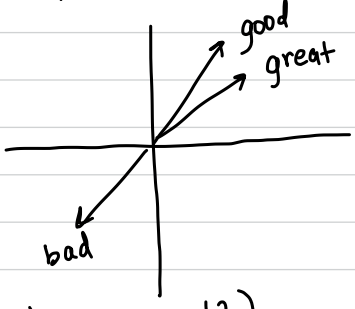
10/13

- Word Embeddings
- one-hot vectors

$S = [1 \ldots 1 \ldots 1]$
good great bad

↳ good/great are as dissimilar as good/bad.

Pre-processing task:
- Turn words into low dimensional vectors (50-300 vals)
- Cosine similarity between vectors represents how similar they are : $\cos\theta = \frac{A \cdot B}{\|A\| \|B\|}$



Ex: good = [1.2 0.8 -0.5 ...]

- word2vec (Mikolov 2013)
  - ↳ Trained w/ continuous bag of words assumption / skipgram
    - ↳ skipgram weights nearby context words more heavily
  - ↳ Model = shallow, 2-layer neural net
  - ↳ Model doesn't matter that much, mostly hyperparameter tuning :
    - context window : 10 for skipgram, 5 for CBOW
    - Dimensionality : diminishing returns
    - sample high-freq words less (threshold)

(Pennington 2014)

GloVe : global vectors for word repr.
- uses global co-occurences instead of local "context" windows

Semantic relationship:
King = Queen - Woman + Man
GloVe > word2vec

Syntactic relationship:
Walking = Sleeping - Sleep + Walk
word2vec > GloVe

no formal explanation for why this works

BERT (2019) : Bidirectional Encoder Representations From Transformers

- Attention model: different word embeddings for different contexts
  "I went to the bank to deposit a check."
  "I had a pichic at the river bank"
- Implemented bidirectionality for the first time
- 93.2% F1 (accuracy measure) on reading comprehension task vs. prev. 91.6% record.
  ↳ Humans: 91.2%
- Bidirectionality: learn left to right, right to left
  ↳ Joint condition on left/right context
- Trained on Wikipedia
- Transformers: allow parallel processing to learn long-term relationships

- can be fine tuned for specific tasks

10/13

- Bag of words

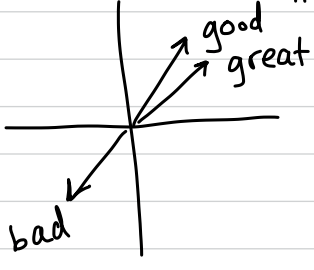
$S = [\ 1 \ \ldots \ 1 \ \ldots \ 1\ ]$
good great bad

good / great are as dissimilar good / bad

Word Embeddings

- words → 50 - 300 value vector
- Preprocessing task
- Cosine similarity between vectors represents how similar they are

$$\cos\theta = \frac{\bar{a} \cdot \bar{b}}{\|a\| \|b\|}$$



- word2vec (Mikolov 2013)
  - ↳ 2-layer neural net
  - ↳ continous bag of words assumption / skipgram

"the movie was great"

context window k=1

- Dimensionality : length of vectors
  diminishing returns : 300+
- Sample high-frequency words less : threshold

GloVe : global vectors for word representation
(Pennington 2014)
- uses global co-occurances instead "local" context windows

Dimensionality Hypothesis ?

JR Firth: "You can tell what a word is like based on the company it keeps"

Semantic Relationships : Meaning

GloVe > word2vec

king = Queen − Woman + Man

Syntactic Relationships

word2vec > GloVe

Walking = Sleeping − Sleep + Walk

BERT : Bidirectional Encoder Representations From Transformers (2019)

- Attention Model: different word embeddings for the same word in different contexts

"I accessed the bank account"

"I went to the river bank"

- Implented bidirectionality for the first
  → learning left to right, right to left
- Trained on wikipedia
- Transformers: allow parallel processing to learn long-term relationships between words
- Reading comprehension
  - 93.2% F1 (accuracy)
  - 91.6%
  - 91.2%
- Fine tune for specific task