L101: Machine Learning for Language Processing

Lecture 6

Guy Emerson

Today's Lecture

- Distributional semantics
  - Count vectors
  - Embedding vectors
- Challenges for distributional semantics

Distributional Semantics

… being hurt by another	horse	especially if some rider …
… beaten by a better	horse	at the distance on …
… these studies that	horses	reared with other …
… reared with other	horses	in a free and …
… ‘Is that all your	horse	gets to eat?’ in …
… cache of cattle and	horse	bones, while from the …
… was a sterling good	horse,	especially at Ascot …
… way as a domestic	horse	that it is stabled …
… 1790 – that is, one	horse	or two cows for …
… as coarse as a	horse	’s tail straying from …

Distributional Semantics

… being hurt by another	horse	especially if some rider …
… beaten by a better	horse	at the distance on …
… these studies that	horses	reared with other …
… reared with other	horses	in a free and …
… ‘Is that all your	horse	gets to eat?’ in …
… cache of cattle and	horse	bones, while from the …
… was a sterling good	horse,	especially at Ascot …
… way as a domestic	horse	that it is stabled …
… 1790 – that is, one	horse	or two cows for …
… as coarse as a	horse	’s tail straying from …

Distributional Semantics

… being hurt by another	horse	especially if some rider …
… beaten by a better	horse	at the distance on …
… these studies that	horses	reared with other …
… reared with other	horses	in a free and …
… ‘Is that all your	horse	gets to eat?’ in …
… cache of cattle and	horse	bones, while from the …
… was a sterling good	horse,	especially at Ascot …
… way as a domestic	horse	that it is stabled …
… 1790 – that is, one	horse	or two cows for …
… as coarse as a	horse	’s tail straying from …

Distributional Semantics

… being hurt by another	horse	especially if some rider …
… beaten by a better	horse	at the distance on …
… these studies that	horses	reared with other …
… reared with other	horses	in a free and …
… ‘Is that all your	horse	gets to eat?’ in …
… cache of cattle and	horse	bones, while from the …
… was a sterling good	horse,	especially at Ascot …
… way as a domestic	horse	that it is stabled …
… 1790 – that is, one	horse	or two cows for …
… as coarse as a	horse	’s tail straying from …

Distributional Semantics

- Linguistic motivation:
  understand language
  - Harris (1954)
  - Firth (1951, 1957)
- Machine learning motivation:
  text is cheap

Context

ASH-993: … saying ‘Is that all your horse gets to eat?’ in amazement …

- Word windows (saying, your, eat, …)
- Dependencies (your-POSS, get-SUBJ)
- Documents (ASH-993)

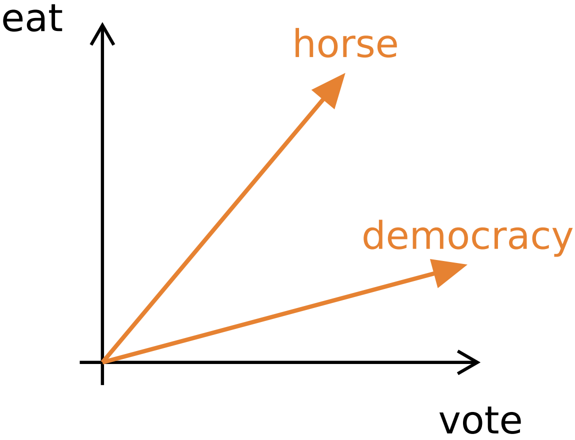
Word Window Hyperparameters

- Window size
- Lemmatisation?
- Stop list?
- Rare words?

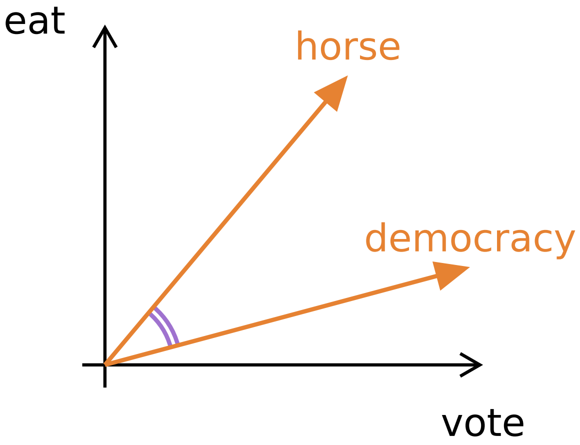
Count Matrix

$$
\begin{array}{cc}
 & \text{contexts} \\
\text{target words} &
\begin{pmatrix}
n_{11} & n_{12} & n_{13} & \dots & n_{1D} \\
n_{21} & n_{22} & n_{23} & \dots & n_{2D} \\
n_{31} & n_{32} & n_{33} & \dots & n_{3D} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
n_{V1} & n_{V2} & n_{V3} & \dots & n_{VD}
\end{pmatrix}
\end{array}
$$

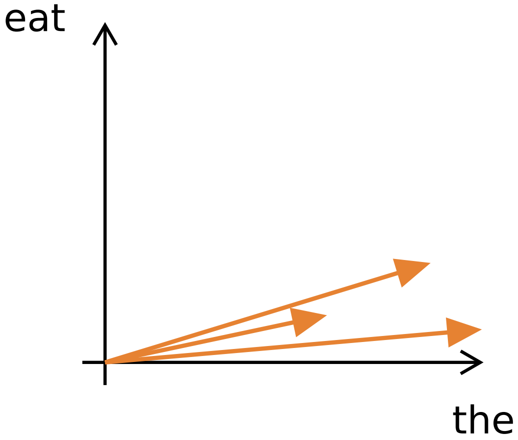
Count Vectors



Count Vectors



Count Vectors



Processing the Counts – TF-IDF

$$v_{ij} = \frac{n_{ij}}{\left|\left\{i' : n_{i'j} > 0\right\}\right|}$$

- Used in document retrieval
- TF: term frequency
- IDF: inverse document frequency

Processing the Counts – PMI

$$v_{ij} = \log \frac{n_{ij} n_{..}}{n_{i.} n_{.j}}$$

- Pointwise Mutual Information (from information theory)
- $\log \dfrac{P(x,y)}{P(x)P(y)}$

Processing the Counts – PMI

$$v_{ij} = \log \frac{n_{ij} n_{..}}{n_{i.} n_{.j}}$$

- Pointwise Mutual Information (from information theory)
- $\log \frac{P(x,y)}{P(x)P(y)}$ – more likely than expected?

Processing the Counts – PPMI

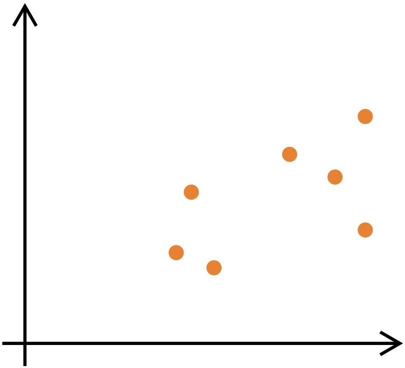
$$v_{ij} = \max\left\{0, \log\frac{n_{ij}n_{..}}{n_{i.}n_{.j}}\right\}$$

- Pointwise Mutual Information (from information theory)
- $\log\dfrac{P(x,y)}{P(x)P(y)}$, positive only
- Avoids negative infinities

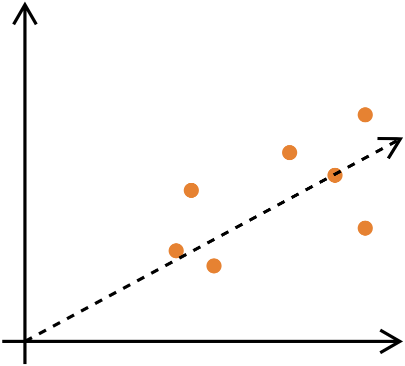
Singular Value Decomposition

- High dimensions difficult to work with
- Find directions of highest variance
- Only use these directions

Singular Value Decomposition



Singular Value Decomposition



Embedding Vectors

- Directly learn lower-dimensional vectors
- Never construct count matrix

Skip-Gram

- Observe target-context pairs (t, c)

Skip-Gram

- Observe target-context pairs (t, c)
- Treat as classification:
  predict context, given target

Skip-Gram

- Observe target-context pairs (t, c)
- Treat as classification:
  predict context, given target
- Discriminative classifier:
  $P(c \mid t) \propto \exp(v_t \cdot u_c)$

Skip-Gram

- Observe target-context pairs (t, c)
- Treat as classification:
  predict context, given target
- Discriminative classifier:
  $P(c \mid t) \propto \exp(v_t \cdot u_c)$
- Like logistic regression, but:
  “input vectors” are learnt, not given

Skip-Gram & Negative Sampling

- $P(c \mid t) \propto \exp(v_t \cdot u_c)$
  requires *all* possible contexts

Skip-Gram & Negative Sampling

- $P(c \mid t) \propto \exp(v_t \cdot u_c)$
  requires *all* possible contexts
- Instead, sample a few other contexts c'

Skip-Gram & Negative Sampling

- $P(c\,|\,t) \propto \exp(v_t \cdot u_c)$
  requires *all* possible contexts
- Instead, sample a few other contexts c'
- Treat as binary classification:
  predict if context is real or sampled

Skip-Gram & Negative Sampling

- $P(\text{real} \mid t, c) \propto \exp(v_t \cdot u_c)$
- $P(\text{sampled} \mid t, c) \propto 1$

Skip-Gram & Negative Sampling

- $P(\text{real} \mid t, c) \propto \exp(v_t \cdot u_c)$
- $P(\text{sampled} \mid t, c) \propto 1$
- $P(\text{real} \mid t, c) = \sigma(v_t \cdot u_c) = \frac{1}{1+\exp(-v_t \cdot u_c)}$

Skip-Gram & Negative Sampling

- Want high: $v_t \cdot u_c$
- Want low: $v_t \cdot u_{c'}$

Count vs. Embedding

- Skip-gram approximately factorises a PMI matrix!

Count vs. Embedding

- Skip-gram approximately factorises a PMI matrix!
- Hyperparameters important

Evaluation

- Lexical semantics
- Compositional semantics
- Downstream tasks

Lexical Semantics

democracy	*water*	*happiness*
aubergine	*flood*	*joy*
computer	*law*	*cat*
earthquake	*lawyer*	*dog*

Lexical Semantics

- Give annotators pairs of words
- Ask to score (e.g. from 1 to 7)

Lexical Semantics

- Give annotators pairs of words
- Ask to score (e.g. from 1 to 7)
- Get system's similarity scores
- Measure Spearman rank correlation

Challenges for Dist. Sem.

- Grounding
- Lexical Semantics
  - Word senses
  - Hyponymy
- Sentence Semantics
  - Composition
  - Logic

Word Senses

… the last kick of the	match.	It was entertaining …
… the Duddon are no	match,	after all, for a route …
… first or second round	matches	of any consequence …
… Tried soaking the	matches	in paint, he wrote, …
… is very much a	match	for Berowne; this is …
… to win and the	match	is therefore …
… to lose you the	match	even though no …
… of an elimination	match	is fought. If this …
… needed to watch the	match,	needed a diversion …
… drop in a burning	match.	The plastic of the …

Word Senses

… the last kick of the	match.	It was entertaining …
… the Duddon are no	match,	after all, for a route …
… first or second round	matches	of any consequence …
… Tried soaking the	matches	in paint, he wrote, …
… is very much a	match	for Berowne; this is …
… to win and the	match	is therefore …
… to lose you the	match	even though no …
… of an elimination	match	is fought. If this …
… needed to watch the	match,	needed a diversion …
… drop in a burning	match.	The plastic of the …

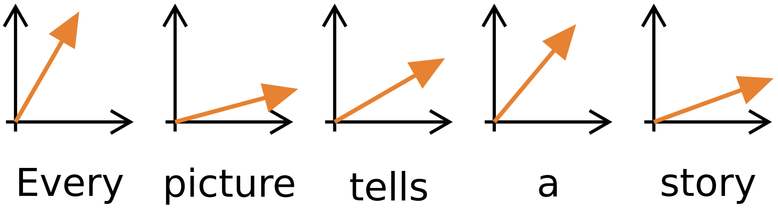
Word Senses

… the last kick of the	match.	It was entertaining …
… the Duddon are no	match,	after all, for a route …
… first or second round	matches	of any consequence …
… Tried soaking the	matches	in paint, he wrote, …
… is very much a	match	for Berowne; this is …
… to win and the	match	is therefore …
… to lose you the	match	even though no …
… of an elimination	match	is fought. If this …
… needed to watch the	match,	needed a diversion …
… drop in a burning	match.	The plastic of the …

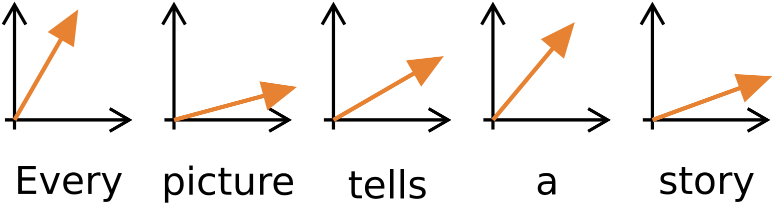
Word Senses

… the last kick of the	match.	It was entertaining …
… the Duddon are no	match,	after all, for a route …
… first or second round	matches	of any consequence …
… Tried soaking the	matches	in paint, he wrote, …
… is very much a	match	for Berowne; this is …
… to win and the	match	is therefore …
… to lose you the	match	even though no …
… of an elimination	match	is fought. If this …
… needed to watch the	match,	needed a diversion …
… drop in a burning	match.	The plastic of the …

Semantic Composition

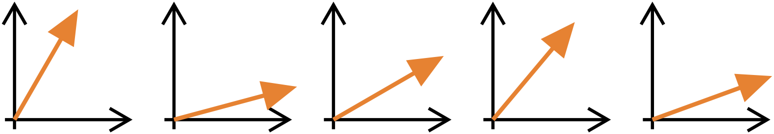


Semantic Composition



- Addition?
- Componentwise multiplication?

Semantic Composition



Every picture tells a story

- Addition?
- Componentwise multiplication?
- Linguistically-motivated approach?

Summary

- Distributional semantics – context
- Count models
  - PPMI, SVD
- Embedding models
  - Skip-gram with negative sampling
- Similarity and relatedness
- Challenges – word senses, composition