VL Embeddings: Overview & Intro

Katja Markert, Ines Rehbein
& Philipp Wiesenbach (Tutor)

Uni Heidelberg

SS 2019

Definition

Embeddings

Representing a linguistic structure such as a character, word, phrase or sentence as a vector of real numbers.

We concentrate on word embeddings (with some extensions towards sentence and phrasal embeddings). Therefore embeddings are a function from a Vocabulary V to the $\mathbb{R}^n$.

The vector for *banana* in Spacy: $(2.022e^{-1}, -7.66e^{-2}, 3.70e^{-1} \ldots)$

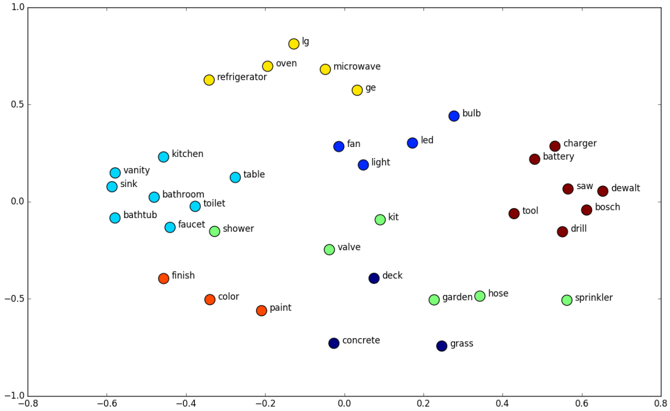
Embeddings

Often people only call dense vectors trained via neural networks as embeddings but there is no real reason not to call sparse vectors or dense vectors generated via matrix factorisation embeddings as well.

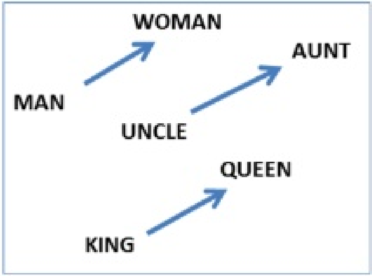
Advantages of representing words as vectors:

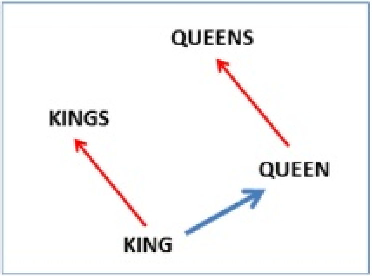
- All vector and matrix operations from linear algebra at our disposal
- Input to machine learning models need to be numbers.

Word embeddings cluster similar words in vector space

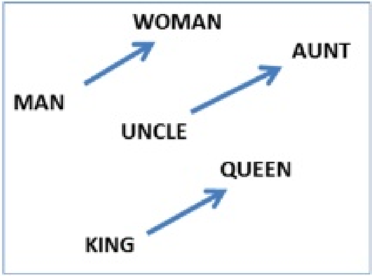


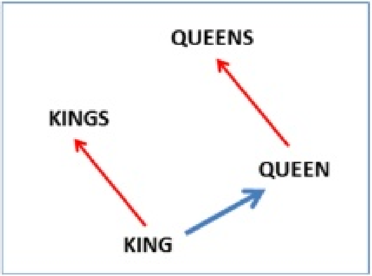
Word embeddings capture analogies





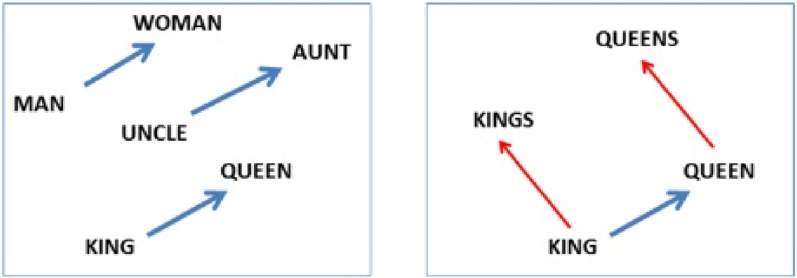
Word embeddings capture analogies





MAN is to WOMAN as KING is to QUEEN

Word embeddings capture analogies

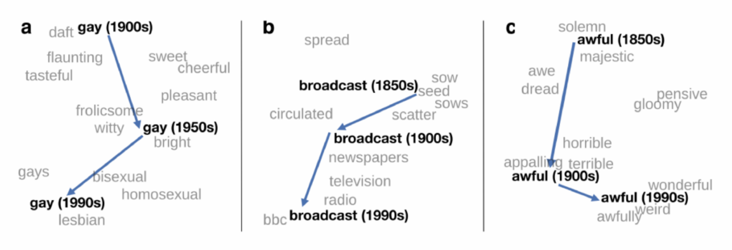


MAN is to WOMAN as KING is to QUEEN

We can solve analogies, using simple arithmetic:

KING - MAN + WOMEN = QUEEN

Applications: Language change



~30 million books, 1850-1990, Google Books data

Work by Hamilton and Jurafsky. See
`https://nlp.stanford.edu/projects/histwords/`

Applications

Embeddings used in almost all current systems as building blocks:

- Coreference resolution: *Donald Trump* ... *Hilary Clinton* ... *the president*.
- Text classification: Present text via word embeddings instead of words → topic classification, sentiment classification ...
- Input as lowest level into sequence-to-sequence models → summarization, generation

Overview VL Embeddings

Topics

- Part I: Lectures on count-based embeddings
- Part II: Lectures on prediction-based embeddings
- Part III: Reading sessions & short student presentations
  - Multi-modal embeddings
  - Multi-lingual embeddings
  - Multi-sense embeddings
  - Bias in neural representations
- Lab sessions
  - Collocations, sparse matrices
  - Matrix factorisation
  - Evaluation and visualisation of word embeddings
  - Multi-modal embeddings

In the course

We will learn

- how the models work

In the course

We will learn

- how the models work
- how to train word embeddings

In the course

We will learn

- how the models work
- how to train word embeddings
- how to evaluate and visualise word embeddings

In the course

We will learn

- how the models work
- how to train word embeddings
- how to evaluate and visualise word embeddings

We will look at

- different types and variations of word embeddings

In the course

We will learn

- how the models work
- how to train word embeddings
- how to evaluate and visualise word embeddings

We will look at

- different types and variations of word embeddings
- embeddings beyond (and below) the word level

In the course

We will learn

- how the models work
- how to train word embeddings
- how to evaluate and visualise word embeddings

We will look at

- different types and variations of word embeddings
- embeddings beyond (and below) the word level
- the relation between matrix factorisation and neural embeddings

Count-based embeddings

Association measures

Association measures between two tokens based on co-occurrence:

- How often do the tokens co-occur?
- What is the distribution of them co-occurring? (mean, variance)
- Do they co-occur more often than chance? (significance tests)
- How much information do the two tokens contribute to each other? (Information theory)

Bigram	$f(w_1)$	$f(w_2)$	$f(w_1, w_2)$	t-test	PMI
unsalted butter	24	320	20	4.47	15.19
over many	13 484	10570	20	2.24	1.01

Sparse matrices

Extension from bigrams to windows leads to matrices:

	astronaut	*cosmonaut*	*tomato*
NASA	4	0	1
Roscosmos	0	4	0
avocado	0	0	7
salad	0	1	10

Problems:

- Long vectors. Length $= |V|$. Many weights to tune in ML.
- Many low frequencies due to Zipfs law.
- (near)-synonyms are in different dimensions: *astronaut/cosmonaut*
- Dense vectors tend to generalize better

Singular value decomposion (SVD)

$$
\underset{m \times n}{\overset{\hat{X}}{\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & \\ \vdots & \vdots & \ddots & \\ x_{m1} & & & x_{mn} \end{pmatrix}}} \approx \underset{m \times r}{\overset{U}{\begin{pmatrix} u_{11} & \cdots & u_{1r} \\ \vdots & \ddots & \\ u_{m1} & & u_{mr} \end{pmatrix}}} \underset{r \times r}{\overset{S}{\begin{pmatrix} s_{11} & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & s_{rr} \end{pmatrix}}} \underset{r \times n}{\overset{V^{\intercal}}{\begin{pmatrix} v_{11} & \cdots & v_{1n} \\ \vdots & \ddots & \\ v_{r1} & & v_{rn} \end{pmatrix}}}
$$

Properties:

- Low-dimensional approximation. $r << n$
- Most important hidden dimensions captured

Maths Background

Concentrating on background that you will need throughout your studies:

- Significance tests
- Information theory (entropy, cross-entropy, mutual information, Kullback-Leibner)
- Linear Algebra
    - Vector operations and normalizations
    - Metrics and distances
    - Matrix operations
    - Matrix factorisation

Neural language models

- Bengio et al. (2003)
  - Extension to traditional n-gram language models (LM)

  $\Rightarrow$ replace conditional probability with neural network (NN):
    - represent each word by small vector
    - jointly estimate parameters of NN and vectors

Neural language models

- Bengio et al. (2003)
  - Extension to traditional n-gram language models (LM)
  - ⇒ replace conditional probability with neural network (NN):
    - represent each word by small vector
    - jointly estimate parameters of NN and vectors
- Collobert and Weston (2008):
  - replace max-likelihood with max-margin approach
  - learn to score correct n-grams higher than random n-grams

Neural language models

- Bengio et al. (2003)
  - Extension to traditional n-gram language models (LM)

  ⇒ replace conditional probability with neural network (NN):
    - represent each word by small vector
    - jointly estimate parameters of NN and vectors
- Collobert and Weston (2008):
  - replace max-likelihood with max-margin approach
  - learn to score correct n-grams higher than random n-grams
- Mikolov et al (2013a,b):
  - efficient log-linear neural language models (Word2vec)
  - remove hidden layers, use larger context windows and negative sampling

Neural language models

- Bengio et al. (2003)
  - Extension to traditional n-gram language models (LM)
  - ⇒ replace conditional probability with neural network (NN):
    - represent each word by small vector
    - jointly estimate parameters of NN and vectors
- Collobert and Weston (2008):
  - replace max-likelihood with max-margin approach
  - learn to score correct n-grams higher than random n-grams
- Mikolov et al (2013a,b):
  - efficient log-linear neural language models (Word2vec)
  - remove hidden layers, use larger context windows and negative sampling

Goal of traditional LM

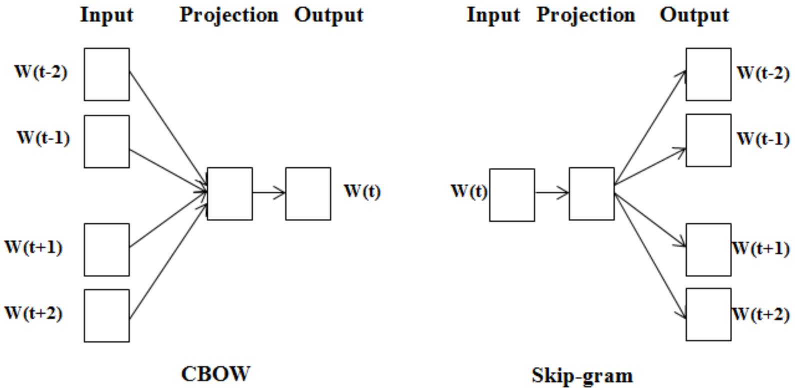
- low-perplexity LM that can predict probability of next word

New goal

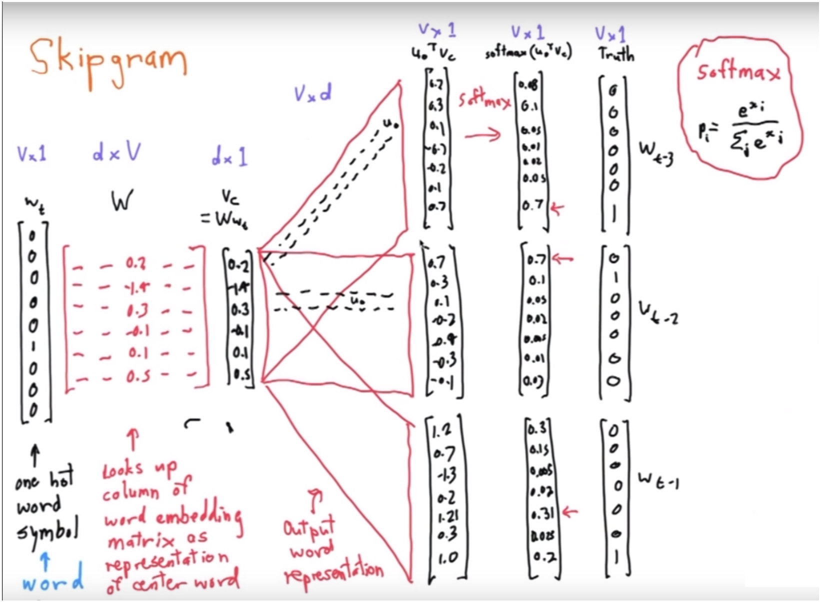
⇒ learn word representations that are useful for downstream tasks

Prediction-based embeddings

Word2vec



Prediction-based embeddings



Prediction-based embeddings

- Word2vec ingredients:
  - softmax, hierarchical softmax, negative sampling
  - gradient-based optimisation (Stochastic Gradient Descend)
  - backpropagation

Evaluation of word embeddings

- Intrinsic evaluation
  - Word similarity and analogy tasks

  $\Rightarrow$ Correlation with human judgments

- Extrinsic evaluation
  - plug-in pretrained embeddings as features for different NLP tasks
  - or let the model learn task-specific embeddings from scratch

Evaluation of word embeddings

- Intrinsic evaluation
  - Word similarity and analogy tasks

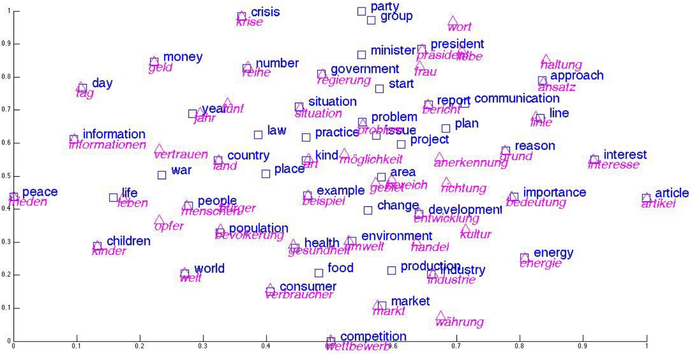
  $\Rightarrow$ Correlation with human judgments

- Extrinsic evaluation
  - plug-in pretrained embeddings as features for different NLP tasks
  - or let the model learn task-specific embeddings from scratch

- Collobert & Weston (2007): Fast Semantic Extraction Using a Novel Neural Network Architecture. Proceedings of ACL 2007.

Different types of word embeddings

Multilingual embeddings



Different types of word embeddings

Multilingual embeddings

- Bilingual mapping
  - Train word representations for each language independantly
  - Learn a mapping to transform representations from one space into the other
  - E.g. Mikolov et al. (2013)
- Monolingual adaptation
  - Given: monolingual embeddings
  - Learn target representations, based on bilingual constraints from MT word alignments
  - E.g. Zou et al. (2013)
- Bilingual training
  - Jointly learn multilingual representations from scratch
  - E.g. Hermann and Blunsom (2014), Luong et al. (2015)

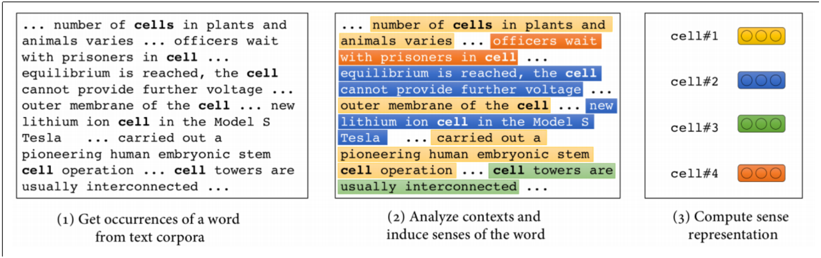
Different types of word embeddings

Multilingual embeddings

- Mikolov, Le & Sutskever (2013):
  Exploiting similarities among languages for machine translation. arXiv:1309.4168, 2013
- Luong, Pham & Manning (2015):
  Bilingual Word Representations with Monolingual Quality in Mind. *Workshop on Vector Space Modeling for NLP*
- Zou, Socher, Cer & Manning (2013):
  Bilingual Word Embeddings for Phrase-Based Machine Translation. *EMNLP 2013*
- Hermann & Blunsom (2014):
  Multilingual Models for Compositional Distributed Semantics. *ACL 2014*

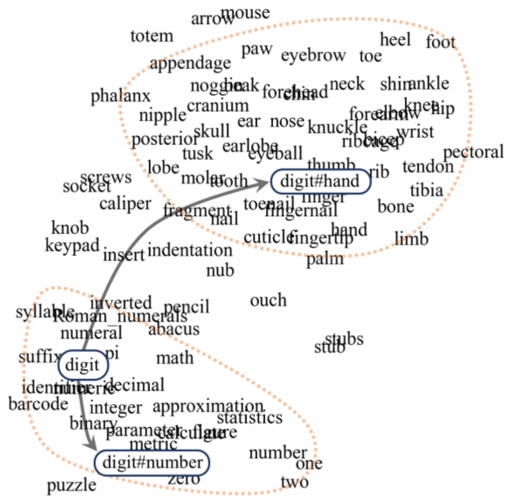
Different types of word embeddings

Multisense embeddings



Different types of word embeddings

Multisense embeddings



Different types of word embeddings

Multisense embeddings

- Multi-prototype neural language model (Huang et al. 2012)
  - Use local and global context to learn multiple representations
  - Cluster representations → learn multi-prototype vectors
  - New dataset: homonymy and polysemy of words in context

Different types of word embeddings

Multisense embeddings

- Multi-prototype neural language model (Huang et al. 2012)
  - Use local and global context to learn multiple representations
  - Cluster representations → learn multi-prototype vectors
  - New dataset: homonymy and polysemy of words in context
- Multi-sense Skip-Gram (Neelakantan et al. 2014)
  - Keep multiple vectors per word
  - Joint word sense discrimination and embedding learning

Different types of word embeddings

Multisense embeddings

- Multi-prototype neural language model (Huang et al. 2012)
  - Use local and global context to learn multiple representations
  - Cluster representations → learn multi-prototype vectors
  - New dataset: homonymy and polysemy of words in context
- Multi-sense Skip-Gram (Neelakantan et al. 2014)
  - Keep multiple vectors per word
  - Joint word sense discrimination and embedding learning
- Evaluation of multi-sense embeddings (Li & Jurafsky 2015):
  - Multi-sense embeddings based on Chinese Restaurant Processes (not part of lecture)
  - How useful are multi-sense embeddings for downstream applications? Evaluate multi-sense embeddings for POS tagging, NER, sentiment analysis, semantic relation identification and semantic relatedness

Different types of word embeddings

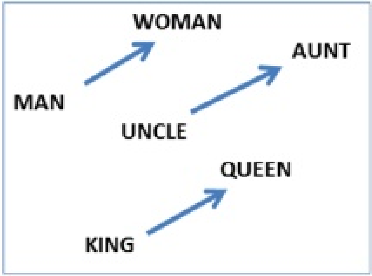
Multisense embeddings

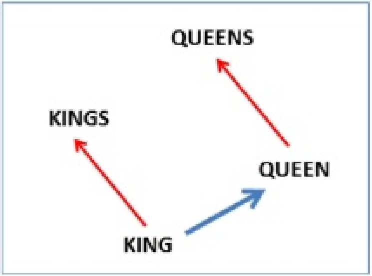
- Huang, Socher, Manning & Ng (2012): Improving word representations via global context and multiple word prototypes. *ACL 2012*
- Neelakantan, Shankar, Passos, & Mccallum (2014): Efficient non-parametric estimation of multiple embeddings per word in vector space. *EMNLP 2014*
- Li & Jurafsky (2015): Do multi-sense embeddings improve natural language understanding? *EMNLP 2015*

Different types of word embeddings

Beyond words – Compositionality

We can use arithmetic operations on word vectors:





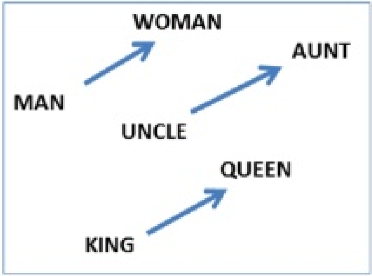
KING - MAN + WOMEN = QUEEN

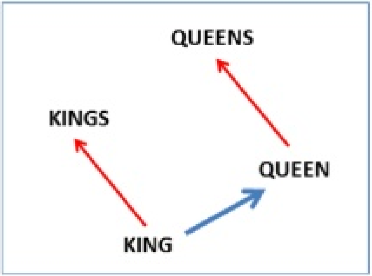
Can we also compute (or learn) representations for phrases?

Different types of word embeddings

Beyond words – Compositionality

We can use arithmetic operations on word vectors:





KING - MAN + WOMEN = QUEEN

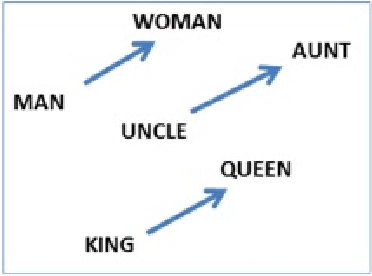
Can we also compute (or learn) representations for phrases?

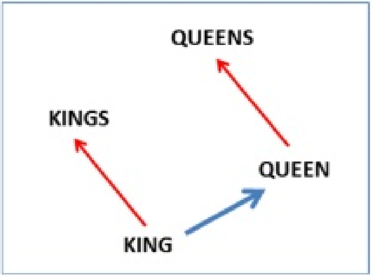
FRAU + MINISTER = MINISTERIN

Different types of word embeddings

Beyond words – Compositionality

We can use arithmetic operations on word vectors:





KING - MAN + WOMEN = QUEEN

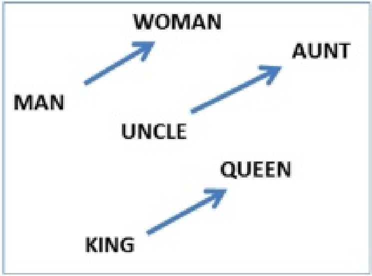
Can we also compute (or learn) representations for phrases?

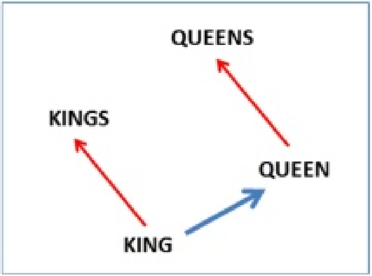
FRAU + SCHAUSPIELER = SCHAUSPIELERIN

Different types of word embeddings

Beyond words – Compositionality

We can use arithmetic operations on word vectors:





KING - MAN + WOMEN = QUEEN

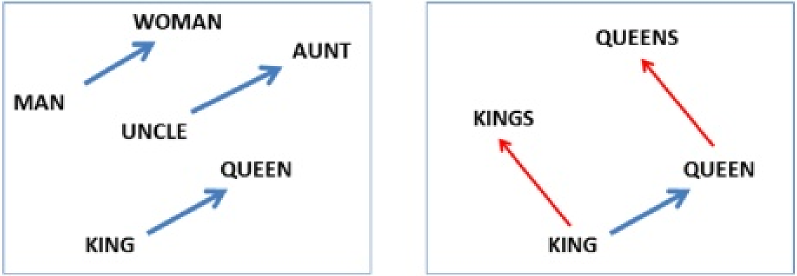
Can we also compute (or learn) representations for phrases?

TOTAL + GUT = SUPERGUT

Different types of word embeddings

Beyond words – Compositionality

We can use arithmetic operations on word vectors:



KING - MAN + WOMEN = QUEEN

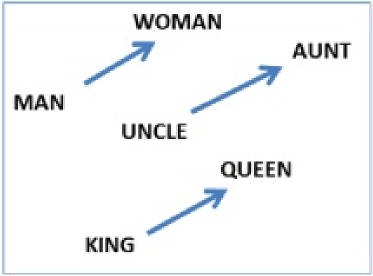
Can we also compute (or learn) representations for phrases?

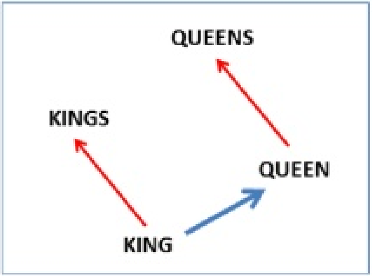
FRAU + MUTTER = EHEFRAU

Different types of word embeddings

Beyond words – Compositionality

We can use arithmetic operations on word vectors:





KING - MAN + WOMEN = QUEEN

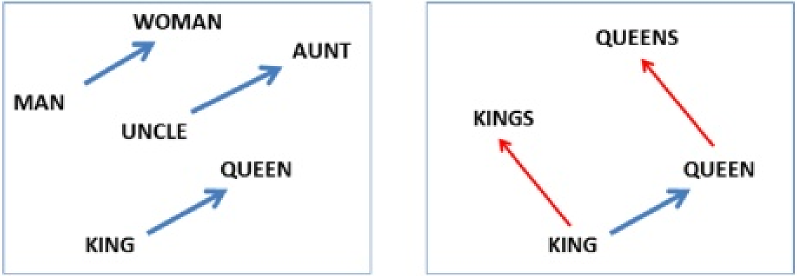
Can we also compute (or learn) representations for phrases?

MANN + VATER = EHEMANN

Different types of word embeddings

Beyond words – Compositionality

We can use arithmetic operations on word vectors:



KING - MAN + WOMEN = QUEEN

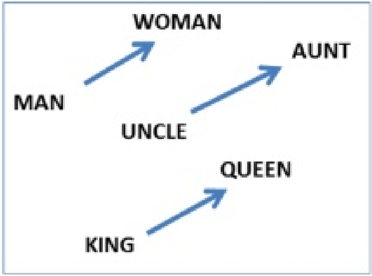
Can we also compute (or learn) representations for phrases?

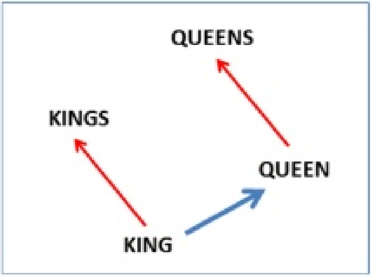
STARK + MANN = FRAU

Different types of word embeddings

Beyond words – Compositionality

We can use arithmetic operations on word vectors:





KING - MAN + WOMEN = QUEEN

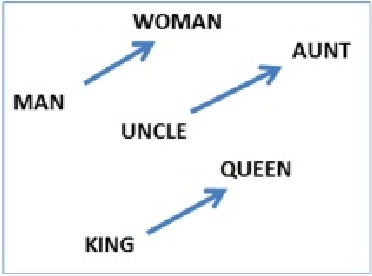
Can we also compute (or learn) representations for phrases?

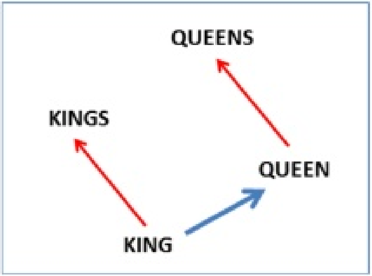
HAUPTSTADT + DEUTSCHLAND = EUROPA

Different types of word embeddings

Beyond words – Compositionality

We can use arithmetic operations on word vectors:





KING - MAN + WOMEN = QUEEN

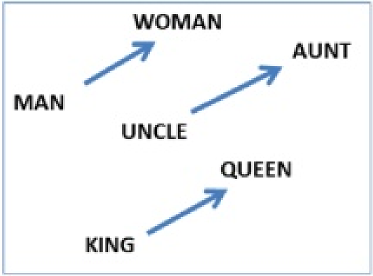
Can we also compute (or learn) representations for phrases?

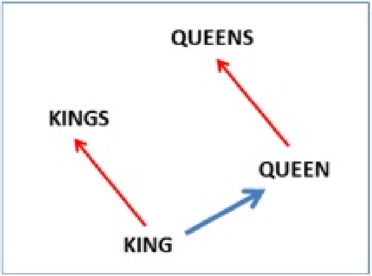
HAUPTSTADT + ITALIEN = BULGARIEN

Different types of word embeddings

Beyond words – Compositionality

We can use arithmetic operations on word vectors:





KING - MAN + WOMEN = QUEEN

Can we also compute (or learn) representations for phrases?

More meaningful representations?
What about sentences or documents?

Different types of word embeddings

Beyond words – Compositionality

- Modeling compositional meaning for phrases and sentences (Blacoe and Lapata 2012)
- Sent2vec (Pagliardini et al. 2018)
  - Learn sentence embedding as a sum of sub-sentence units
  - Uses average over ngrams in the sentence
- *Space: General purpose neural embeddings (Wu et al. 2018)
  - Learn entity embeddings with discrete feature representations from relations between those entities
    - *entities* (e.g. sentences, paragraphs, docs)
    - *features* (e.g. words, characters, char-ngrams, ...)

Different types of word embeddings

Beyond words – Compositionality

- Blacoe and Lapata (2012): A comparison of vector-based representations for semantic composition. *EMNLP 2012*
- Wu, Fisch, Chopra, Adams, Bordes and Weston (2018): StarSpace: Embed all the things! *AAAI 2018*
- Pagliardini, Gupta and Jaggi (2018): Unsupervised learning of sentence embeddings using compositional n-gram features. *NAACL-HLT 2018*

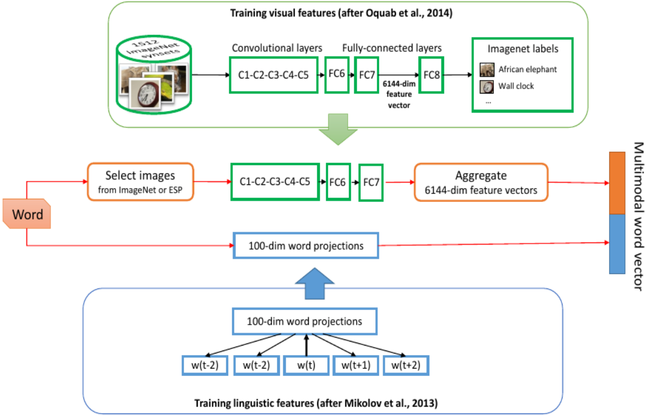
Image embeddings

- Images can be represented as vectors as well
- Therefore similarity between images can be computed as well
- If words can be mapped onto images, we can then use images to compute word similarity

word *alsatian* → ImageNet
`http://www.image-net.org/`

word *husky* → ImageNet
`http://www.image-net.org/`

Combining image and word embeddings



Picture from Kiela and Bottou (2014): Learning image embeddings using convolutional neural networks for improved multi-modal semantics. *Proceedings of EMNLP*

Typical questions for multimodal embeddings

1. How to retrieve images for words?
2. How to compute image vectors?
3. How to aggregate vectors from several images?
4. How to combine word and image vectors?
5. How to combine word/image vectors into sentence vectors?
6. When does it help? When are image vectors better and when word vectors?

NB: We will not go into the details of neural computer vision! If you want to do that, look at the seminal paper Simonyan, K. and A. Zisserman (2014). Very deep convolutional networks for large-scale image recognition.

Papers for short presentations: Multi-modal embeddings

- Bruni et al (2012): Distributional semantics in technicolor. *Proceedings of ACL*
- Kiela and Bottou (2014): Learning image embeddings using convolutional neural networks for improved multi-modal semantics. *Proceedings of EMNLP*
- Glavas et al (2017): If sentences could see: Investigating visual information for semantic textual similarity. *Proceedings of IWCS-2017*

Bias

Bias Definition I

Inconsistent behaviour of a system towards input from different demographic groups (adapted from Hardt et al 2016. Equality of opportunity in supervised learning. NIPS 2016)

Bias Definition II

Model is biased if it learns inappropriate stereotypical correlations of concepts

For us Definition 2 is relevant.

Bias

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Aus Bolukbasi et al (2016)

Or from Caliskan et al (2017)

- African-American names (*Leroy, Shaniqua*) had a higher similarity with unpleasant words (*abuse, stink, ugly*)
- European American names (*Brad, Greg, Courtney*) had a higher cosine with pleasant words (*love, peace, miracle)*

Papers for short presentations: Bias

Main question: How to measure bias in embeddings?

- Caliskan et al (2017): Semantics derived automatically from language corpora contain human-like biases. *Science 2017*
- Garg et al (2018): Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of sciences*
- Bolukbasi et al (2016): Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Proc of NIPS*

Papers for short presentations: Bias

Main question: How to mitigate bias?

- Zhao et al (2018): Learning gender-neutral word embeddings. *EMNLP 2018*
- Park et al (2018): Reducing Gender Bias in Abusive Language Detection *EMNLP 2018*
- Zhao et al (2018): Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. *NAACL 2018*