Natural Language Processing

What is NLP?

- NLP = Natural Language Processing
  - Natural Language = Human languages
  - Processing of human languages using computers
  - Human language = speech + text
- Other names include:
  - Computational Linguistics
  - NLE = Natural Language Engineering
  - HLT = Human Language Technology

Is NLP useful?

- NLP can be used in the following areas:
  - machine generated dictionaries, new word discovery, spelling checking, automatic generation of words(terms)
  - parsing and grammar checking, sentence generation, NLG(Natural language generation)
  - machine translation
  - speech synthesis and speech recognition(speech input/output to computers, speech transmission through internet)
  - handwriting recognition(pen computers)
  - Q&A, automatic speech response, voice command
  - information retrieval, extraction, summarization, data mining, knowledge discovery. Internet makes text processing a MUST.
  - many more

Levels of Language

- **phonetic** - pronunciation
- **orthographic** - writing
- **word** - lexicons
- **sentences**-phrases, short sentences, multiple sentences
- **paragraphs** - consisting of a number of related sentences
- **articles** - consisting of a number of paragraphs and a main theme

Areas of study

- **syntax** - concern about form, ways of expression
- **semantics** - concern about the meaning
- **pragmatics** - concern about the actual usage of the language.

Why NL Understanding is hard?

- Natural language is extremely rich in form and structure, and **very ambiguous**.
  - How to represent meaning,
  - Which structures map to which meaning structures.
- One input can mean many different things. Ambiguity can be at different levels.
  - Lexical (word level) ambiguity -- different meanings of words
  - Syntactic ambiguity -- different ways to parse the sentence
  - Interpreting partial information -- how to interpret pronouns
  - Contextual information -- context of the sentence may affect the meaning of that sentence.
- Many input can mean the same thing.
- Interaction among components of the input is not clear.

Knowledge of Language

- **Phonology** – concerns how words are related to the sounds that realize them.
- **Morphology** – concerns how words are constructed from more basic meaning units called morphemes. A morpheme is the primitive unit of meaning in a language.
- **Syntax** – concerns how can be put together to form correct sentences and determines what structural role each word plays in the sentence and what phrases are subparts of other phrases.
- **Semantics** – concerns what words mean and how these meaning combine in sentences to form sentence meaning. The study of context-independent meaning.

Knowledge of Language (cont.)

- **Pragmatics** – concerns how sentences are used in different situations and how use affects the interpretation of the sentence.

- **Discourse** – concerns how the immediately preceding sentences affect the interpretation of the next sentence. For example, interpreting pronouns and interpreting the temporal aspects of the information.

- **World Knowledge** – includes general knowledge about the world. What each language user must know about the other's beliefs and goals.

Ambiguity

<u>I made her duck.</u>

- How many different interpretations does this sentence have?
- What are the reasons for the ambiguity?
- The categories of knowledge of language can be thought of as ambiguity resolving components.
- How can each ambiguous piece be resolved?
- Does speech input make the sentence even more ambiguous?
  - Yes – deciding word boundaries

Ambiguity (cont.)

- Some interpretations of : **`I made her duck.`**
  1. I cooked *duck* for her.
  2. I cooked *duck* belonging to her.
  3. I created a toy duck which she owns.
  4. I caused her to quickly lower her head or body.
  5. I used magic and turned her into a *duck*.
- `duck` – morphologically and syntactically ambiguous: noun or verb.
- `her` – syntactically ambiguous: dative or possessive.
- `make` – semantically ambiguous: cook or create.
- `make` – syntactically ambiguous:
  - Transitive – takes a direct object. => 2
  - Di-transitive – takes two objects. => 5
  - Takes a direct object and a verb. => 4

Characteristic of human language

- It is discrete, symbolic
- It is only an approximation - fuzzy
- It is an open set
- It is changing with time and locality
  - It is a moving set
- It is different for different person
  - misunderstanding often occur between people of same kind speaking the same language
- Great variations in form, in expression

Related Fields

- Man-machine Interface - Computer I/O.
  - Human engineering
  - Display technology
- Knowledge Acquisition and Data Mining, Scientific discovery
- Computer-aided Learning
- Security - authenticity, information services
- Psycho-linguistics
  - study of human psychological process in language - include perception, acquisition, learning, processing, understanding, response etc.

Topics of interest

- Word characters and statistics
  - Entropy, Mutual Information, N-gram
- Word formation - Morphology
- Meaning of Words
- New Words and Detection
- WordNet and Application - Word associations
  - May not cover all topics.

Word Structure - Morphology

Word

- Word is the smallest unit that has ‘complete meaning’.
  - this definition is ambiguous as the so-called ‘complete meaning’ is not well defined.
  - it provides a unit of rather complete ‘concept’.
- Two types of words:
  - simple words: student, teacher,
  - compound words:grass-hopper, code-book, day-light, etc.
- Simple words are constructed from smaller units called morphemes.
  - understanding: under-stand-ing
  - incompleteness: in-com-plete-ness
- Compound words made up of a string of simple words

Morphology

- Morphology is a study of **construction** of words
  - Its major concern is on the ***expression of words***, their ***surface*** structure.
- Morphology consists of
  - Inflection - creates various forms of each word
    - *Examples: run, runs, ran*
    - **Chinese words have no inflectional structure**
  - Derivation - creates new words from existing words, often of different syntactic categories
    - *complete, completeness*

Morphemes and Allomorphs

- ***Morphs*** are the smallest meaningful segments into which a word is divided.
- ***Allomorphs*** are the various forms of any morpheme.
- English words consist of zero or more prefixes, followed by the root and then zero or more suffixes
- A state engine can be used to express the construct of an English word

$$\begin{array}{lccccccc} & \text{Prefix} & & \text{Root} & & \text{Suffix} & & \text{Suffix} \\ untouchables = & un & + & touch & + & able & + & s \end{array}$$

Morphemes

- Morphemes are word sub-units. Every morpheme has its own 'meaning'.
- Most English morphemes are derived from **Latin**.
  - Examples: abil-, able appear in able, ability, inability, disable, disability, unable, enable, able-bodied, agreeable, amicable, capable, durable, impeccable, indispensable, inevitable, insatiable, laudable, palpable, portable, suitable.
- There are 3 types of morphemes:
  - root - able, cur-, cou-, sur-lic- lei..
  - prefix - a-, ab-, ad-, anti-, ana-...
  - suffix - -ate, -ble, -dom, -er, -ery, -ic

Inflection

- Inflection is an alteration of the form of a word indicating grammatical features, such as number, person, or tense.
- **number**: singular and plural, such as *box-boxes, man-men, table-tables*
- **person** - *I-me, he-him, she-her*
- **tense** - *run-runs-ran, tell-tells-told, speak-speaks-spoke-spoken, eat-eats-ate-eaten*
- Some changes are regular; but many are irregular.
  - Rules or ATN(augment transition net) can be constructed for regular words; for irregular words, look-up tables can be used.

English Inflection - Verbs

- Verbs - regular forms
  - plain form, with no ending
    - *Fido will bark*
  - *s* form
    - *Fido barks.*
  - *ing* form
    - *a barking dog.*
  - *ed* form
    - *Fido barked.*

- Verb - irregular forms
  - past tense does not end in ed; instead it is formed by vowel change
    - *sing, sang*
    - *ear, ate*
  - *en* form, used after has and other auxiliary verbs
    - *Fido has eaten.*

Others

- adjectives and adverbs
  - take suffixes *er* and *est*.
- Nouns
  - add s or es
  - some irregular
    - child, children
    - ox, oxen
- Possessive ’s as *clitic* not suffix
  - Clitic is a syntactically different word that is always pronounced as part of previous word.
  - Example: the boy who came Early’s lunch (‘s is NOT a suffix of early).

Spelling Rules - **Morphographemics**

- Final e detection
  - silent final e disappears before any suffix that begins with a vowel, as in *rake+ing=raking, rake+ed=raked.*
- y to i rule
  - final y changes to I before any suffix that does not begin with I, as in *carry+ed=carried, carry+ing=carrying*.
- s to es rule
  - the suffix s, on both nouns and verbs, appears as es after s, z,x,sh,ch and after y which has changed to i, as in *grass+s=grasses, dish+s=dishes, carry+s=carries.*
- final consonant doubling
  - a single final consonant doubles before any suffix that begins with a vowel, as in *grab+ed=grabbed, grab+ing=grabbing, big+er=bigger.*
  - There are exceptions as in *offering, chamfering*.

Nominalization(名詞化)

- Turning verbs, adjectives into nouns.
  - add -ing to verb, such as running, housing, keeping
  - add -ness, ty to adjectives, such as competitiveness, activity, ability.

Compound words

- English compound words are formed by concatenation of a number of words. Examples:
  - grasshopper, congresswoman, codebook, cordwood. corncakes, backdoor, baseman, basketball, daybreak, daylight, deadlock, deepfreezer, trademark …
- Compound words are written as one word. Phrases such as *school teachers* are written as a number of words with white spaces in between the words.

Phrasal Words

- Word phrases are formed by concatenation of more than one word, such as
  - school teacher, text book, Housing Development Board, Singapore Airport
- Most phrasal words are nouns
  - Phrasal words are most formed by concatenation of nouns

Word Category

- Words can be divided into the following categories, call part of speech(POS)
- Words are traditionally classified according to their functions in context into the following POS: noun, pronoun, verb, adjective, adverb, preposition, conjunction, and interjection, and sometimes the article.
- However, in NLP, words can be classified in much different ways, such as:
  - in ENGTWOL (Constraint English Grammar): adjective, abbreviation, adverb coordinating conjunction(and), subordinating conjunction(that), determiner, infinitive marker(to), interjection, noun, negative particle(not), numeral, -ing form, -ed.-en form, preposition, pronoun, verb
  - in LINK Grammar, words are classified into verb, noun, determiners, adjective, preposition, adverb, conjunction…
- There is no universally agreed ways of classified and a set of so-called standard POS. Each NLP system designs its own set of POS.

New Words Discovery

- About 10-100 new words appear every day
  - In a Chinese text corpus, as high as 10% of the total words are out-of-vocabulary or new words. Treatment of new words is very crucial to Chinese text analysis.
- a good lexicon analysis tool is needed in the detection of new words.
- New word discovery is a very important topic in NLP because of its value in
  - discovery of new trend in science
  - new trend in social, political and economical activities

How new words are generated?

- Three major approaches:
  - Constructed from existing morphemes or words
    - Earlier, most new words are formed in this way
  - Abbreviations
    - Laser, RAM, ROM,
    - HDB, USRP, NUS
  - Overload existing words with new meaning
    - update, pull technology, surfer, server, etc.

Computational Linguistic approach to Morphology

- Two major issues in computational morphology
  - discovery of word structure, spelling checking
  - generation of new words
- Spelling checking - letter tree
- Abstract Morphology - 2 level approach
- Algorithms
  - ATN parser
  - N-gram and MI approach
  - HMM parser

N-gram approaches for word segmentation

- a combination of N-gram and N-gram MI can be used to identify possible word segments that are prefix, suffix or root.
- Method
  - collect a large number of English words from a huge corpus
  - obtain n-grams, n=1… 10 and their MIs
  - rank the n-grams using their frequencies and MIs
  - select the first 3000 n-grams according to the rank
  - examine them and decide if there can be suffices, prefixes and roots. You can do this by
    - manual inspection, some expertise in linguistic needed
    - checking with a lexicon dictionary of roots
    - develop an algorithm to decide if an n-gram is a root.
    - train a neural net/HMM? to recognize roots.

Formal Language Theory
Grammar and Syntax

Phrase Structure

The structure of an English sentence can be best represented in a hierarchical tree. For example:

The dog chased a cat into the garden

GRAMMAR: a set of PS rules

S → NP VP

NP → D N

VP → V NP |V NP PP

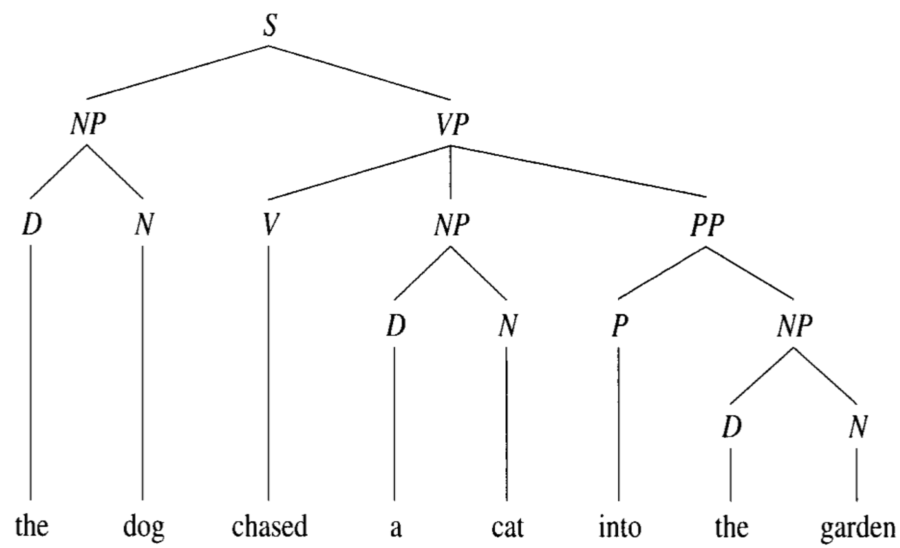
PP → P NP

D → the | a

N → dog|cat|garden

V → chased|saw

P → into



An equivalent expression:

$[_S[_{Np}[_D\text{the}][_N\text{dog}]]_{VP}[_V\text{chased}][_{NP}[_D\text{a}][_N\text{cat}]][_{PP}[_P\text{into}[_{NP}[_D\text{the}][_N\text{garden}]]]$

Phrase Structure Rules
-two important parts

- POS - Part-of-Speech or Categories
  - Words are classified in POS such as noun, verb, adjective, adverb, auxiliary, preposition etc
- Rules are re-writing rules in the form of X→A B C ..
  - X can be re-written into components A, B and C.
  - X, A, B, C… are POSs
  - Example: S →VP NP
- These rules are rules of ***generation***.

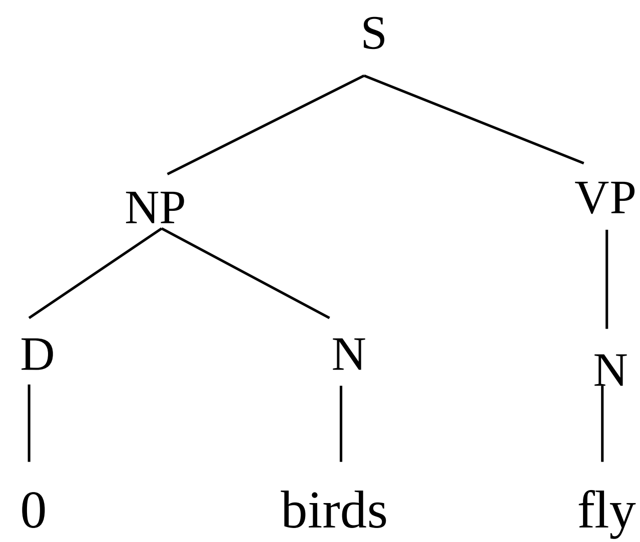
PS Rules and CFG

- PS rules are expressed in CFG - context free grammar - so called because the rules do not refer to the **context** where it is applied.
  - There is only a single item on the left hand side.
  - On the right hand side, if the number of item is always 2 (or 1 in the terminate case), then the expression is in the Chomsky Normal Form (or CNF).
- In a context-sensitive grammar, the rule will only be applicable when it is in the correct context. It has more than one item on its left-hand side.

  $Q_1AQ_2 \rightarrow Q_1PQ_2$, Q_1 and Q_2 are the context that rule $A \rightarrow P$ is applied.
- CFG is a type 0 generative grammar. $G=(V_N, V_T, S, F)$ where V_N are non-terminal symbols, V_T are terminal symbols, S is the starting symbol and F is a set of production rules).

Phrase-structure Formalism

- **Non-terminal** and **terminal** rules
  - non-terminal: S $\rightarrow$ VP NP, NP $\longrightarrow$ D N
  - terminal: N $\rightarrow$ garden, V $\rightarrow$ saw
- There can be more than one rule expanding the same symbol
- A symbol can also be expanded to **NULL**
  - D $\rightarrow$ 0
- Abbreviations
  - VP $\rightarrow$ V(NP)
    - NP is optional, equivalent to 2 rules
      VP $\rightarrow$ V NP and VP $\rightarrow$ V
  - Combination of rules is allowed.
    - N $\rightarrow$ dog | garden | … equivalent to rules
      N $\rightarrow$ dog, N $\rightarrow$ garden ...



Recursion

- PS Rules are recursive

  The dog chased the cat

  The girl thought the dog chased the cat

  The butler said the girl thought the dog chased the cat

  The gardener claimed the butler said the girl thought the dog chased the cat

- Recursion is a common phenomena in NLP
- PS rules allow recursion by rules like
  - S → NP VP,
  - VP → V S

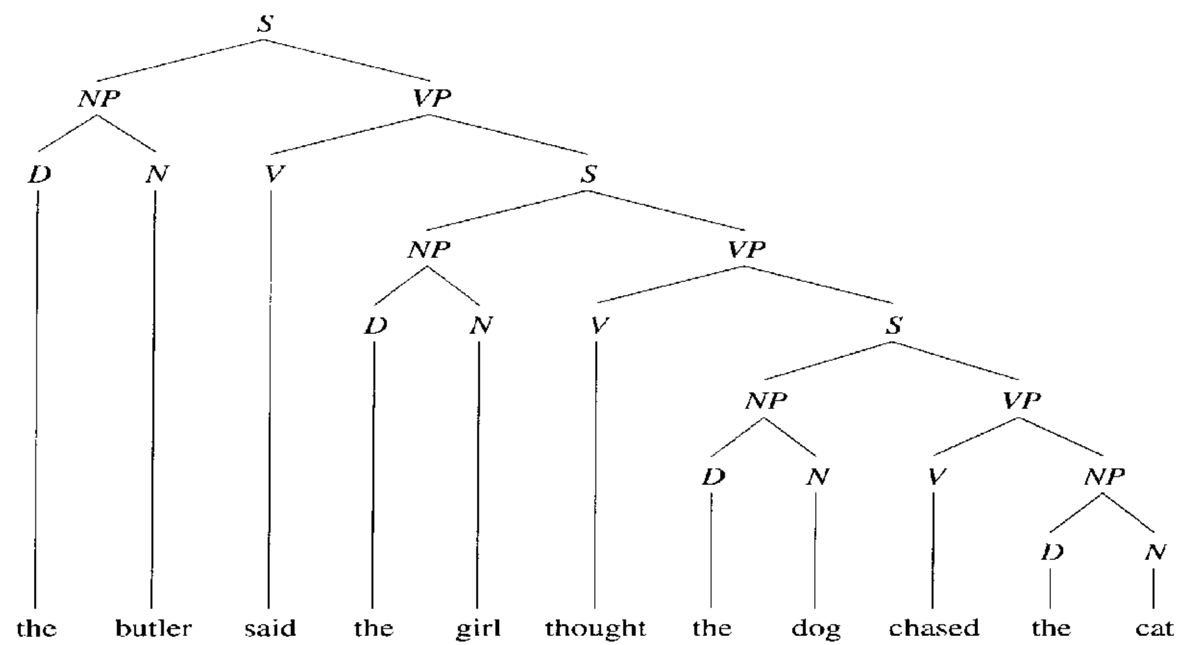
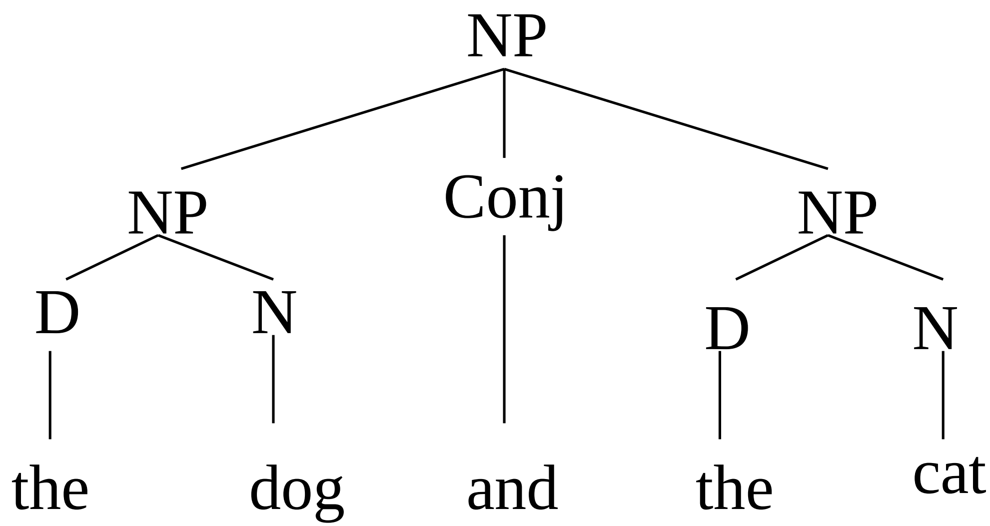


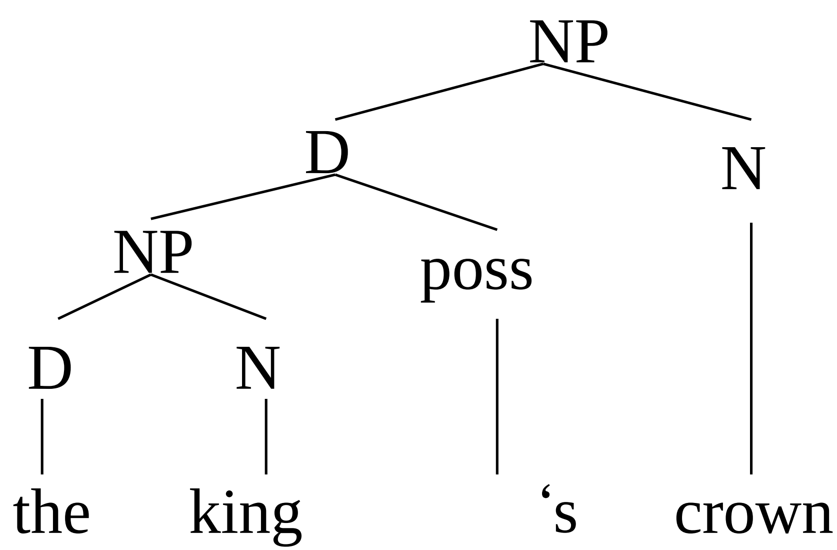
Figure 3.3 Recursive PS rules can generate sentences within sentences.

Loops

NP→ NP Conj NP
NP → D N
D → [the]
N → [dog];[cat]



NP → D N
D → NP poss
NP → D N
D → [the]
N → [king];[crown]
poss → [‘s]



two structures in English where a constituent begins with a constituent of the same type.

Agreement

- In English, verb and its subject must agree in number
- To implement this, we attach arguments to NP and VP

  N(singular) →[dog];[cat];[mouse].

  N(plural) →[dogs];[cats];[mice].

  V(singular) →[chases];[sees].

  V(plural) →[chase];[see].

- expressed in number(in Prolog):

  NP(Number) → D, N(Number).

  VP(Number) → V(Number), NP(Number).

  S → NP(Number), VP(Number)

Case Marking

- some English pronouns are marked for case.
  - He sees him. *Him sees he.
  - She sees her. *Her sees she.
  - They see them. *Them see they.
- The forms that come before verb (he, she, they) are called nominative (主格的) and the forms that come after are called accusative.

  $S \rightarrow NP\ VP$ introduces a nominative

  $VP \rightarrow V\ NP$ introduces an accusative
- Case Marking in CFG

```
pronoun(singular, nominative) -->[he];[she].
pronoun(singular, accusative) -->[him];[her].
pronoun(plural, nominative) --> [they].
pronoun(plural, accusative) -- > [them].
```

and the NP rules

```
np(Number, Case) --> pronoun(Number, Case).
np(Number, _) --> d, n(Number).
s --> np(Number, nominative), vp(Number).
vp(Number) -->v(Number), np(_, accusative).
```

Subcategorization

- VP requires subcategorization

Verb	Complement
sleep, bark	None
chase, see	One NP
give, sell	Two NPs
say, claim	sentence

- We cannot represent them using a single rule such as:

  $VP \rightarrow V(NP)(NP)(S)$

- Instead we need 4 rules + a way to associate the right rule to each verb.

  $VP \rightarrow V$

  $VP \rightarrow V\ NP$

  $VP \rightarrow V\ NP\ NP$

  $VP \rightarrow V\ S$

Subcategorization - one solution(bad)

- separate the verbs

$vp \rightarrow v(1).$

$vp \rightarrow v(2), np.$

$vp \rightarrow v(3), np, np.$

$vp \rightarrow v(4), s.$

$v(1) \rightarrow$ [barked];[slept].

$v(2) \rightarrow$ [chased];[saw].

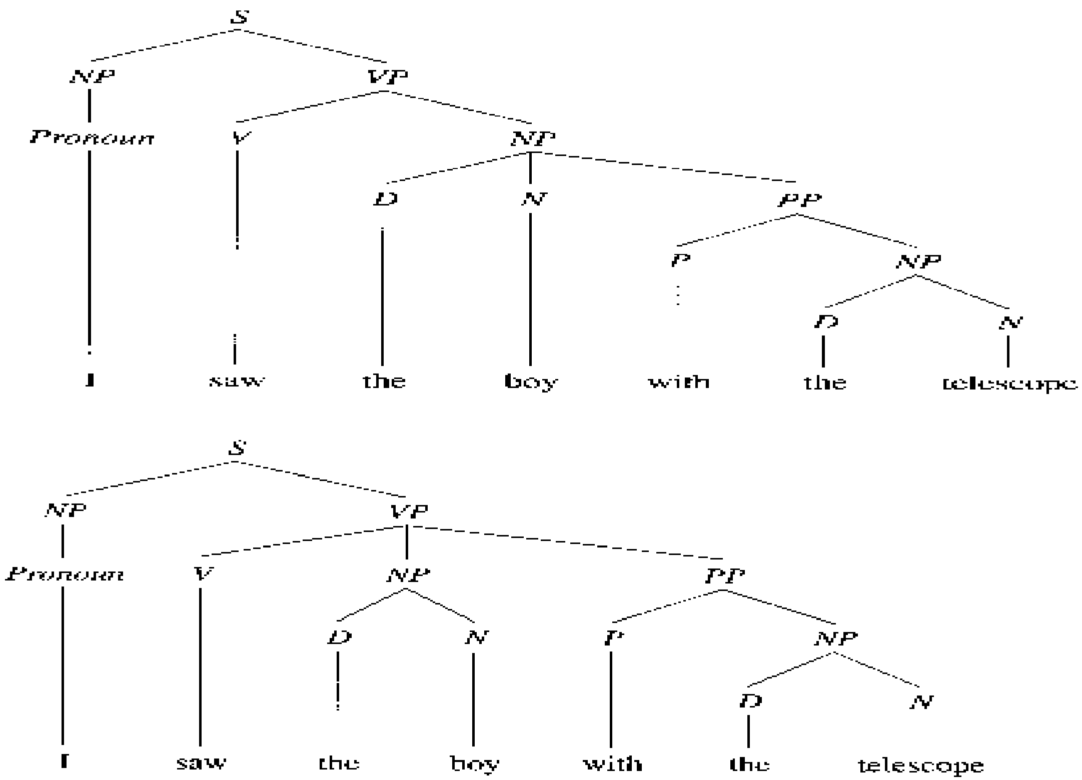
$v(3) \rightarrow$ [gave];[sold].

$v(4) \rightarrow$ [said];[thought].

- Problem : loss of generalization the common property of verbs cannot be expressed. This problem will be solved using **feature unification**.

Structure Ambiguity - a common language phenomenon

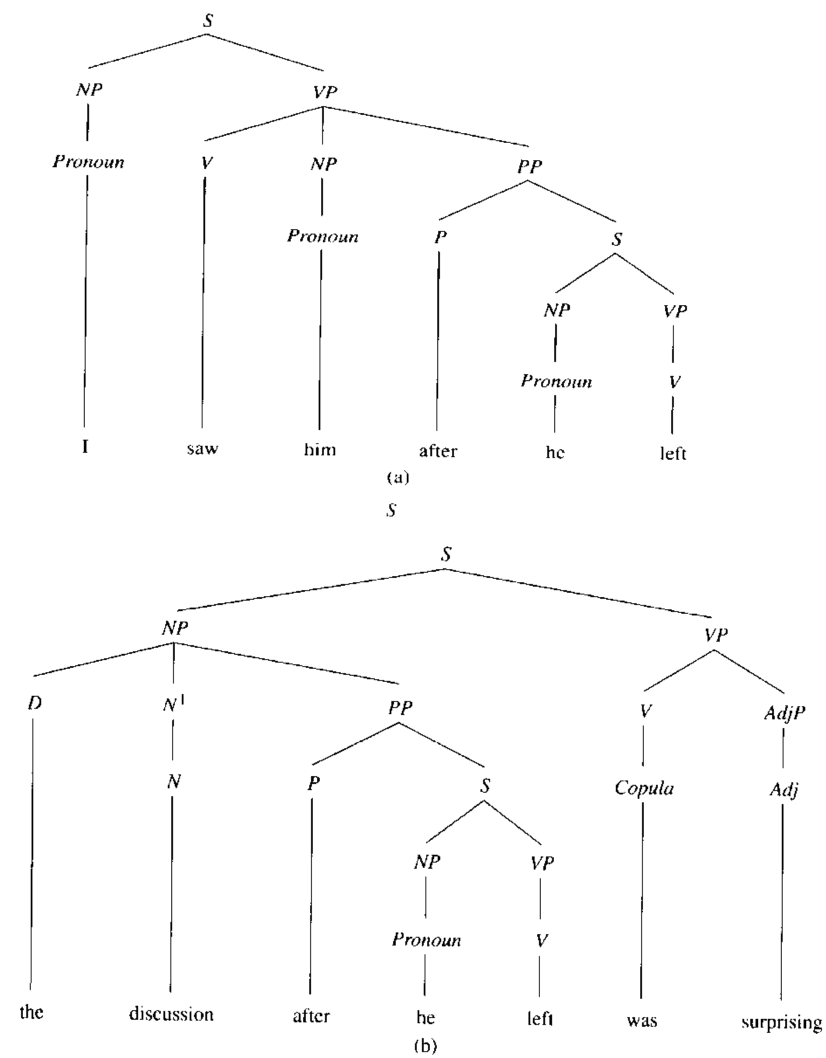
I saw the boy with the telescope



This sentence has 2 meanings

Sentential PP

- subordinating conjunctions
  - before, after, when, whenever, because
- allow a sentence to modify some part of another sentence
  - I saw him after [$_s$ *he left*]
  - The discussion after [$_s$ *he left*] was surprising
- subordinating conjunctions are prepositions



Stochastic Language Models

N-Gram

What are the most common English words?

1	the	article	69975	69792.94
2	be	verb	39175	39109.95
3	of	prep.	36432	35786.01
4	and	co. conj.	28872	28821.11
5	a	article	23073	22984.95
6	in	prep.	20870	20685.17
7	he	pers. pro.	19427	17280.77
8	to	inf. mark.	15025	14990.82
9	have	verb	12458	12192.06
10	to	prep.	11165	11129.57
11	it	pronoun	10942	10836.51
12	for	prep.	8996	8899.55
13	I	pers. pro.	8387	6885.48
14	they	pers. pro.	8284	8162.08
15	with	prep.	7286	7267.37
16	not	neg. adv.	6976	6739.48
17	that	sub. conj.	6468	6373.68
18	on	prep.	6183	6151.18
19	she	pers. pro.	6039	4378.51
20	as	sub. conj.	6029	5982.09
21	at	prep.	5377	5317.20
22	by	prep.	5246	5066.04
23	this	sing. det.	5145	5064.57
24	we	pers. pro.	4865	4699.87
25	you	pers. pro.	4620	3644.51
26	from	prep.	4371	4358.51
27	do	verb	4367	4141.96
28	but	co. conj.	4226	4123.17
29	or	co. conj.	4204	4064.58
30	an	article	3727	3695.88
31	which	wh-det.	3560	3435.25
32	would	modal aux.	3062	2896.00
33	say	verb	2765	2331.97
34	all	pre-quant.	2758	2733.03
35	one	card. num.	2737	2714.29
36	will	modal aux.	2686	2579.81

Entropy of a Symbol

- By definition

For a symbol of probability P_i $H_i = -P_i \log_2 P_i$

For a complete system H=-$\sum_{i=1}^{N} P_i \log_2 P_i$

What is entropy

- A measure of *randomness*
  - Entropy for a total random system $H=\log_2 N$, where N is the number of items in the system.
  - Maximum entropy refers to total randomness.
- A measure of *amount of information*
  - the more significant the meaning(more information, *heavier* information) the larger the entropy of a symbol.
- A system can be compressed to *H* bits of data by using variable code length Huffman coding scheme.
  - **Huffman code length=*H***

Redundancy

- Redundancy measures the '***structureness***' of a system.
- The higher the redundancy, the less amount of information provided.

$$R = 1 - \frac{H}{H_{\max}} = 1 + \frac{\sum_{i=1}^{N} P_i \log_2 P_i}{\log_2 N}$$

Mutual Information

- A measure of how one symbol is associated with another symbol in occurrence.
- Let *H(a,b)* be the entropy of symbols a and b together, *H(a)* and *H(b)* are the entropies of a and b respectively, mutual information is defined as:
  - *MI=H(a,b)-H(a)-H(b)*
- The MI is the **loss of information** when a and b appear together.
- The MI for N items can be written as

$$MI = H(1,2,\ldots N) - \sum_{i=1}^{N} H(i)$$

N-grams

- N-grams are segments of N consecutive items in a string of language symbols
- We can have:
  - alphabet N-gram-such as ‘word’ has *wo, or, rd* 3 bigrams, ‘university’ has *uni, niv, ive, ver, ers, rsi, sit, ity* trigrams
  - word N-gram- in the first sentence, the word bigrams are: N-gram are, are segments, segments of, of N ...
  - Part-of-speech N-gram
  - many others …
- Generally, there are L-N n-grams, L=string length.
- Sometimes, the N-gram, MI are refereed as ***Language Model***.

N-grams

- For word string W, P(W) can be decomposed as

$$
\begin{aligned}
P(W) &= P(w_1, w_2, ..., w_n) \\
&= P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_1, w_2)...P(w_n \mid w_1, w_2, ..., w_{n-1}) \\
&= \prod_{i=1}^{n} P(w_i \mid w_1, w_2, ..., w_{i-1})
\end{aligned}
$$

- For a trigram model, the trigram can be estimated by observing the frequencies or counts of the word pair $C(w_{i-2}, w_{i-1})$ and triplet $C(w_{i-2}, w_{i-1}, w_i)$ as follows:

$$
P(w_i \mid w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})}
$$

Some NLP Applications

- Machine Translation – Translation between two natural languages.
  - See the Babel Fish translations system on Alta Vista.
- Information Retrieval – Web search (uni-lingual or multi-lingual).
- Query Answering/Dialogue – Natural language interface with a database system, or a dialogue system.
- Report Generation – Generation of reports such as weather reports.
- Some Small Applications –
  - Grammar Checking, Spell Checking, Spell Corrector

Natural Language Understanding

↓ Words

Morphological Analysis

↓ Morphologically analyzed words *(another step: POS tagging)*

Syntactic Analysis

↓ Syntactic Structure

Semantic Analysis

↓ Context-independent meaning representation

Discourse Processing

↓ Final meaning representation

Natural Language Generation

↓ Meaning representation

Utterance Planning

↓ Meaning representations for sentences

Sentence Planning and Lexical Choice

↓ Syntactic structures of sentences with lexical choices

Sentence Generation

↓ Morphologically analyzed words

Morphological Generation

↓ Words

Discourse

- Discourses are collection of coherent sentences (not arbitrary set of sentences)
- Discourses have also hierarchical structures (similar to sentences)
- **anaphora resolution** -- to resolve referring expression
  - `Mary bought a book for Kelly.` **She** `didn't like` **it**.
    - **She** refers to Mary or Kelly. -- possibly Kelly
    - **It** refers to what -- book.
  - `Mary had to lie for Kelly.` **She** `didn't like` **it**.
- Discourse structure may depend on application.
  - Monologue
  - Dialogue
  - Human-Computer Interaction

Natural Language Generation

- NLG is the process of constructing natural language outputs from non-linguistic inputs.
- NLG can be viewed as the reverse process of NL understanding.
- A NLG system may have two main parts:
  - **Discourse Planner** -- what will be generated. which sentences.
  - **Surface Realizer** -- realizes a sentence from its internal representation.
- **Lexical Selection** -- selecting the correct words describing the concepts.

Machine Translation

- Machine Translation -- converting a text in language A into the corresponding text in language B (or speech).
- Different Machine Translation architectures:
  - interlingua based systems
  - transfer based systems
- How to acquire the required knowledge resources such as mapping rules and bi-lingual dictionary? By hand or acquire them automatically from corpora.
- Example Based Machine Translation acquires the required knowledge (some of it or all of it) from corpora.