COMP 3225

Natural Language Processing

Word2Vec

Stuart E. Middleton

sem03@soton.ac.uk

University of Southampton

Overview

- Word2Vec

- <break - discussion point>

- Visualizing Embeddings
- Semantic Properties of Embeddings
- Bias and Embeddings

Word2Vec

- Dense embeddings work better than sparse embeddings
  - It is still not 100% clear why, but the weight of evidence is strong
- Static embeddings
  - Word2Vec embeddings are static embeddings, with a single embedding per word that does not change
  - This contrasts to contextual embeddings such as BERT where the vector for a word varies based on its context
- Self-supervision
  - Train a binary classifier to predict is word A is likely to show up near word B
  - No need for tagged training datasets, this problem formulation can be run on an unlabeled corpus
  - The classifier is not relevant, we want the learnt embedding weights

Word2Vec

- Word2Vec classifier uses a skip-gram model
  - Skip-gram algorithm is part of the word2vec package
- Implementations
  - Tensorflow https://www.tensorflow.org/tutorials/text/word2vec
  - Gensim https://radimrehurek.com/gensim/models/word2vec.html
- Skip-gram model
  - Target words + neighbouring context words are positive examples
  - Random sampling to create negative examples
  - Logistic regression to train classifier
  - Learned weights provide the embeddings

Word2Vec

- Input is a target word w and a window of L words to make some context words c

```
... lemon,  a [tablespoon of apricot jam,      a] pinch ...
               c1         c2    w     c3       c4
```

- P(+|w,c) = probability c is a context word for w
- P(-|w,c) = probability c is not a context word for w

$$P(+|w,c) \qquad P(-|w,c) = 1 - P(+|w,c)$$

Word2Vec

- Probability is based on embedding vector similarity of w and c
  - Dot product + logistic function σ to make it into a probability

$$\sigma(x) = \frac{1}{1+\exp(-x)}$$

$$P(+|w,c) = \sigma(c \cdot w) = \frac{1}{1+\exp(-c \cdot w)}$$

$$\begin{aligned} P(-|w,c) &= 1 - P(+|w,c) \\ &= \sigma(-c \cdot w) = \frac{1}{1+\exp(c \cdot w)} \end{aligned}$$

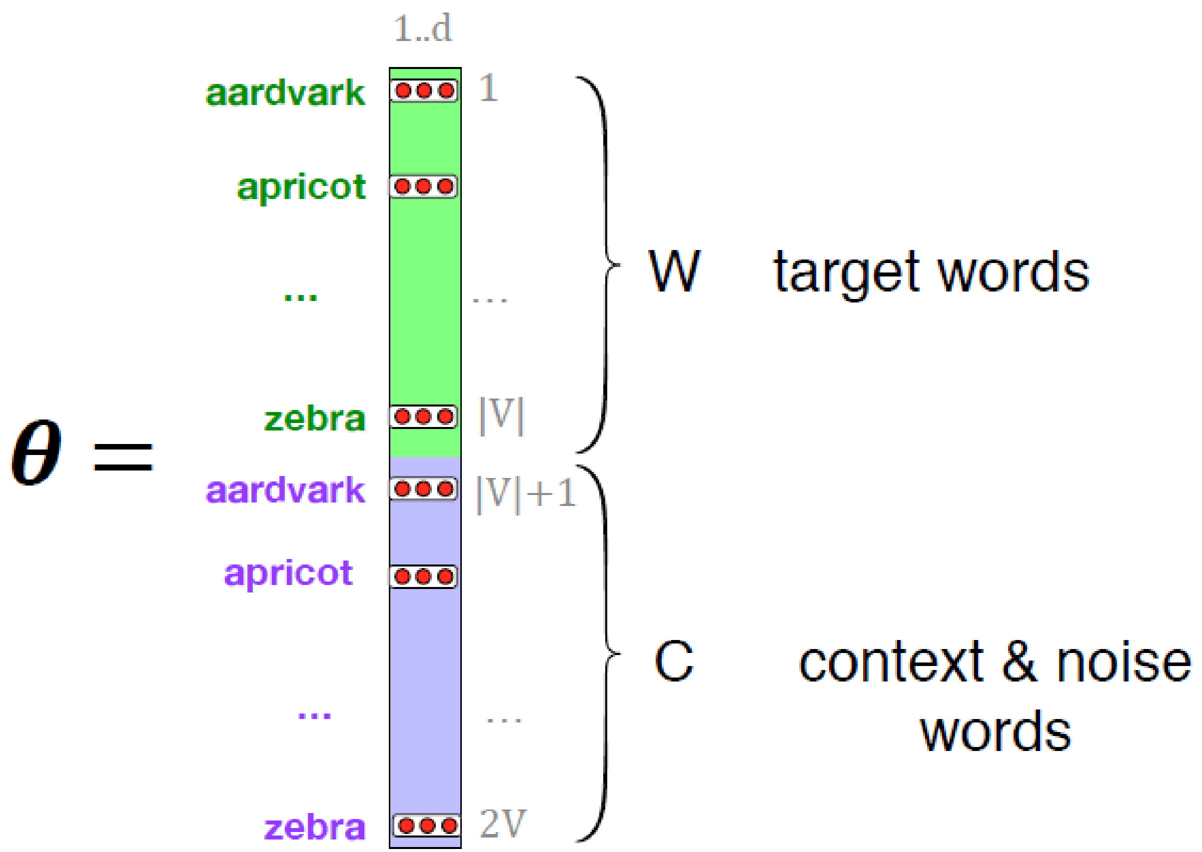
- Simplifying assumption that all context words are independent allows us to just multiple the probabilities for all of window L

$$P(+|w,c_{1:L}) = \prod_{i=1}^{L} \sigma(\ c_i \cdot w)$$

$$\log P(+|w,c_{1:L}) = \sum_{i=1}^{L} \log \sigma(\ c_i \cdot w)$$

Word2Vec

- Skip-gram model stores an target embedding matrix W for target words and an context embedding matrix C for context and noise words
- Embedding matrices have a dimension d whose size is found empirically



Word2Vec

- Learning skip-gram embeddings
- A context window L = 2 would provide 4 context words per target word. This yields 4 positive training examples
- We create negative context words by randomly sampling for corpus lexicon (minus the real context words) to provide k negative examples

- Example below with k = 2 (2 negative examples for 1 positive)

positive examples +

w	c_{pos}
apricot	tablespoon
apricot	of
apricot	jam
apricot	a

negative examples -

w	c_{neg}	w	c_{neg}
apricot	aardvark	apricot	seven
apricot	my	apricot	forever
apricot	where	apricot	dear
apricot	coaxial	apricot	if

Word2Vec

- To avoid a strong bias towards common words for noise an α weighted unigram sample frequency is used to select noise words

$$P_\alpha(w) = \frac{count(w)^\alpha}{\sum_{w'} count(w')^\alpha}$$

- Embeddings are learnt by minimizing a loss function using stochastic gradient decent

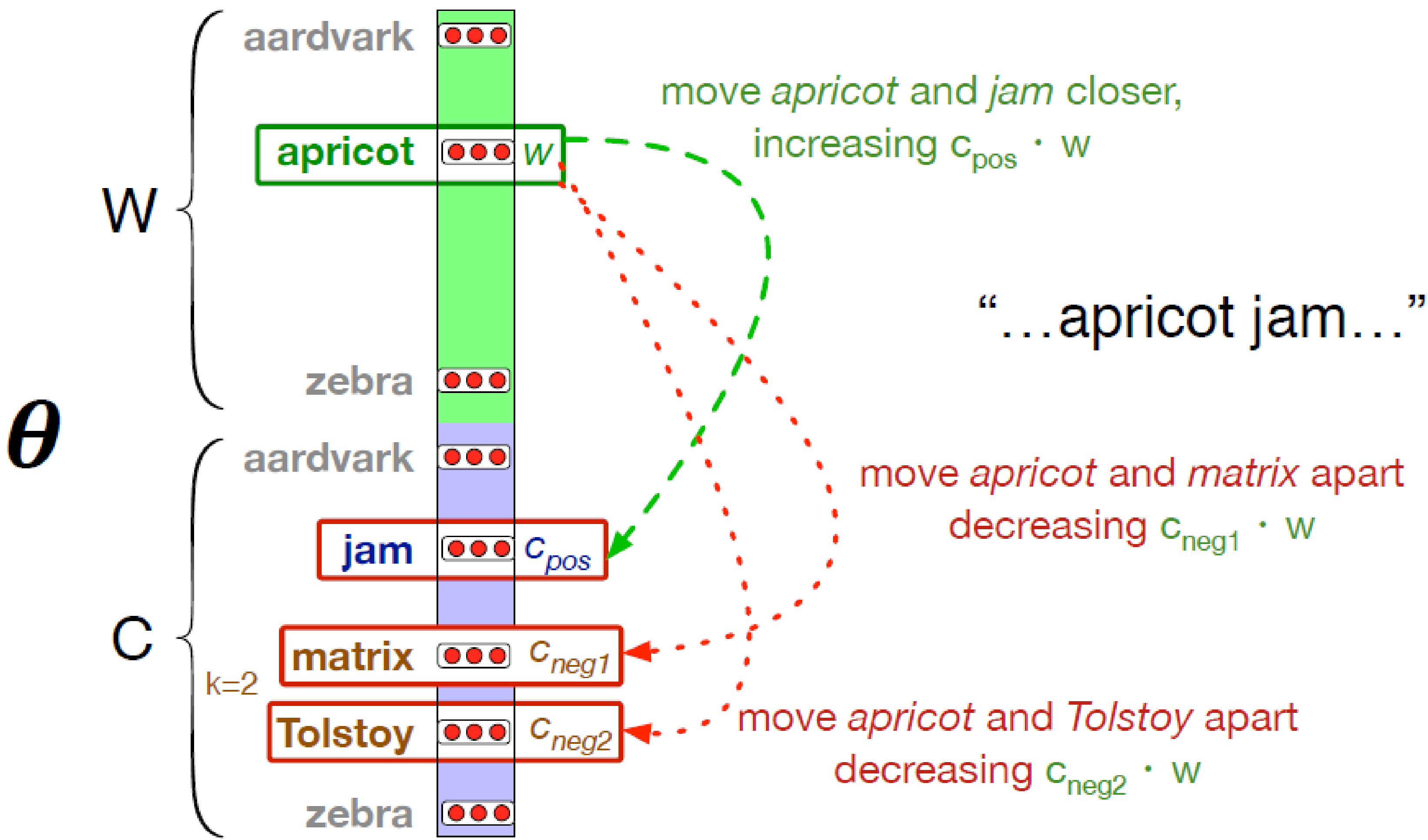
Word2Vec

- Given target w, positive context c_{pos}, (negative context c_{neg}) x k
- Loss function >> max(prob target is close to pos example) AND max(prob target is not close to neg example)

$$
\begin{aligned}
L_{CE} &= -\log\left[P(+|w,c_{pos})\prod_{i=1}^{k}P(-|w,c_{neg_i})\right] \\
&= -\left[\log P(+|w,c_{pos})+\sum_{i=1}^{k}\log P(-|w,c_{neg_i})\right] \\
&= -\left[\log P(+|w,c_{pos})+\sum_{i=1}^{k}\log\left(1-P(+|w,c_{neg_i})\right)\right] \\
&= -\left[\log\sigma(c_{pos}\cdot w)+\sum_{i=1}^{k}\log\sigma(-c_{neg_i}\cdot w)\right]
\end{aligned}
$$

Word2Vec

- Move 'apricot' weights closer to 'jam' and further from 'matrix'



Word2Vec

- Target matrix W and context matrix C are randomly initialized prior to training
- Word embedding matrix = W + C, or sometimes just W

Word2Vec

- There are other types of static embeddings
  - Fasttext uses subword models (character n-grams) to help overcome out of vocabulary and rare word problems
    - https://github.com/facebookresearch/fastText
    - https://fasttext.cc
    - Pre-trained models for Wikipedia and Web Crawl for 157 languages
  - Global Vectors (GloVe) uses ratios of probabilities from the word-word co-occurrence matrix applied to a large global corpus
    - https://nlp.stanford.edu/projects/glove/
    - Pre-trained models for Wikipedia, Gigaword (news), Common Crawl (web) and Twitter (social media)

Break

- Panopto Quiz - discussion point

- Which of these models use static embeddings?

Sentiment analysis model fine tuning pre-trained BERT word embeddings
Music recommender system using word embeddings loaded from a skip-gram model pre-trained on Spotify corpus
Legal document classifier using a sentence embedding pre-trained on Wikipedia
Legal document classifier using a sentence embedding pre-trained on Wikipedia and then fine tuned with legal text

Break

- Panopto Quiz - discussion point

- Which of these models use static embeddings?

Sentiment analysis model fine tuning pre-trained BERT word embeddings
Music recommender system using word embeddings loaded from a skip-gram model pre-trained on Spotify corpus
Legal document classifier using a sentence embedding pre-trained on Wikipedia
Legal document classifier using a sentence embedding pre-trained on Wikipedia and then fine tuned with legal text

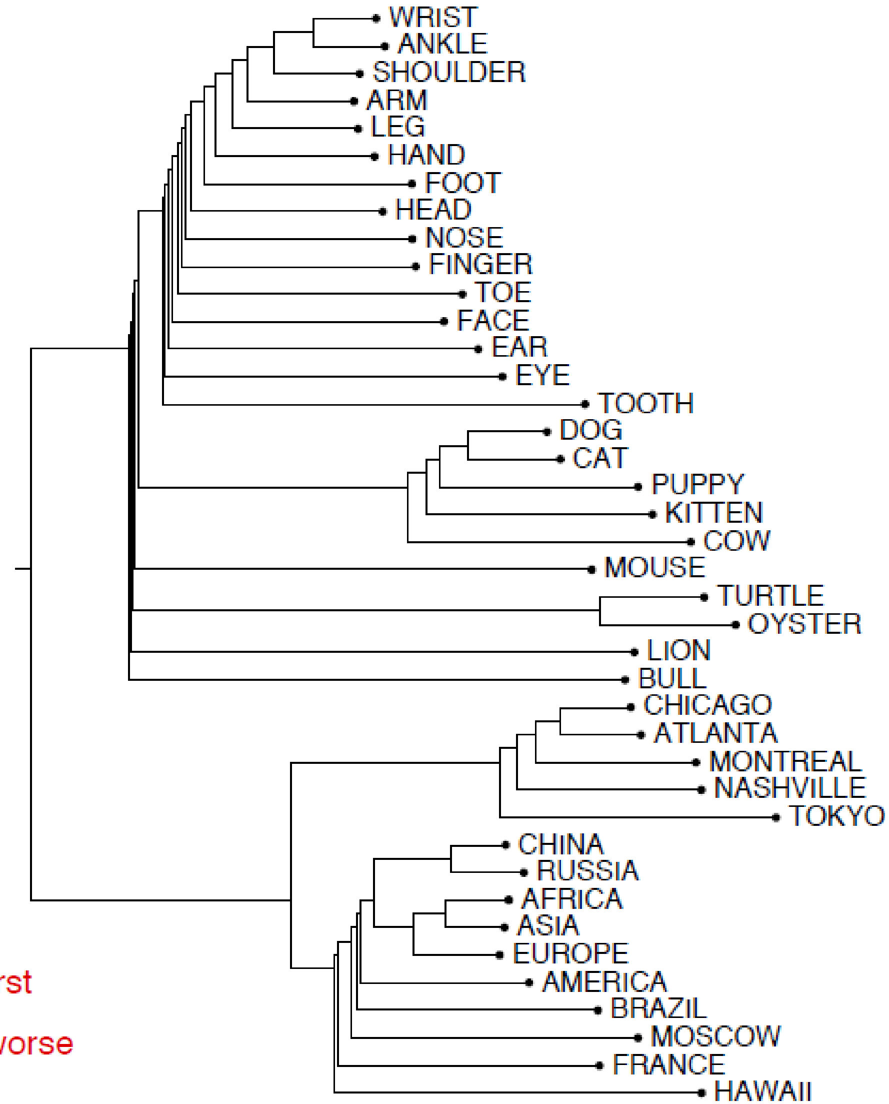
Static embeddings are pretrained and fixed, often using a model different from the target problem model (e.g. with a skip-gram model on Wikipedia corpus)

Contextual embeddings are usually the hidden layers trained (or 'fine tuned') for the target problem (contextually trained using a target domain corpus)

Static embeddings which are allowed to change will become contextual embeddings (e.g. fine tuning BERT embeddings on a specific problem corpus such as sentiment analysis)

Visualizing Embeddings

- Given a target word print top N closest words
- Hierarchical clustering of embedding space
- 2D projections of N dimensional embedding space



GloVe embeddings for frog

2D projection of embeddings trained for sentiment analysis

Semantic Properties of Embeddings

- Context window size L is based on desired goals
- Similarity
  - Small L (2-4) captures similar words (e.g. list of similar things)
- Association
  - Larger L (5+) captures longer distance topical relationships

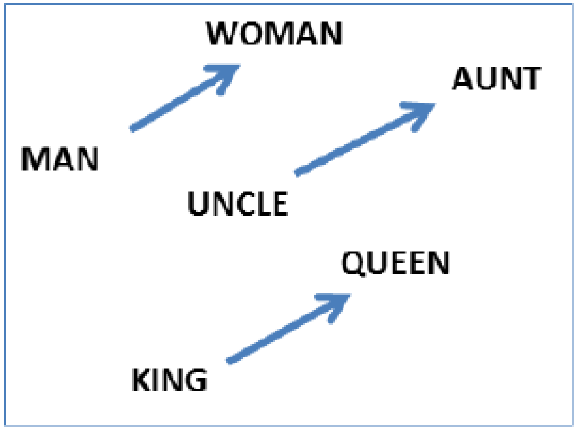
Word2Vec trained on Wikipedia

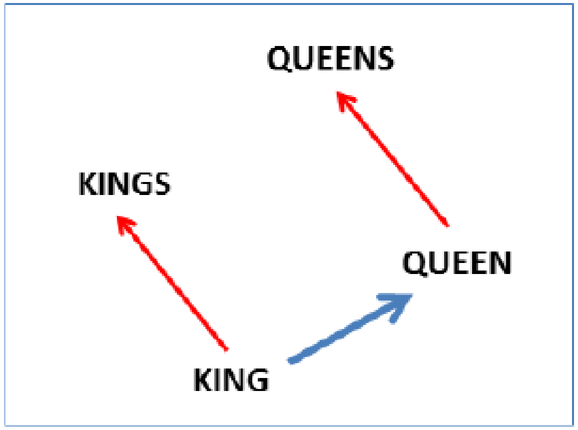
Target Word	BoW5	BoW2
batman	nightwing aquaman catwoman superman manhunter	superman superboy aquaman catwoman batgirl
hogwarts	dumbledore hallows half-blood malfoy snape	evernight sunnydale garderobe blandings collinwood
turing	nondeterministic non-deterministic computability deterministic finite-state	non-deterministic finite-state nondeterministic buchi primality
florida	gainesville fla jacksonville tampa lauderdale	fla alabama gainesville tallahassee texas
object-oriented	aspect-oriented smalltalk event-driven prolog domain-specific	aspect-oriented event-driven objective-c dataflow 4gl
dancing	singing dance dances dancers tap-dancing	singing dance dances breakdancing clowning
	L = 5	**L = 2**

Cite: Levy, O. Goldberg, Y. Dependency-Based Word Embeddings, ACL 2014

Semantic Properties of Embeddings

- Analogy
  - A is to B; as C is to ?
  - Compute vector offset of A -> B
  - Apply vector offset to C to discover ?





Single layer RNN trained on Broadcast News
(right) gender relation (left) single/plural relation

- Need to exclude morphological variants of target word
  e.g. cherry -> red; potato -> ~~potato~~|~~potatoes~~|brown
- Works best for frequent words with lots of examples and small L sizes

Cite: Mikolov, T. Yih, W. Zweig, G. Linguistic Regularities in Continuous Space Word Representations, NAACL 2013

Bias and Embeddings

- Embeddings encode the bias within the training sets used to compute them
- A historical news corpus will encode the bias of that time
  - Man -> Computer Programmer; Woman -> Homemaker
- Allocation harm
  - Biases in algorithms which result in unfair real world outcomes (e.g. load application pre-filtering algorithm)
- Bias amplification
  - Embeddings tend to exaggerate patterns making encoded bias more extreme than the original training resource
  - Implicit bias can be captured in embeddings and exaggerated (e.g. racial bias, ageism)

Cite: Bolukbasi, T. Chang, K. Zou, J. Saligrama, V. Kalai, A.T.
Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, NIPS 2016

Bias and Embeddings

- Representational harm
  - Harm caused by a system demeaning or even ignoring some social groups
- Debiasing
  - Manipulating embeddings to remove unwelcome stereotypes
  - May reduce bias, but will not eliminate it >> open problem
- Be mindful about bias in algorithms
  - Training data questions (historical? partial demographic? extreme views?)
  - Algorithms questions (patterns linking to stereotypes? uncertainty reported? unbiased ground truth available? explainable results?)
  - Check for bias
  - Mitigate bias if you can
  - Declare bias if you cannot
  - Decision makers using algorithms need to be aware of bias in results

Cite: Bolukbasi, T. Chang, K. Zou, J. Saligrama, V. Kalai, A.T.
Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings, NIPS 2016

Required Reading

- Vector Semantics and Embeddings
  - Jurafsky and Martin, Speech and Language Processing, 3rd edition (online) >> chapter 6

Questions

- Panopto Quiz - 1 minute brainstorm for interactive questions

  Please write down in Panopto quiz in **1 minute** two or three questions that you would like to have answered at the next interactive session.

  Do it **right now** while its fresh.

  Take a screen shot of your questions and **bring them with you** at the interactive session so you have something to ask.