Sources

Distributional Semantics <https://www.cs.utexas.edu/users/mooney/cs388/slides/dist-sem-intro-NLP-class-UT.pdf>:

* Distributional hypothesis: meaning of a word is the set of contexts in which it appears
  + Important aspects of the meaning can be approximated by the set of contexts in which it appears
* Weak and strong version of hypothesis:
  + Weak: Quantitative method for semantic analysis/lexical resource induction
  + Strong: Cognitive hypothesis about the form and origin of semantic representation
* Distributional semantic models (DSM)
  + Meaning of words normally represented as vectors recording the word’s distributional history
  + General purpose models that don’t require supervision
  + Usually scalable, flexible, very empirical
  + Literature focuses on doing this with the goal of computing word-to-word similarity
* Steps to build DSMs
  + Process source (tokenization, linguistic tagging, stemming, stopword removal, other preprocessing)
    - Disadvantages to deep linguistic tagging: language dependent, errors at each stage, more parameters to tune
    - Computational linguistics tends to use linguistic tagging, but no clear evidence of their advantage
    - Cognitive science instead tends to use broader document/text-based contexts
  + Collect co-occurrence matrix of words
    - Each row represents a distributional vector
    - Each column represents the elements of the context
    - Element contains the number of times word appears within the context of
  + Transform the matrix
    - Attach weights to the raw frequency counts
      * Positive point-wise mutual information robust and widely used
      * Other weighings: mutual information, log likelihood ratio, tf-idf
    - Dimensionality reduction (so an matrix turns into an matrix where )
  + Computing word similarities
    - Traditionally the goal even though it’s not for us
    - Compute distances between two words by getting the cosine of their distributional vectors (commonly accepted)
* Definitions of ‘context’
  + Document ID
  + Different ‘window’ length (entire document, a sentence, 2 words)
  + Different type of words (all words, non-stopwords only, words with a syntactic path to the target only example on slide 23)
  + Different types of encoded attributes to words in context (part of speech, syntactic path to target)
  + All non-very common words in sentence (a, the, beyond, although, other)

<http://www.aclweb.org/anthology/W09-0214>

* Context vectors
  + Sum of the distributional vectors of all the words within x window of the target
    - Unique context vector for each occurrence of a target word
    - Subtract overall document distributional vector so document style doesn’t bias context vectors from that document
  + Assumed to encode the aggregate meaning of the context (so used for word sense disambiguation)
  + Compared in the same manner as distributional vectors
    - Subtract from the cosine distance of two vectors the cosine distance of the documents in which they appear (so contexts don’t vary by document topics)
  + Can be used to detect usage variety (by computing average distance between contexts)
    - With lots of appearances, computing average distances between all contexts can result in lots of comparisons, so a monte carlo-based estimation approach was used by the authors for frequent words
* Paper used context vectors to detect specific categories of semantic change (broadening, narrowing, pejoration)