NLP

NATURAL LANGUAGE PROCESSING

Contents

NLP

- Natural Language Understanding
  - Taking some spoken/typed sentence and working out what it means
- Natural Language Generation
  - Taking some formal representation of what you want to say and working out a way to express it in a natural human language like English
- Fundamental goal: deep understand of broad language
  - Not just string processing or keyword matching
- Target end systems
  - speech recognition, machine translation, question answering…
  - spelling correction, text categorization…

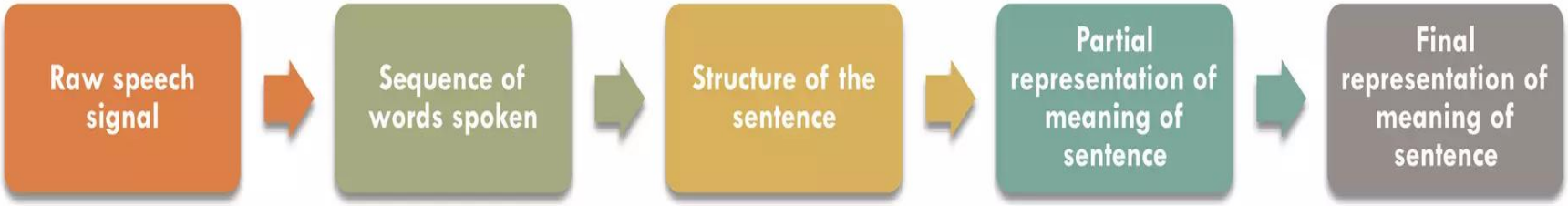
Applications

- Text Categorization - classify documents by topics, language, author, spam filtering, sentiment classification (positive, negative)
- Spelling & Grammar Corrections
- Speech Recognition
- Summarization
- Question Answering
- Better search engines
- Text-to-speech
- Machine aided translation
- Information Retrieval
  - Selecting from a set of documents the ones that are relevant to a query
- Extracting data from text
  - Converting unstructured text into structure data

Natural Language Understanding

- Answering an essay question in exam
- Deciding what to order at a restaurant by reading a menu
- Realizing you've been praised
- Appreciating a poem

Natural Language Understanding



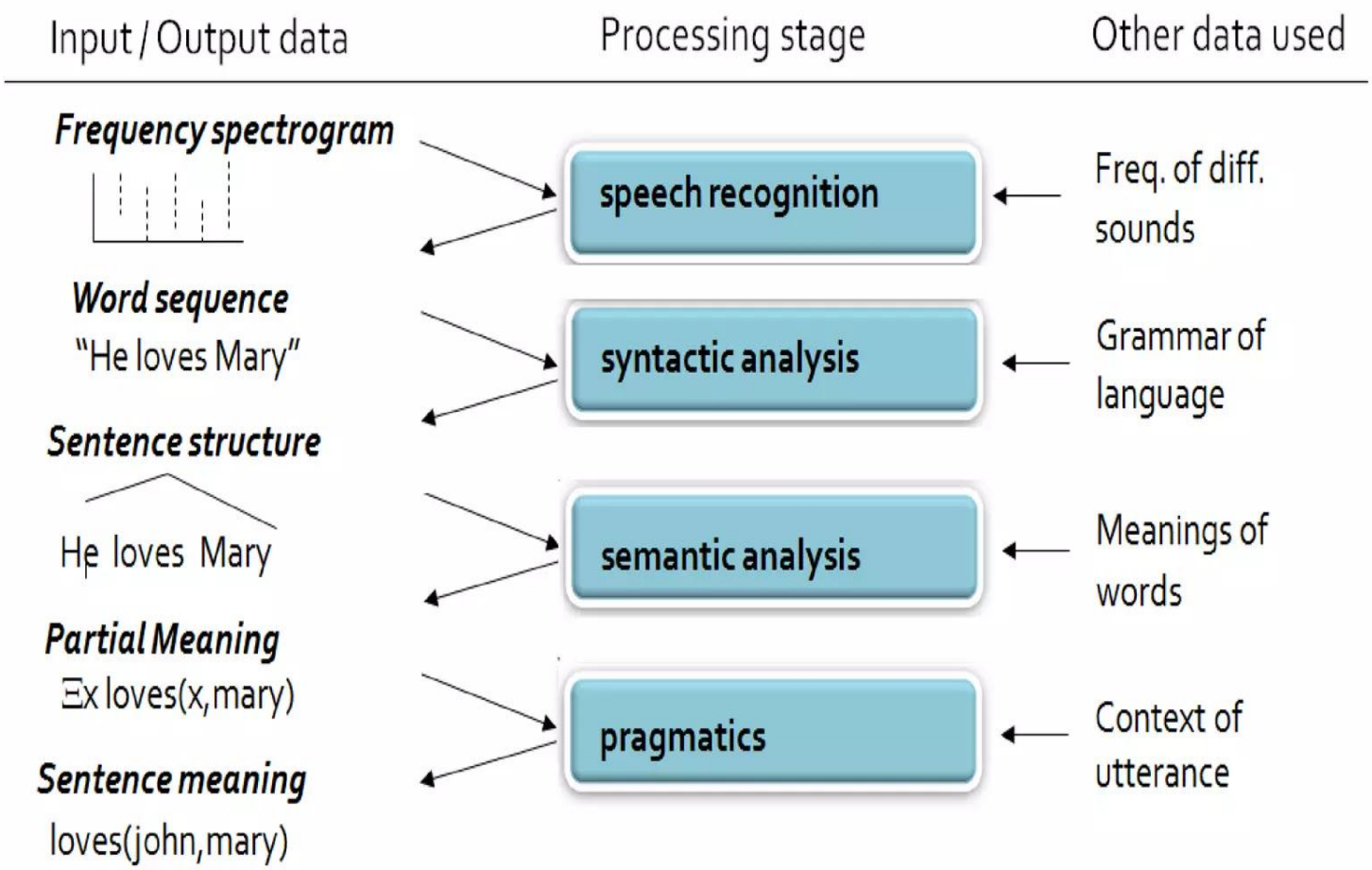
Speech recognition

Syntactic analysis using knowledge of the grammar

Semantic analysis using information about meaning of words

Pragmatic analysis using info. about context

Natural Language Understanding



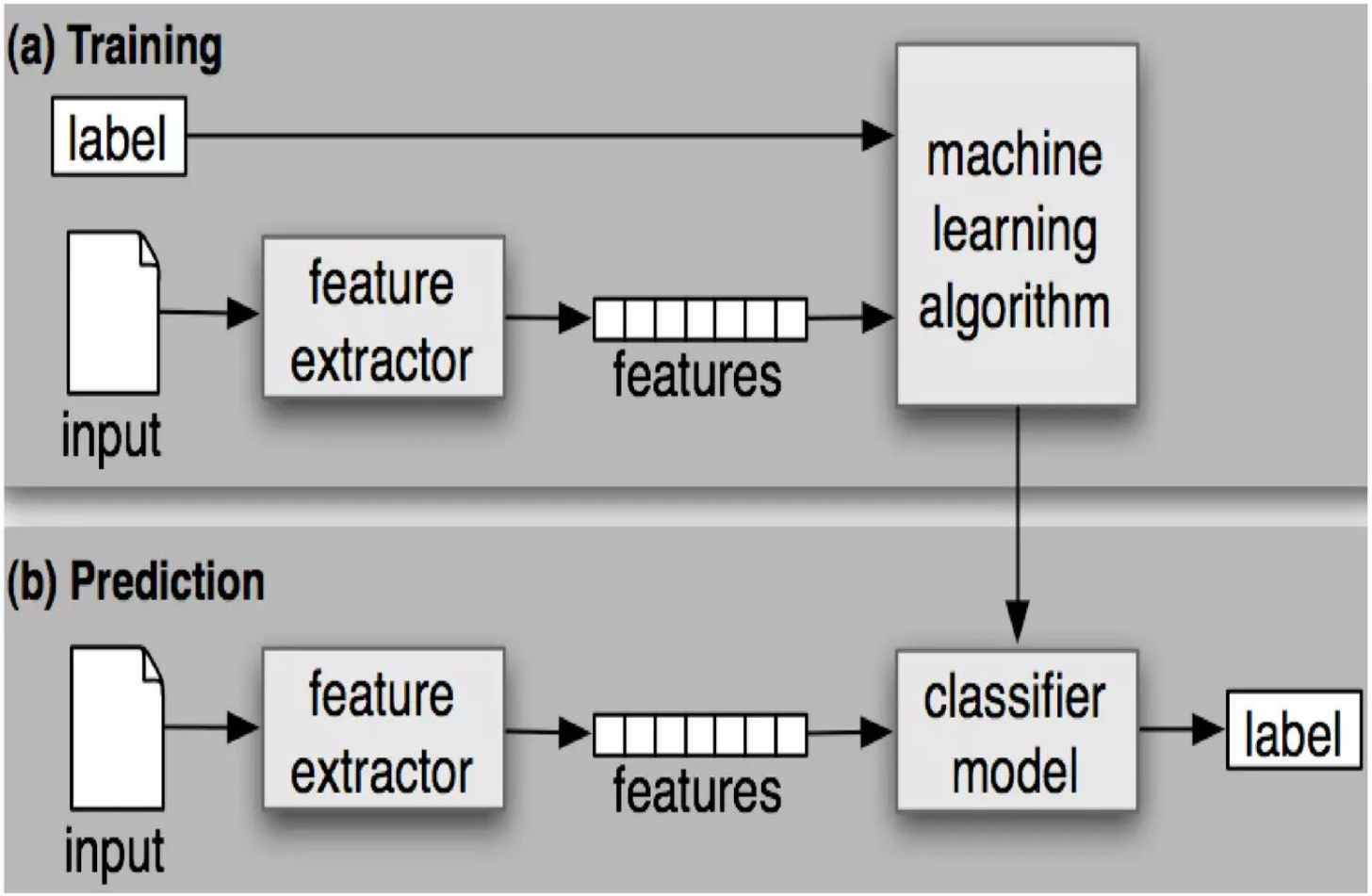
Text Categorization

- Text annotation - classify entire document
- Sentiment classification
  - What features of the text could help predict # of likes?
  - How to identify customer opinions?
  - Are the features hard to compute? (syntax? sarcasm?)
- Is it spam?
- What medical billing code for this visit?
- What grade for an answer to this essay question?

Text Categorization

- Is it interesting to this user?
  - News filtering; helpdesk routing
- Is it interesting to this NLP program?
  - If it's Spanish, translate it from Spanish
  - If it's subjective, run the sentiment classifier
  - If it's an appointment, run information extraction
- Where should it be filed?
  - Which mail folder? (work, friends, junk, urgent ...)
  - Yahoo! / Open Directory / digital libraries

Text Categorization



Syntactic Analysis

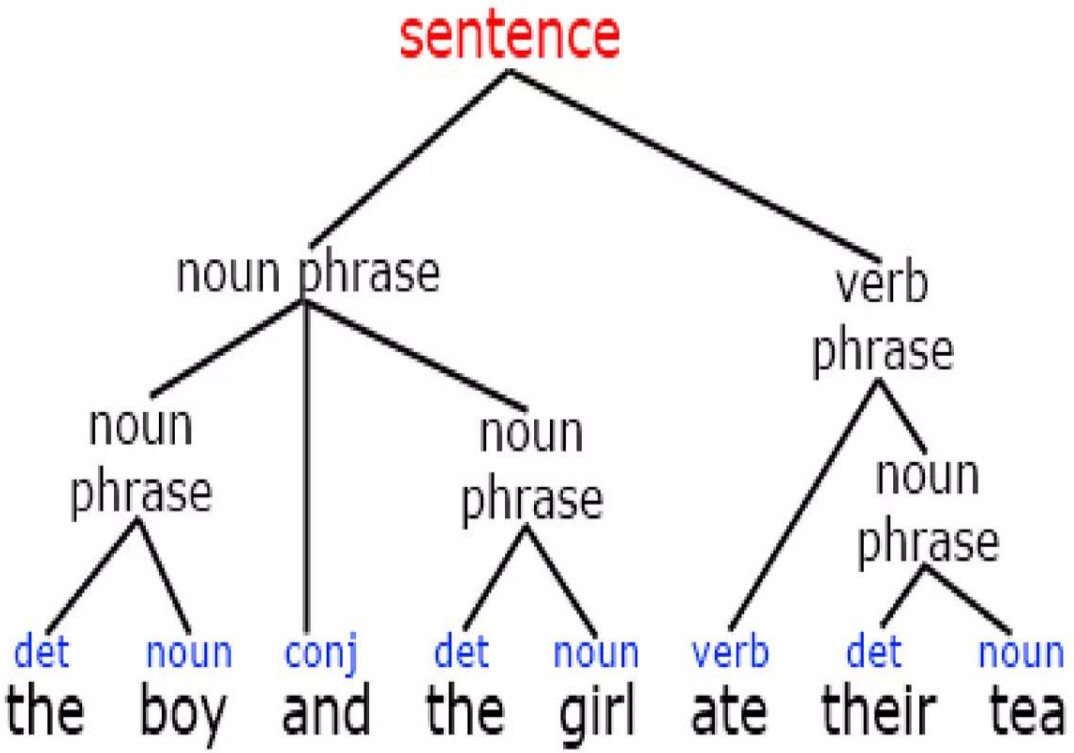
- Syntax mapped into semantics
  - Nouns ↔ things, objects, abstractions.
  - Verbs ↔ situations, events, activities.
  - Adjectives ↔ properties of things, ...
  - Adverbs ↔ properties of situations, ...

Symbols in Grammar

- S - Sentence
- NP- Noun Phrase
- PN- Proper Noun
- N-Noun
- VP-Verb Phrase
- Adv-Adverb
- V-Verb
- Adj-Adjective
- Prep-Preposition
- Art-Article

- Pro-Pronoun
- PP-Prepositional Phrase
- * Ungrammatical Sentence
- → Consists of / rewrites as
- ( ) Optional Constituent
- { } Only one of these constituents must be selected

Syntactic Analysis - Grammar



Semantic Analysis

- Generates meaning/representation of the sentence from its syntactic structures
- Represents the sentence in meaningful parts
- Uses possible syntactic structures and meaning
- Builds a parse tree with associated semantics
- Semantics typically represented with logic

Pragmatic Analysis

- Uses context
- Uses partial representation
- Includes purpose and performs disambiguation
- Where, when, by whom an utterance was said
- Uses context of utterance
  - Where, by who, to whom, why, when it was said
  - Intentions - inform, request, promise, criticize, …

Pragmatic Analysis

- Handling Pronouns
  - Mary eats apples. She likes them
    - She = Mary, them = apples
- Handling ambiguity
  - Pragmatic ambiguity - you're late - What's the speaker's intention - informing or criticizing?

NLP Challenges

- NLP systems needs to answer the question “who did what to whom”
- MANY hidden variables
  - Knowledge about the world
  - Knowledge about the context
  - Knowledge about human communication techniques
    - Can you tell me the time?
- Problem of scale
  - Many (infinite?) possible words, meanings, context

NLP Challenges

- Problem of sparsity
  - Very difficult to do statistical analysis, most things (words, concepts) are never seen before
- Long range correlations
- Key problems
  - Representation of meaning
  - Language presupposes knowledge about the world
  - Language only reflects the surface of meaning
  - Language presupposes communication between people

NLP Challenges

- Fictitious worlds - people on mars can fly
- Defining scope
  - people like ice-cream - does this mean all people like ice cream?
- Language is changing and evolving
- Complex ways of interaction between the kinds of knowledge
- Exponential complexity at each point in using the knowledge

Meaning

- From NLP viewpoint, meaning is a mapping from linguistic forms to some kind of representation of knowledge of the world
  - Physical referent in the real world
  - Semantic concepts, characterized also by relations.
- It is interpreted within the framework of some sort of action to be taken

Meaning – Representation and Usage

- I am Italian
  - From lexical database (WordNet)
  - Italian =a native or inhabitant of Italy→ Italy = republic in southern Europe [..]
- I am Italian
  - Who is "I"?
- I know she is Italian/I think she is Italian
  - How do we represent I know and I think
  - Does this mean that I is Italian?
  - What does it say about the I and about the person speaking?
- I thought she was Italian
  - How do we represent tenses?

Corpus-based Statistical Approaches

- How can a machine understand these differences?
  - Decorate the cake with the frosting
  - Decorate the cake with the kids
- Rules based approaches
  - Hand coded syntactic constraints and preference rules
  - The verb decorate require an animate being as agent
  - The object cake is formed by any of the following, inanimate entities
    - cream, dough, frosting…..

Corpus-based Statistical Approaches

- These approaches are time consuming to build, do not scale up well and are very brittle to new, unusual, metaphorical use of language
- To swallow requires an animate being as agent/subject and a physical object as object
  - I swallowed his story
  - The supernova swallowed the planet
- A Statistical NLP approach seeks to solve these problems by automatically learning lexical and structural preferences from text collections (corpora)

Corpus-based Statistical Approaches

- Statistical models are robust, generalize well and behave gracefully in the presence of errors and new data
- Steps
  - Get large text collections
  - Compute statistics over those collections
    - The bigger the collections, the better the statistics

Corpus-based Statistical Approaches

- Decorate the cake with the frosting
- Decorate the cake with the kids
- From (labeled) corpora we can learn that
  - #(kids are subject/agent of decorate) > #(frosting is subject/agent of decorate)
- From (unlabelled) corpora we can learn that
  - #("the kids decorate the cake") >> #("the frosting decorates the cake")
  - #("cake with frosting") >> #("cake with kids")
- Given these facts, we need a statistical model for the attachment decision

Corpus-based Statistical Approaches

- Topic categorization: classify the document into semantics topics

Document 1	Document 2
The nation swept into the Davis Cup final on Saturday when twins Bob and Mike Bryan defeated Belarus's Max Mirnyi and Vladimir Voltchkov to give the Americans an unsurmountable 3-0 lead in the best-of-five semi-final tie. Topic = sport	One of the strangest, most relentless hurricane seasons on record reached new bizarre heights yesterday as the plodding approach of Hurricane Jeanne prompted evacuation orders for hundreds of thousands of Floridians and high wind warnings that stretched 350 miles from the swamp towns south of Miami to the historic city of St. Augustine. Topic = Natural Event

Corpus-based Statistical Approaches

- Topic categorization: classify the document into semantics topics

Document 1 (sport)	Document 2 (Natural Event)
The nation swept into the Davis Cup final on Saturday when twins Bob and Mike Bryan …	One of the strangest, most relentless hurricane seasons on record reached new bizarre heights yesterday as….

- From (labeled) corpora we can learn that
  - #(sport documents containing word cup) > #(nature event documents containing word cup) - feature
- We then need a statistical model for the topic assignment

Corpus-based Statistical Approaches

- Feature extractions
  - Usually linguistics motivated
- Statistical models
- Data
  - Corpora, labels, linguistic resources

Measuring Performance

- Classification accuracy
  - What % of messages were classified correctly?
- Is this what we care about?
- Which system is better?

	Overall accuracy	Accuracy on spam	Accuracy on gen
System 1	95%	99.99%	90%
System 2	95%	90%	99.99%

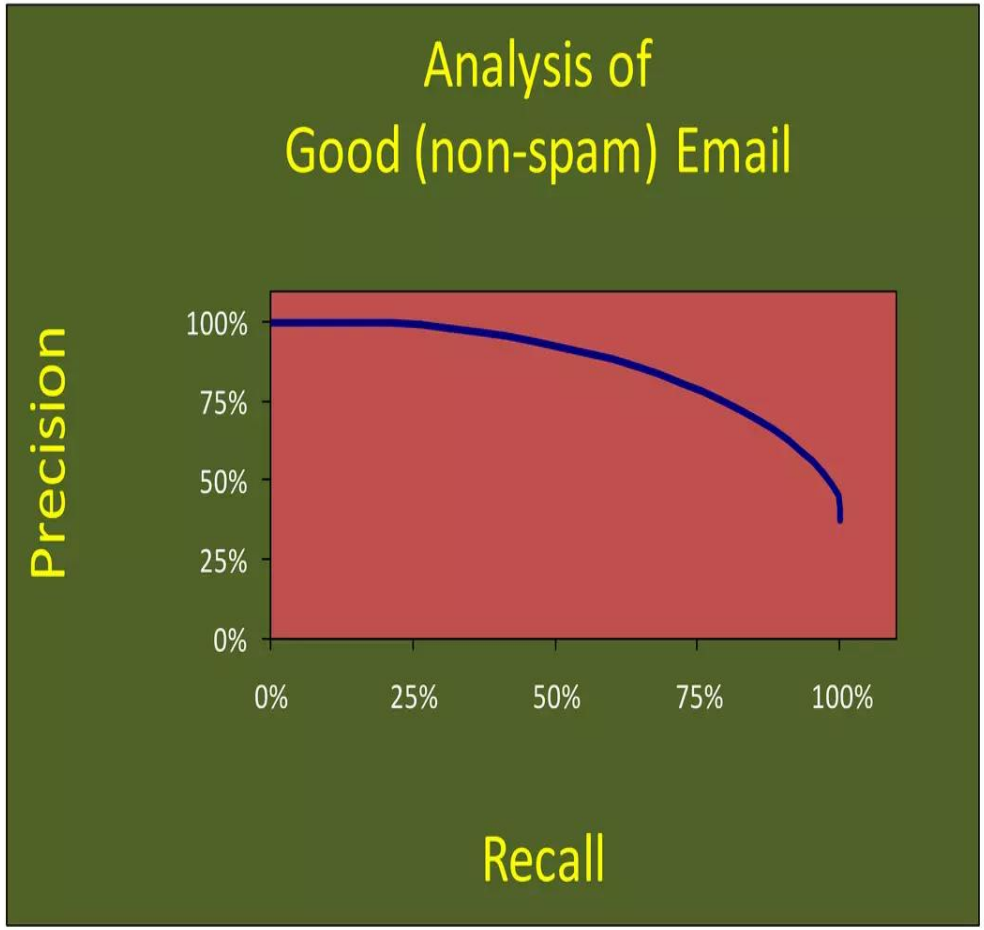
Measuring Performance

- Precision = $\frac{\text{good messages kept}}{\text{all messages kept}}$

- Recall = $\frac{\text{good messages kept}}{\text{all good messages}}$

- Move from high precision to high recall by deleting fewer messages
  - delete only if spam content > high threshold

Measuring Performance



NLP - Supervised Learning Methods

- Conditional log-linear models
  - Feature engineering - Throw in enough features to fix most errors
  - Training - Learn weights θ such that in training data, the true answer tends to have a high probability
  - Test - Output the highest-probability answer
  - If the evaluation metric allows for partial credit, can do fancier things
    - minimum-risk training and decoding

Part of Speech Tagging

- Treat tagging as a token classification problem
  - Tag each word independently given
    - features of context
    - features of the word's spelling (suffixes, capitalization)
- Use an HMM
  - the tag of one word might depend on the tags of adjacent words
- Combination of both
  - Need rich features (in a log-linear model), but also want feature functions to depend on adjacent tags
  - So, the problem is to predict **all** tags together.

Named Entity Recognition

- Deals with the *detection* and *categorization* of proper names
- Labeling all occurrences of named entities in a text
  - Named Entity = People, organizations, lakes, bridges, hospitals, mountains…
- Well-understood technology, readily available and works well
- Uses a combination of enumerated lists (often called gazetteers) and regular expressions

Named Entity Recognition

Commonly Used Named Entity Types

NE Type	Examples
ORGANIZATION	*Georgia-Pacific Corp., WHO*
PERSON	*Eddy Bonte, President Obama*
LOCATION	*Murray River, Mount Everest*
DATE	*June, 2008-06-29*
TIME	*two fifty a m, 1:30 p.m.*
MONEY	*175 million Canadian Dollars, GBP 10.40*
PERCENT	*twenty pct, 18.75 %*
FACILITY	*Washington Monument, Stonehenge*
GPE	*South East Asia, Midlothian*

Complex Entities and Relationships

- Uses Named Entities as components
- Pattern-matching rules specific to a given domain
- May be multi-pass - One pass creates entities which are used as part of later passes
  - First pass locates CEOs, second pass locates dates for CEOs being replaced…
- May involve syntactic analysis as well, especially for things like negatives and reference resolution

Simple CF Grammars

- sentence ↵
- noun_phrase verb_phrase ↵
- proper_name verb_phrase ↵
- Jim verb_phrase ↵
- Jim verb noun_phrase prep_phrase ↵
- Jim ate noun_phrase prep_phrase ↵
- Jim ate article noun prep_phrase ↵
- Jim ate a noun prep_phrase ↵
- Jim ate a pizza prep_phrase ↵
- Jim ate a pizza preposition noun_phrase ↵
- Jim ate a pizza on noun_phrase ↵
- Jim ate a pizza on article noun ↵
- Jim ate a pizza on the noun ↵
- Jim ate a pizza on the bus

Simple CF Grammars

- Other examples of sentences generated by this grammar:
  - `Jim ate a pizza`
  - `Dan yawns on the bus`
- These *wrong* data will also be recognized:
  - `Jim ate an pizza`
  - `Jim yawns a pizza`
  - `Jim ate to the bus`
  - `the boys yawns`
  - `the bus yawns`
- ... but not these, obviously correct:
  - `the pizza was eaten by Jim`
  - `Jim ate a hot pizza`
- and so on, and so forth.

N-grams

- Letter or word frequencies - 1-grams - unigrams
  - useful in solving cryptograms - ETAOINSHRDLU…
- If you know the previous letter - 2-grams - bigrams
  - “h” is rare in English (4%)
  - but “h” is common after “t” (20%)
- If you know the previous two letters - 3-grams - trigrams
  - “h” is really common after “(space) t”

References

1. Ruslan Mitkov, The Oxford Handbook Of Computational Linguistics, Oxford Universitty Press, 2003.
2. Robert Dale, Hermani Moisi, Harold Somers, Handbook Of Natural Language Processing, Markcel Dekker Inc.
3. James Allen, Natural Language Processing, Pearson Education, 2003.
4. Manning C, Raghavan P, Schuetze H. Introduction to Information Retrieval. Cambridge, UK: Cambridge University Press, 2008
5. Douglas Biber, Susan Conrad, Randi Reppen, Corpus Linguistics – Investigating Language Structure And Use, Cambridge University Press, 2000.
6. David Singleton, Language And The Lexicon: An Introduction, Arnold Publishers, 2000.
7. Allen, James, Natural Language Understanding, second edition (Redwood City: Benjamin/Cummings, 1995).
8. Ginsberg, Matt, Essentials of Artificial Intelligence (San Mateo: Morgan Kaufmann, 1993)
9. Hutchins W. The First Public Demonstration of Machine Translation: the Georgetown-IBM System, 7th January 1954. 2005.
10. Chomsky N. Three models for the description of language. IRE Trans Inf Theory 1956;2:113–24
11. Aho AV, Sethi R, Ullman JD. Compilers: Principles, Techniques, Tools. Reading, MA: Addison-Wesley, 1988
12. Chomsky N. On certain formal properties of grammars. Inform Contr 1959;2:137–67
13. Friedl JEF. Mastering Regular Expressions. Sebastopol, CA: O'Reilly & Associates, Inc., 1997