Semantic Analysis

Word sense

Lexeme → set of words, same fundamental meaning (run, runs, ran → lexeme RUN)

Lemma → Lexeme you'd put in a lexicon

One lemma → multiple lexemes (word senses)

Homonymy

Same pronunciation/spelling, different meaning

Polysemy

Two senses of a lemma are semantically linked.

Synonymy

When two senses of two different lemmas are (nearly) identical.

If can substitute word A with word B without changing the meaning of the sentence, A & B are synonymous.

Antonymy

Opposite of synonyms. A & B are opposites.

Hyponymy

More specific (car → vehicle, mango → fruit).

Hyponym is the lower word in the word tree.

Hypernymy

Less specific (furniture → chair, fruit → mango).

Hypernym is the upper word in the word tree.

WordNet

Website. Three databases: nouns, verbs, adjectives + adverbs.

Each lemma has synset, a set of one or more senses.

Simplified Lesk algorithm

Given word in a context + number of senses for word.

Textual overlap of non-stopwords between context and sense → score of sense.

Word similarity

How similar is word A to word B? Synonym is boolean relation, want numeric representation.

Distributional hypothesis

Distance between two word senses by finding words with similar distributions in a corpus.

Represent words as vectors.

Co-occurrence matrix

		context words						
		crown	throne	reign	Sweden	match	goal	play
target words	queen	4	1	1	2	0	0	0
	king	3	2	1	3	1	0	0
	soccer	1	0	0	4	3	4	2
	hockey	0	1	0	1	2	1	1

Simulate the Simplified Lesk algorithm

Count non-stopword similarities between context and senses, take highest count.

Compute the path length-based similarity of two words

similarity = (word1, word2) → return 1 / (1 + pathlength(word1, word2));

pathlength = number of edges in shortest path between word1 and word2.

pathlength = Basically count the number of words you meet along the way minus the original word.

Derive a co-occurrence matrix from a document collection

Each cell → number of documents in which target word (row) co-occurs with context word (col).