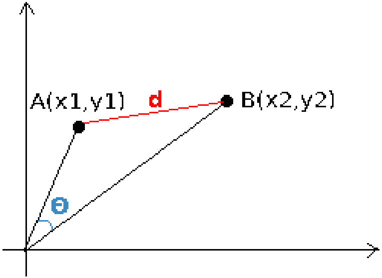
- Angle (digital, information) $<$ (apricot, information). When two vectors are similar, their angle is smaller, but cosine is larger

Cosine distance in natural language processing

- Many distance metrics exist (most prominent are Euclidean, Cosine and Manhattan)
- Cosine distance between vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ is defined as vector dot product over vector norm:

$$\cos(\boldsymbol{v}, \boldsymbol{w}) = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{|\boldsymbol{v}|\,|\boldsymbol{w}|} = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}}$$

Figure 3: Relationship between Euclidean and Cosine distances (source)



- Cosine distance normalises for vector length, allowing us to compare texts of uneven size

Outline

1. Understanding the meaning of the words
   - Context matters
   - Is TCM able to capture meaning?

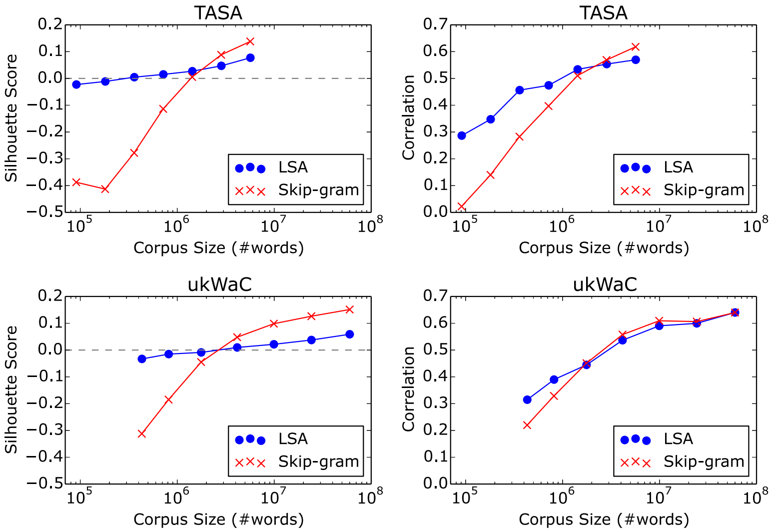
2. Using the TCM
   - Computing semantic distances
   - Factorising TCM to arrive at document vectors
   - Word2vec model to learn word emebeddings

Creating document vectors from word vectors

1. TCM is a symmetric #words×#words matrix. We can SVD-decompose it to arrive at word vectors.
2. Then we can utilise the DTM to aggregate TCM-produced word vectors (perhaps, after weighting)
3. This will give us document vectors from appropriately averaged word vectors.
   - Paradigmatic relations between words $\Rightarrow$ paradigmatic relations between documents with the aid of syntagmatic relations between words within documents

In practice, Step 1 (SVD reduction of TCM) doesn't work very well with enough data:

Figure 4: LSA vs skipgram word2vec model performance (Altszyler et al., 2016, fig. 1)



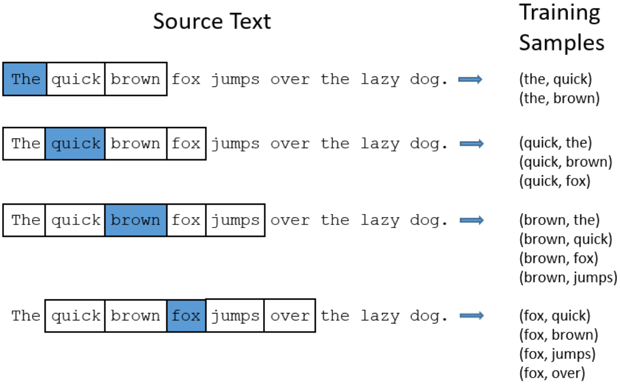
Outline

1. Understanding the meaning of the words
   - Context matters
   - Is TCM able to capture meaning?
2. Using the TCM
   - Computing semantic distances
   - Factorising TCM to arrive at document vectors
   - Word2vec model to learn word emebeddings

On word2vec model

- word2vec model by Mikolov et al. (2013) revolutionised the field. At core: a neural network that seeks to predict next word in context window

Figure 5: One-hot encoding of training data (source)



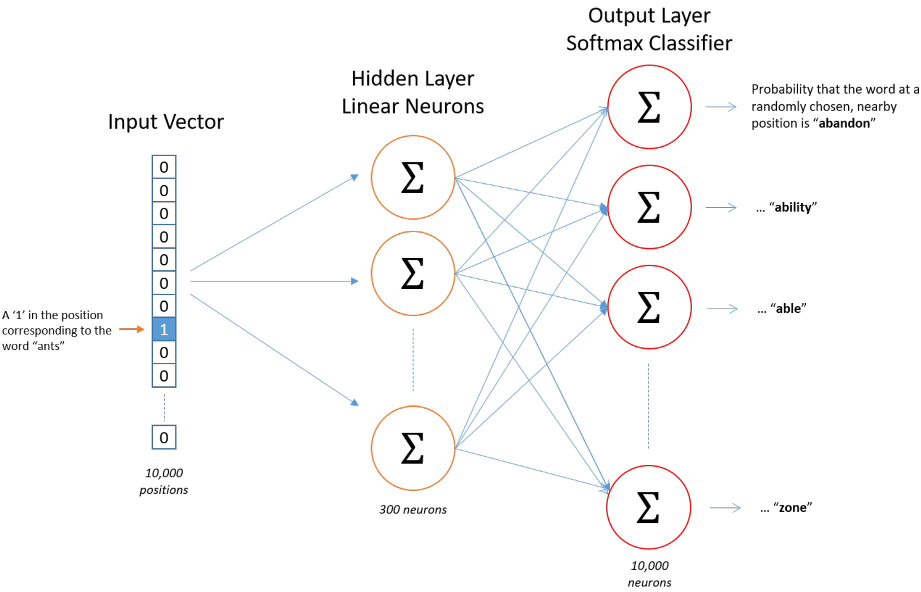
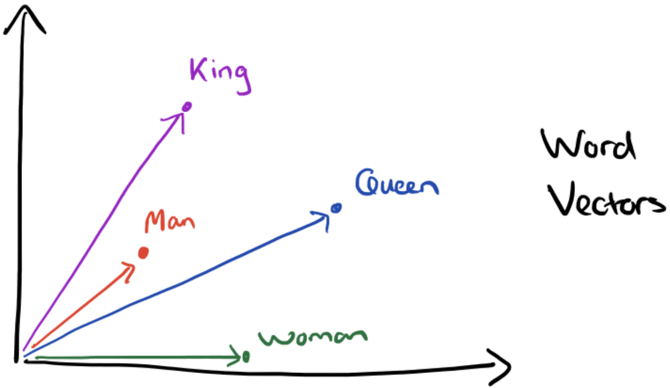


Figure 6: Skip-gram neural net architecture (source)

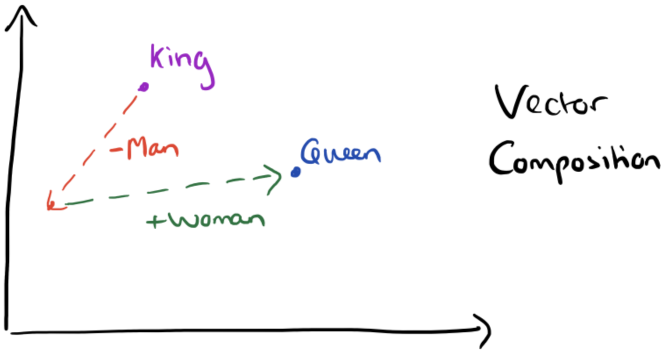
Vector operations

Figure 7: Trained word embeddings for words {King, Queen, Man, Woman, ...} (source)



Vector operations with word embeddings

Figure 8: Vector "King - Man + Woman" is closest to vector of word Queen (source)



An aside: predicting vowels with neural network

- http://playground.tensorflow.org/

green — input/output data, blue* — weights, brown*— bias (courtesy of http://bit.ly/2EBVDYr)

- Architecture is set by a user
- Very rough intuition: Neural net optimises (*) weights and bias to minimise loss function on test set

An aside: predicting vowels with neural network

- https://jsfiddle.net/memoryfull/g5zumxmh/4/
- One-hot-encoding of text, standardised input and output [0...1]
- No architecture: Neuroevolution of Augmenting Topologies (NEAT)
- Genotype: element types, connections and weights
- Phenotype: network architecture
- Parameters:
  - `equal: true`
  - `population: 50`
  - `elitism: 5`
  - `iterations: 1500`
  - `error: 0.03`
  - `clear: true`

Take-aways

- Linguistic items with similar distributions have similar meanings — Distributional hypothesis
- By collecting information on which words tend to share neighbours we learn paradigmatic relations between words
- State-of-art in computing word vectors is the word2vec model (fastText flavour)
- Word vectors can be averaged at document level to arrive at document vectors capturing paradigmatic relations between documents

Thank you for your attention!

References — I

Altszyler, E., M. Sigman, S. Ribeiro, and D. F. Slezak (2016). Comparative study of lsa vs word2vec embeddings in small corpora: a case study in dreams database. *arXiv preprint arXiv:1610.01520*.

Firth, J. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*.

Harris, Z. (1954). Distributional structure. *Word 10*(23), 146–162.

Jurafsky, D. and J. Martin (2017). *Speech and language processing* (3 ed.). Pearson.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies 20*, 33–53.