

Practical Natural Language Processing

Kan Min-Yen
National University of Singapore

Teaching staff

- **Lecturer:**

Min-Yen Kan (“Min”)

kanmy@comp.nus.edu.sg

Office: AS6 05-12

++65 6516-1885

Hobbies:

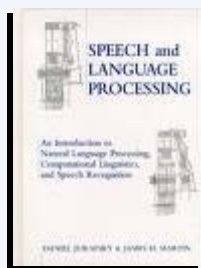
rock climbing, ballroom dancing, and inline skating...



Lost in Hakodate, Japan

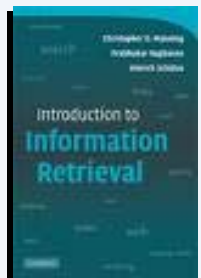
Textbooks Used

- **J&M – Jurafsky and Martin**



- Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition
- Colorado + other contributors

- **MRS – Manning, Raghavan, Schütze**



- Introduction to Information Retrieval
- Stanford and Yahoo!
- Whole book (.PDF) available from authors website:
- <http://www-csli.stanford.edu/~hinrich/information-retrieval-book.html>

Course Outline

Day 1

AM

- Applications' Input / Output
- Resources

PM

- Selected Toolkits
- Python Intro
- NLTK Hands-on

Day 2

AM

- Evaluation
- Annotation
- Information Retrieval
- ML Intro

PM

- Machine Learning
- SVM Hands-on

Day 3

AM

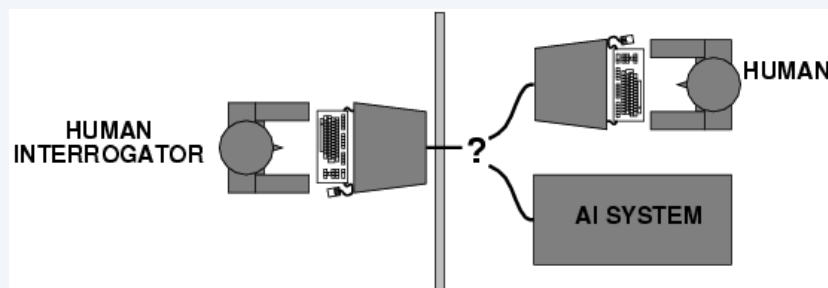
- Sequence Labeling
- CRF++ Hands-on

PM

- Dimensionality Reduction
- Clustering
- Trends & Issues

Acting humanly: Turing Test

- Turing (1950) "Computing machinery and intelligence":
- "Can machines think?" → "Can machines behave intelligently?"
- Operational test for intelligent behavior: the Imitation Game



Credits: wikipedia

- Predicted that by 2000, a machine might have a 30% chance of fooling a lay person for 5 minutes
- Anticipated all major arguments against AI in following 50 years
- Suggested major components of AI: knowledge, reasoning, language understanding, learning

NLP from an academic POV: communication

- **Communication**

- **Intentional** exchange of information brought about by the production and perception of signs drawn from a shared system of conventional signs

- **Humans use language to communicate most of what is known about the world**

- **Communication as Action**

- **Speech act**

- Language production viewed as an action

- **Speaker, hearer, utterance**

- **Examples:**

- Query: “Who’s going to be elected president in November?”

- Inform: “I’m teaching a course offsite today.”

- Request: “Please help me make 10 copies.” “I could use some help with photocopying.”

- Acknowledge: “OK”

- Promise: “I’ll be there by 9:30 a.m.”

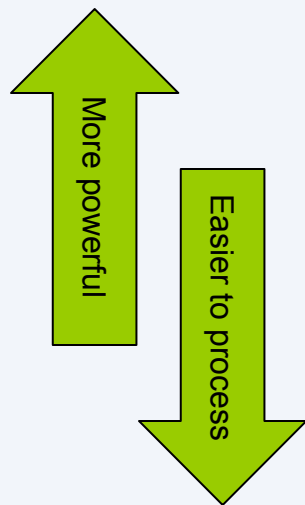
Fundamentals of Language

- **Formal language:** A (possibly infinite) set of strings
- **Grammar:** A finite set of rules that specifies a language
- **Rewrite rules**
 - nonterminal symbols (S, NP, etc)
 - terminal symbols (he)
 - $S \rightarrow NP VP$
 - $NP \rightarrow \text{Pronoun}$
 - $\text{Pronoun} \rightarrow \text{he}$



Chomsky Hierarchy

Four classes of grammatical formalisms:



- **Recursively enumerable grammars**

- Unrestricted rules: both sides of the rewrite rules can have any number of terminal and nonterminal symbols

$$AB \rightarrow C$$

- **Context-sensitive grammars**

- The RHS must contain at least as many symbols as the LHS

$$ASB \rightarrow AXB$$

- **Context-free grammars (CFG)**

- LHS is a single nonterminal symbol

$$S \rightarrow XYa$$

- **Regular grammars**

$$X \rightarrow a$$

$$X \rightarrow aY$$



Component Steps of Communication

SPEAKER:

- **Intention**
 - $\text{Know}(H, \neg \text{Alive}(\text{Wumpus}, S3))$
- **Generation**
 - “The wumpus is dead”
- **Synthesis**
 - $[\text{thaxwahmpaxsihzdehd}]$



Component Steps of Communication

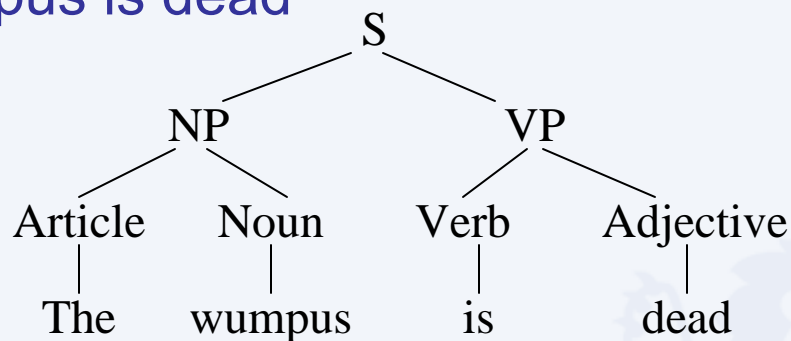
HEARER:

- Perception:

- “The wumpus is dead”

- Analysis

- (Parsing):



- (Semantic Interpretation):

- Alive(Wumpus, Now)

- Tired(Wumpus, Now)

- (Pragmatic Interpretation):

- Alive(Wumpus1, S3)

- Tired(Wumpus1, S3)

Component Steps of Communication

HEARER:

- **Disambiguation:**

¬Alive(Wumpus1,S3)

- **Incorporation:**

TELL(KB, ¬Alive(Wumpus1,S3))



Not so great newspaper headlines

- Squad helps dog bite victim.
- Helicopter powered by human flies
- Portable toilet bombed; police have nothing to go on.
- British left waffles on Falkland Islands.
- Teacher strikes idle kids.



Ambiguity!

Core issue in many fields of AI

Ambiguity in every level of NLP.

Can you think of some examples?

- Words -
- Syntax -
- Semantics -
- Pragmatics -

“One morning I shot an elephant in my pajamas
How he got into my pajamas I don’t know”

-- Groucho Marx, Animal Crackers 1930


- **Skewness in the ambiguity**
(DeRose 88, J&M pp 299)

Unambiguous	1 tag	35340
Ambiguous	2 tags	4100
	3 tags	3760
	4 tags	61
	5 tags	12
	6 tags	2
	7 tags	1 (“still”)

Approaches to Solving NLP problems

- **Rule Based (Symbolic)**

- Developed like traditional expert systems: hand coded rules
- Pro: fast to develop, doesn't require large datasets
- Con: fragile, costly to maintain



Typical of
resource-poor
languages

- **Statistics Based (Empirical)**

- Annotate data based on **standard** tagsets, then machine learn a model
- Pro: current trend, robust, performs better
- Con: extensive up front cost, requires lots of data, improvement may not correct obvious errors

- **Hybrid systems**

- Often blend rule-based pre- and post-processing with ML core

- **Human Intuition**

- plays a large role in both, either in coding the rules directly or in deciding what features to use
- can be driven by error analysis

Natural Language Processing – Back to you

What is NLP in your context?

How is it related to Information Retrieval?

How is it related to Machine Learning?

How is it related to your customers?



Whirlwind Application Tour

Applications – Input and Output

- **Words**

- Morphological Processing
- Spelling Correction
- Word segmentation
- Language Identification

- **Syntax**

- POS Tagging
- Parsing

- **Semantics**

- Word Sense Disambiguation
- Named Entity Recognition
- Semantic Role Labeling

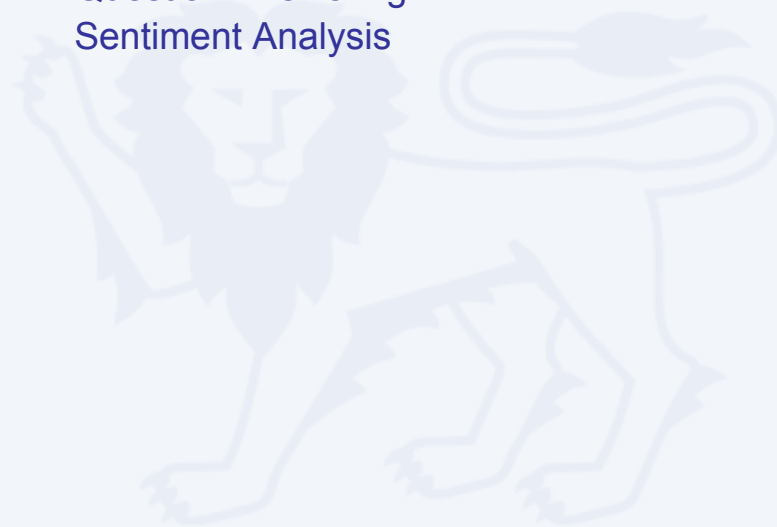
- **Pragmatics**

- Reference Resolution
- Generation*

- **Applications**

- Information Extraction
- Summarization
- Machine Translation

- Information Retrieval
 - Genre Analysis
 - Question Answering
 - Sentiment Analysis



Morphological Processing – Ch 3 J&M

- **Input** : Given a set of words (sentence)
- **Output** : Decide the stems (lemmas), prefixes and suffixes
- **Inflectional** – syntactic function such as agreement “prices soared”
- **Derivational** – change the class of the word “derive → derivational”
- **Used in stemming packages for conflating related words**
- **Morphotactics** – model of morpheme ordering
- **Solve with**
 - Orthographic Rules – how to combine morphemes
 - Finite State Transducers (FST)

Spelling Correction – Ch 5 J&M

- **Input:** Uncorrected sentence / words
- **Output:** Corrected words (in context?)
- **Malapropisms:** words correctly spelled but incorrectly used
- **Solve with**
 - Edit distance for operations
 - Incorporate corpus frequency
 - Hidden Markov Model (HMM) to deal with context



Word segmentation – Ch 5, pp 180-4 J&M

- **Input:** Given a sentence
- **Output:** Decide where the words are

日本章鱼怎么说?

1. 日_(day) 本章_(essay) 鱼_(fish) 怎么_(how) 说_(say)?
2. 日本_(japanese) 章鱼_(octopus) 怎么_(how) 说_(say)?

- **More prevalent than you might think:**
 - Multiword expressions (MWE) “make a call”, “push off”, “don’t”
“as and when”, “in terms of”
- **Solve with:**
 - Sequence Labeling – Hidden Markov Model
 - Be aware of multiple coding points or encoding standards (e.g. for Chinese characters)
 - Both dictionary and context as features

What about doubled words
in Malay?
rumah-rumah (houses)

Language Identification

- **Input:** Document or segmented document
- **Output:** detected language of each segment or document
- **Code switching:** changing languages within document
- **Considered a **solved** problem with a few sentences of text**
- **Solve with:**
 - Encoding
 - Character n-grams as vectors and cosine similar
 - Can sometimes check for genre, dialect of text as well

POS Tagging – Ch 8 J&M

- **Input:** Segmented word sequence
- **Output:** Syntactically-labeled word sequence

NN	Noun, sing. or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PP	Personal Pronoun
PP\$	Possessive Pronoun

- Inventory of tags: coarse or fine-grained?
- Solve with:
 - Rule-Based (pp. 302)
 - Transformational based learning (TBL; pp 308-09),
 - Sequence labeling

Parsing – Ch 9-12 J&M

- **Input:** Labeled word sequence
- **Output:** Sentence structure in some form

- **Issues: Long Distance Dependencies (*cf* morphology)**
- **Too much ambiguity, must constrain processing**
- **Solve with:**
 - Context Free Grammars (CFG)
 - Constituent / Phrase-Structure Parse – relations between constituents
 - Dependency Parse – relations between words

Parsing (cont'd)

- **Earley vs. CYK vs. GHR**

- Earley (**Deterministic** Chart)
- CYK (**Probabilistic** CFG)
- Collins (**Lexicalized** Probabilistic CFG)

- **Considerations:**

- Relationship to programming language compilation:
 - Shift reduce parsers for context free regular languages
 - Is natural language context free?
- Dependency parse: for free word order since constituency doesn't really matter

Named Entity Recognition

- **Input:** Text
- **Output:** Labeled text spans

- **Related to Parsing via: Chunking, Shallow Parsing**
- **Solve with:**
 - Don't use parsing (poly-time); opt for linear time complexity
 - FASTUS Cascade (pp 580)
 - Sequence labeler



Word Sense Disambiguation

- **Input:** Word sequence
- **Output:** Sense marked word sequence

- **Issues:**
 - Homonymy, polysemy, synonymy (Ch 16.1)
 - Not covered: creativity (metaphor, metonymy, Ch 16.4)

- **Solve by:**
 - Context, selectional restrictions
 - Machine learning
 - Heuristics – “One sense per collocation”
 - Bootstrapping a labeled corpus



Word Sense Disambiguation (cont')

- **Considerations:**

- Relationship to conflation – dimensionality reduction
- WSD benchmark tasks
- Discrepancy between WSD of words varies highly
- Depends on set of words: All words, set of words



Semantic Role Labeling – Sec 16.3 J&M

- **Input:** Sentence
- **Output:** Thematic roles to phrases within sentence
- **Issues:**
 - Used on top of (mostly) constituent parsing, chunking
 - Related to WSD in problem scope, dependency parsing
 - Alternations, Selectional restrictions
- **Solve with:**
 - ML on annotated data

Reference Resolution – Ch 18 J&M

- **Input:** Discourse
- **Output:** Reference resolved discourse
- **Issues**
 - Anaphora: indefinite, definite, pronouns
 - Centering (*cf* discourse)
 - Pleonastic uses (It is raining)
 - Coherence vs. cohesion (*cf* MIT fake conference submissions)
- **Solve with:**
 - ML on annotated and processed data

Generation – Ch 20 J&M

- **Input:** Facts in some symbolic form (logical form) + intention
- **Output:** Natural Language Output
- **Considerations:**
 - Related to machine translation
 - An entire pipeline, with many levels of processing
 - Used in description generation for museums, personalized course instruction
- **Solve with:**
 - **Surface / Sentential level:**
Functional Unification Grammar (FUG), Forest based scoring (PCFG based) with ML backbone
 - **Discourse level:** RST, Centering

Summarization

- **Input:** A text
- **Output:** A shorter version of the input text
- **Issues:**
 - Multi vs. single, is an update?, Query vs. generic, indicative vs. informative.
 - Ordering, Cohensiveness, Content, Fluency (repairs)
 - End application or use
- **Solve with:**
 - Sentence selection (view as selection or ranking problem)
 - Discourse motivated repairs (*cf* generation)

Information Extraction – Sec 15.5 J&M

- **Input:** tagged, parsed, NER text
- **Output:** relationships between NEs, factual tuples suitable for ingestion in a database

- **Issues:**
 - usually needs domain specific information
 - requires NER as NEs participate in roles

- **Solve with:**
 - Heuristic systems
 - Machine Learning with heuristic features

Machine Translation – Ch 21 J&M

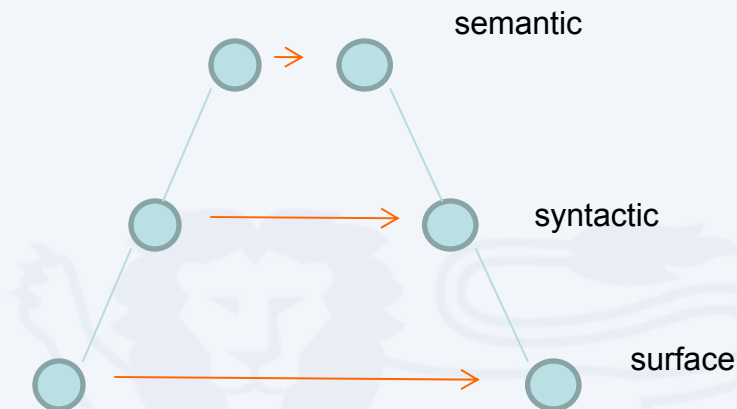
- **Input:** input sentence in source language (e)
- **Output:** output sentence in target language (f)

- **Architectures**

- Interlingua – *cf* generation
- Parsing, transfer, generation
- Direct - SMT

- **Solve by:**

- Large corpora for English
- Example Based MT (memoization for some constituents?)
- Transformation Based Learning (TBL, see Tagging)



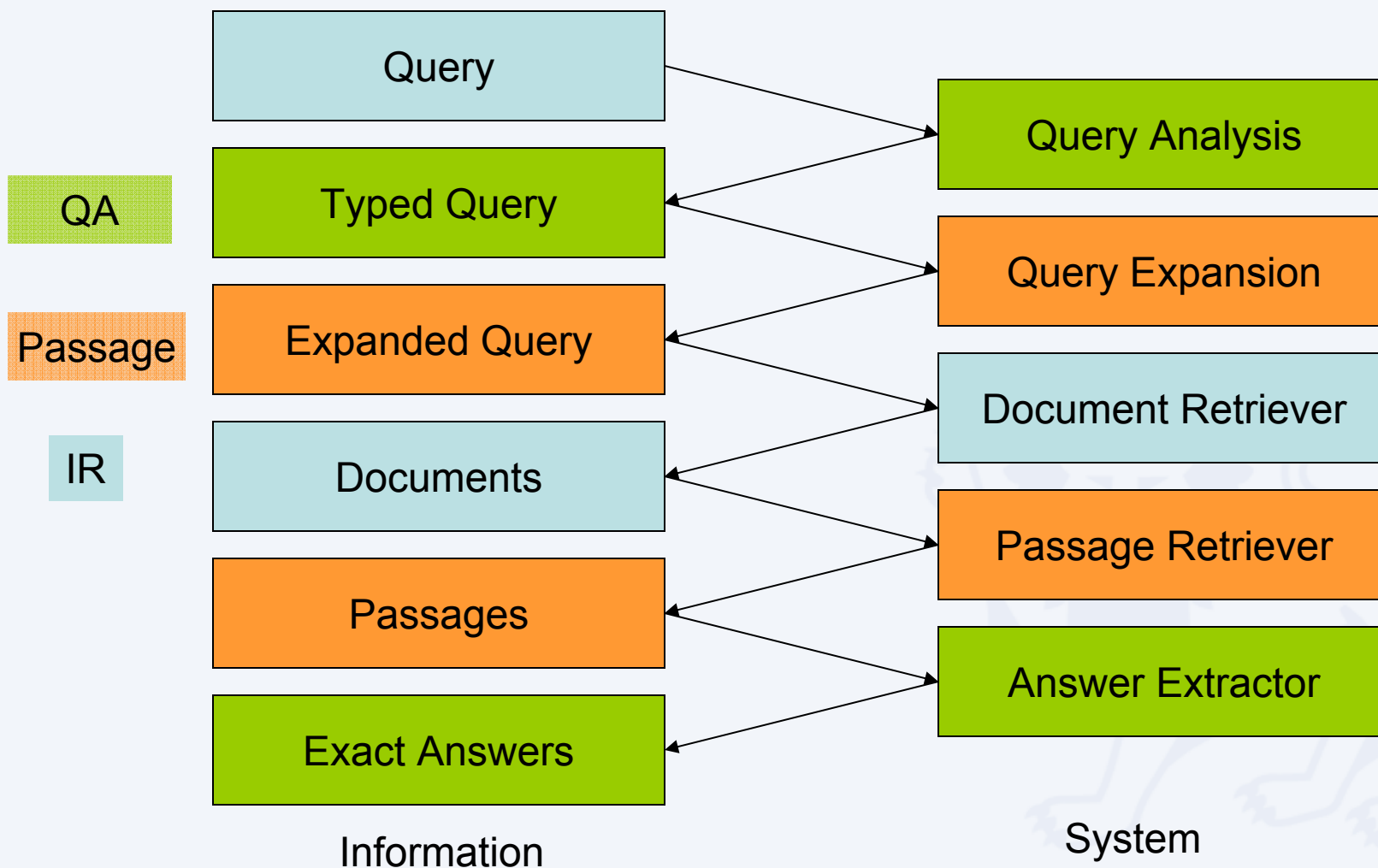
Information Retrieval - MRS

- **Input:** a query
- **Output:** ranked set of documents relevant to the query

- **Issues:**
 - ranking words, use of hyperlinks, internal structure of documents

- **Solve with:**
 - Vector Space Model, Language Model
 - Hyperlink: prestige model (Pagerank) or other model
 - Query analysis, customization, clickthrough data

Generic IR Architecture



Question Answering

- **Input:** a natural language question
- **Output:** an exact answer

- **Considerations:**
 - Factoid vs. List
 - Does question have an answer? Equivalent answers?

- **Solve with:**
 - Cascaded document retrieval, passage retrieval, exact answer retrieval
 - Need both question analysis and answer justification

Register, Genre and Stylistics

- **Input:** text
- **Output:** type of text

- **Text = Content + Presentation**

- **Handling different forms of text**
 - Email/SMS/IM: threading, emoticons, lexical differences
 - Blog: Link structure, trackback, social network analysis
 - Formal report, web pages: formatting, conventional presentation style → segmentation and segment classification
 - **More on this later**

Sentiment Analysis

- **Input:** Text
- **Output:** Opinionated? Positive or Negative?
- **Considerations**
 - Actually a subclass of **text classification**
 - Double negatives infrequent
 - Words carry opinion implicitly (“3G” for a mobile phone)
 - Recent trends: Attribution to **opinion holder**, **aspect** of item being editorialized
- **Solve with:**
 - ML on annotated data

Summary

- **Intro to many issues and parts of NLP**

Words → Phrases → Syntax → Semantics → Discourse →
Pragmatics → Applications

- **Many parts can be solved using machine learning techniques**

- Critical part of clean annotation and feature engineering

- **Academic research often doesn't concern**

- Memory or time efficiency

- In such cases, rule-based heuristics may be better if limited domain (exploit specific domain characteristics)

NLP Resources

- Corpora
- Lexicons

(English) WordNet

- A hierarchical lexicon

S: (v) jump, leap, bound, **spring** (move forward by leaps and bounds) *"The horse bounded across the meadow"; "The child leapt across the puddle"; "Can you jump over the fence?"*

- Organizes in terms of synset
- Includes gloss - definition
- Used to compute similarity between words, sentences
- Other projects to build (manually, automatically)
WordNets in other languages

WordNet – Ch 16.2 J&M

Semantic relation	Description	Part of speech				Example
		N	V	Adj	Adv	
Synonym	A concept that means exactly or nearly the same as another. <i>WordNet</i> considers immediate hypernyms to be synonyms.	×	×	×	×	{ <i>sofa, couch, lounge</i> } are all synonyms of one another. { <i>seat</i> } is the immediate hypernym of the synset.
Antonym	A concept opposite in meaning to another.	×	×	×	×	{ <i>love</i> } is the antonym of { <i>hate, detest</i> }.
Hypernym	A concept whose meaning denotes a superordinate.	×	×			A { <i>feline, felid</i> } is a hypernym of { <i>cat, true cat</i> }.
Hyponym	A concept whose meaning denotes a subordinate.	×	×			A { <i>wildcat</i> } is a hyponym of { <i>cat, true cat</i> }.
Substance meronym	A concept that is a substance of another concept.	×				A { <i>snowflake, flake</i> } is substance of { <i>snow</i> }.

Used to build the hierarchy

WordNet – Ch 16.2 J&M

Semantic relation	Description	Part of speech				Example
		N	V	Adj	Adv	
Part meronym	A concept that is a part of another concept.	×				A { <i>crystal, watch crystal, watch glass</i> } is a part of a { <i>watch, ticker</i> }.
Member meronym	A concept that is a member of another concept.	×				An { <i>associate</i> } is a member of an { <i>association</i> }.
Substance of holonym	A concept that has another concept as a substance.	×				A { <i>tear, teardrop</i> } has { <i>water, H2O</i> } as a substance.
Part of holonym	A concept that has another concept as a part.	×				A { <i>school system</i> } has a { <i>school, schoolhouse</i> } as a part.
Member of holonym	A concept that has another concept as a member.	×				{ <i>organized crime, gangland, gangdom</i> } has { <i>gang, pack, ring, mob</i> } as a member.
Attribute	An adjective that is the value of a noun.	×				{ <i>fast (vs. slow)</i> } is a value of { <i>speed, swiftness, fastness</i> }

WordNet – Ch 16.2 J&M

Semantic relation	Description	Part of speech				Example
		N	V	Adj	Adv	
Cause to	A verb that is the cause of a result.		×			{ <i>give</i> } is the cause of the result { <i>have, have got, hold</i> }
Entailment	A verb that involves unavoidably a result.		×			To { <i>die, decease, perish, go, exit, pass away, expire</i> } involves unavoidably to { <i>leave, leave behind</i> }.
Troponym	A verb that is a particular way to do another.		×			To { <i>samba</i> } is a particular way to { <i>dance, trip the light fantastic</i> }.
Pertainym	An adjective or adverb that relates to a noun.			×	×	{ <i>criminal</i> } relates to { <i>crime</i> }.
Attribute	An adjective that is the value of a noun.	×				{ <i>fast (vs. slow)</i> } is a value of { <i>speed, swiftness, fastness</i> }
Value	A noun that has an adjective for a value.			×		{ <i>weight</i> } has { <i>light (vs. heavy)</i> } as a value.



Role Labeled Data

Annotated data to learn semantic roles in sentences

- **FrameNet** – case frame representation (lexicalized)
 - semantic roles
- **PropBank** – predicate argument structures
 - more syntactically motivated, centered on the verb, more coarse grained / robust , not lexicalized
- **VerbNet** – Merger of both semantic roles and predicate arguments for limited set of verbs

(Tree) Banks

Structure of language for creating NL algorithms from training data

- Penn Treebank – Syntactic information
- Discourse Treebank – discourse information
- SenseEval – Sense disambiguated data
- NomBank – Similar to **role labeled data** but for nouns
 - “IBM lecture”
 - Lecture about IBM?
 - Lecture given by IBM personnel?

NLP / Speech Corpora

Consortiums that license data for commercial development

- **Linguistic Data Consortium (LDC)**

- US based
- most research on these corpora
- better tuned to US intelligence interests
- more diversified genre collection

- **Evaluations and Language resources Distribution Agency (ELDA)**

- European based
- more language diversity

IR Corpora

- **Reuters 21578**
 - Default classification dataset, too small for today's investigation purposes
 - Subsequent work in building Reuters RCV1
- **TREC / INEX / CLEF / NTCIR**
 - Yearly tests of IR systems
 - TREC: oldest, most variety, also TRECvid
 - INEX: XML retrieval
 - CLEF / NTCIR: Multilingual retrieval
- **WebKB, Open Directory Project**
 - Web page classification
 - Harder to get datasets → commercial concerns, AOL gaffe

Summary

- **Resources / corpora necessary if you don't want to reinvent the wheel**
- **Worth the licensing fee and investigation**
- **Pulling data from the web as-is without consent may constitute copyright violation**

