UNIT V

5. APPLICATIONS

Communication – Communication as Action – Formal Grammar for a Fragment of English – Syntactic Analysis – Augmented Grammars – Semantic Interpretation – Ambiguity and Disambiguation – Discourse Understanding – Grammar Induction – Probabilistic Language Processing – Probabilistic Language Models – Information Retrieval – Information Extraction – Machine Translation

5.1. Introduction of Communication

Communication is the planned replace of data brought regarding by the making and awareness of **symbols** warn from a collective method of traditional symbols.

What groups' persons separately from a distinct animal is the difficult method of prearranged communication called **language** that allows us to spread most of what we aware regarding the universe.

5.2. Communication as Action

The actions existing to an agent is to make language. This is known as **speech-act**. The general expressions mentioning to every form of communication is

Speaker → Utterance → Hearer

The different types of terms used in speech act are:

- Inform
- Query
- Request
- Acknowledge
- Promise

Basics of Language

A **formal language** is described as a group of strings. Apiece of string is a order of terminating signs is known as **words**.

A **grammar** is a fixed group of regulations that gives a language. The grammar is a group of modify the rules.

Example

Sentence - S

Noun phrase - NP

Verb phrase - VP

These are called **Non Terminal Symbols.**

The Constituent Steps of Conversation

A distinctive conversation event, in which spokesman ‘S’ needs to notify listener ‘L’ regarding proposition ‘P’ by applying words ‘W’, is collected of **7** steps.

1. Intention
2. Generation
3. Synthesis
4. Perception
5. Analysis
6. Disambiguation
7. Incorporation

The bellow figure explains the 7 methods mixed up in communication by applying the example statement: The wumpus is expired.

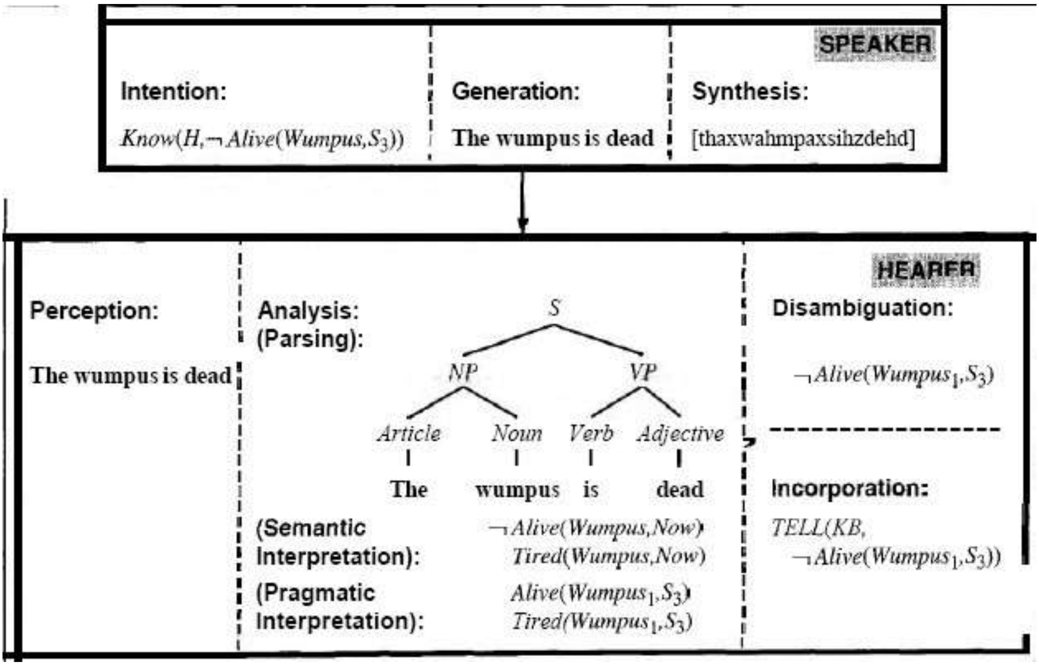


Figure: Seven Processes Involved in Communication

5.3. Proper Grammar for a Portion of English

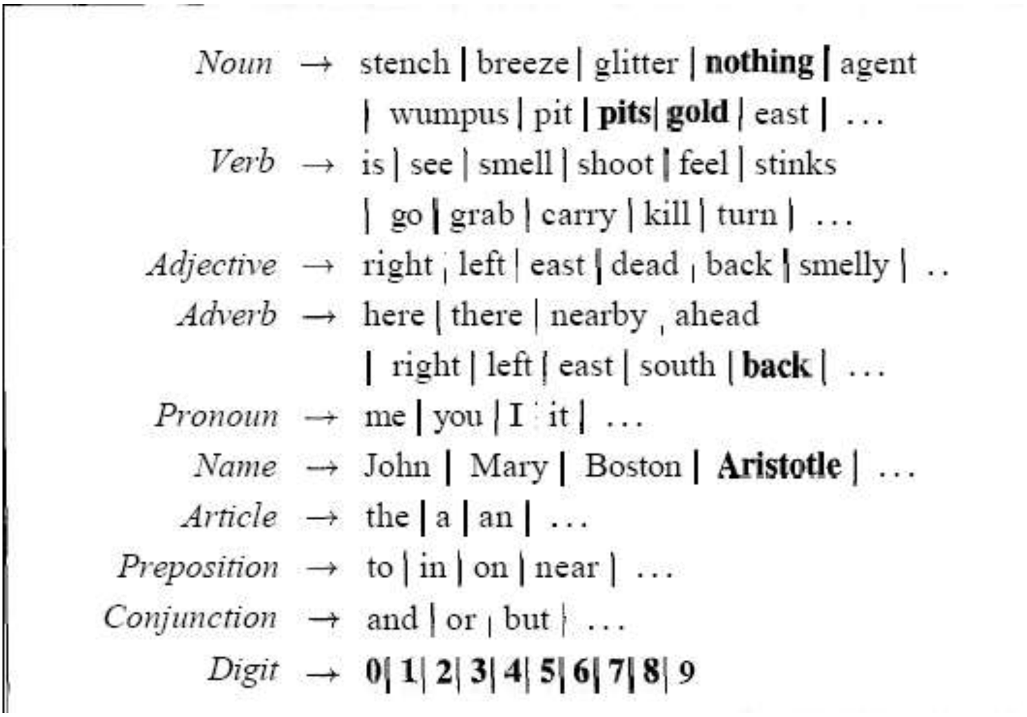
A proper grammar for a portion of English to create syntaxes regarding the wumpus world is build and this language. $\mathcal{E}_0$

The Lexicon of: $\mathcal{E}_0$

A index of permissible words, collected into groups such as

- Nouns
- Pronouns
- Verbs to indicate events
- Names to things
- Adverbs to change verbs
- Adjectives to change nouns
- Preposition
- Conjunction
- Articles

The following figure shows a small lexicon.



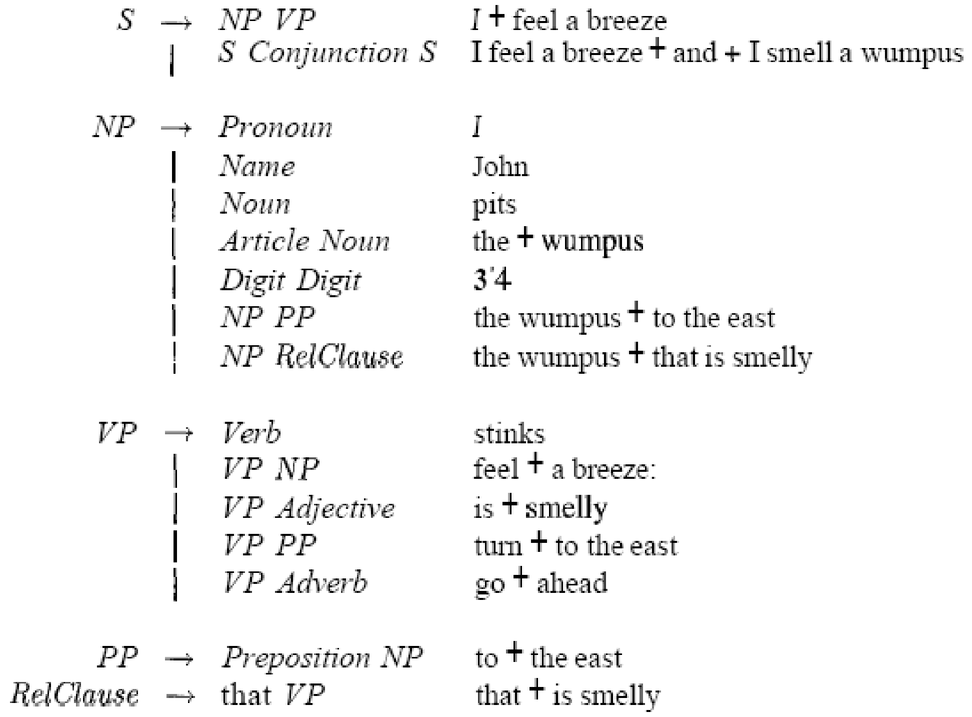
Noun → stench | breeze | glitter | **nothing** | agent
| wumpus | pit | **pits** | **gold** | east | . . .
Verb → is | see | smell | shoot | feel | stinks
| go | grab | carry | kill | turn | . . .
Adjective → right | left | east | dead | back | smelly | . .
Adverb → here | there | nearby | ahead
| right | left | east | south | **back** | . . .
Pronoun → me | you | I | it | . . .
Name → John | Mary | Boston | **Aristotle** | . . .
Article → the | a | an | . . .
Preposition → to | in | on | near | . . .
Conjunction → and | or | but | . . .
Digit → **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | 9

The Grammar of $\mathcal{E}_0$.

It defines how to combine the word and phrases, using 5 non-terminal symbols. The dissimilar types of expressions are:

- Sentence
- Verb phrase(VB)
- Noun phrase (NP)
- Relative clause(Reclause)
- Prepositional phrase(PP)

The figure explains a grammar for $\mathcal{E}_0$.



S	→	*NP VP*	I + feel a breeze
	|	*S Conjunction S*	I feel a breeze + and + I smell a wumpus
NP	→	*Pronoun*	I
	|	*Name*	John
	|	*Noun*	pits
	|	*Article Noun*	the + wumpus
	|	*Digit Digit*	3'4
	|	*NP PP*	the wumpus + to the east
	|	*NP RelClause*	the wumpus + that is smelly
VP	→	*Verb*	stinks
	|	*VP NP*	feel + a breeze:
	|	*VP Adjective*	is + smelly
	|	*VP PP*	turn + to the east
	|	*VP Adverb*	go + ahead
PP	→	*Preposition NP*	to + the east
RelClause	→	that *VP*	that + is smelly

Figure: The Grammar for $\mathcal{E}_0$.

5.4. Syntactic Analysis (Parsing)

Syntactic analysis is the phase in which an input statement is exchanged into a hierarchical-structure that communicates to the elements of meaning in the statement. This procedure is known as **parsing**.

Parsing should be observed as a procedure of discovering for a parse-tree. There are 2 methods for identifying the search-space.

1. Top-down parsing
2. Bottom-up parsing

1. Top-down Parsing

Initiate with the begin symbol and execute the grammar rules forward up to the symbols at the ends of the tree communicate to the elements of the statement initiating parsed.

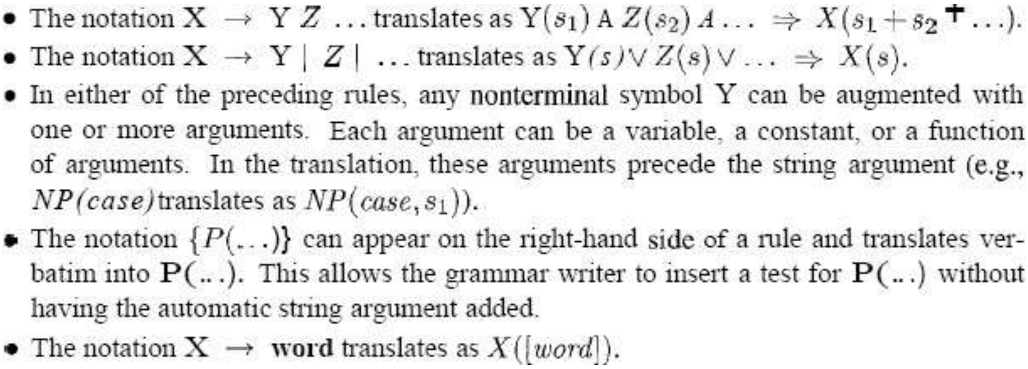
2. Bottom-up Parsing

Initiate with the statement to be parsed and execute the grammar rules backward up to an individual tree whose ends are the words of the statement and whose top node is the institute symbol has been formed.

5.5. Augmented Grammars

The procedure of expanding the presented rules of the grammar in place of initiating latest rules. This formality for expansion is known as **definite clause grammar** or **DCG**. The major advantage of DCG is which we should expand the group signs with extra expand.

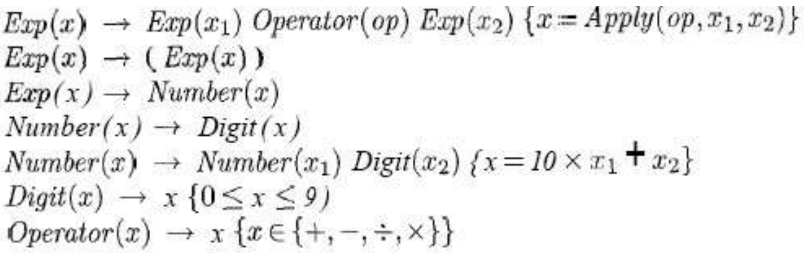
We define DCG as follows:



- The notation X $\rightarrow$ Y Z ... translates as $\mathrm{Y}(s_1)$ A $Z(s_2)$ $A \ldots \Rightarrow X(s_1 + s_2 + \ldots)$.
- The notation X $\rightarrow$ Y | Z | ... translates as $\mathrm{Y}(s) \vee Z(s) \vee \ldots \Rightarrow X(s)$.
- In either of the preceding rules, any nonterminal symbol Y can be augmented with one or more arguments. Each argument can be a variable, a constant, or a function of arguments. In the translation, these arguments precede the string argument (e.g., $NP(case)$ translates as $NP(case, s_1)$).
- The notation $\{P(\ldots)\}$ can appear on the right-hand side of a rule and translates verbatim into $\mathbf{P}(\ldots)$. This allows the grammar writer to insert a test for $\mathbf{P}(\ldots)$ without having the automatic string argument added.
- The notation X $\rightarrow$ **word** translates as $X([word])$.

5.6. Semantic Interpretation

Semantics – the origin of the meaning of statements. Semantic interpretation is the procedure linked an FOL statement with an expression.



$$
\begin{array}{l}
Exp(x) \rightarrow Exp(x_1)\ Operator(op)\ Exp(x_2)\ \{x = Apply(op, x_1, x_2)\} \\
Exp(x) \rightarrow (\ Exp(x)\ ) \\
Exp(x) \rightarrow Number(x) \\
Number(x) \rightarrow Digit(x) \\
Number(x) \rightarrow Number(x_1)\ Digit(x_2)\ \{x = 10 \times x_1 + x_2\} \\
Digit(x) \rightarrow x\ \{0 \leq x \leq 9\} \\
Operator(x) \rightarrow x\ \{x \in \{+, -, \div, \times\}\}
\end{array}
$$

Figure: A Grammar for Arithmetic Operations, Augmented with Semantics

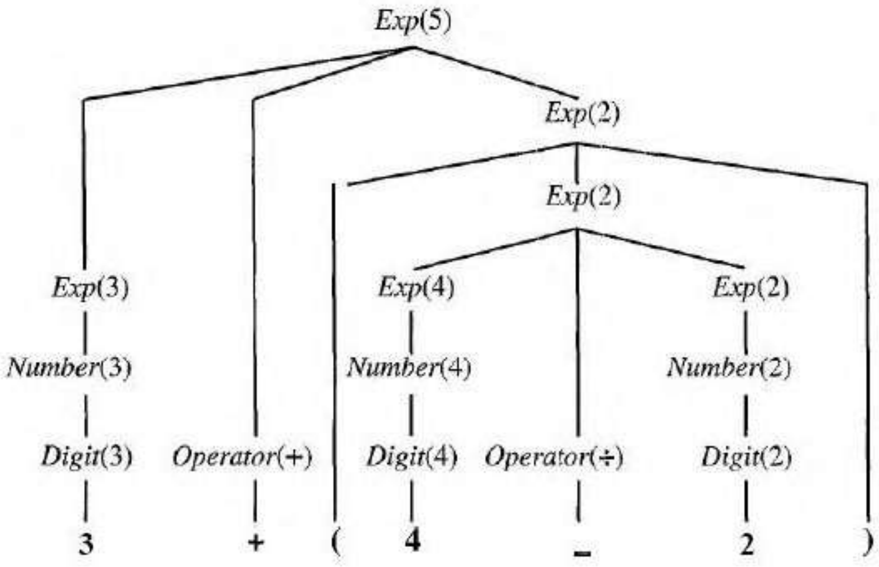


Figure: Parse-tree with Semantic-interpretations for the Expression 3 + (4 / 2)

The Semantics of an English Section

$S(rel(obj)) \rightarrow NP(obj)\ VP(rel)$
$VP(rel(obj)) \rightarrow Verb(rel)\ NP(obj)$
$NP(obj) \rightarrow Name(obj)$

$Name(John) \rightarrow$ **John**
$Name(Mary) \rightarrow$ **Mary**
$Verb(\lambda y\ Ax\ Loves(x,y)) \rightarrow$ **loves**

Figure: A grammar that should obtain a parse-tree and semantic-interpretation for "John loves Mary".

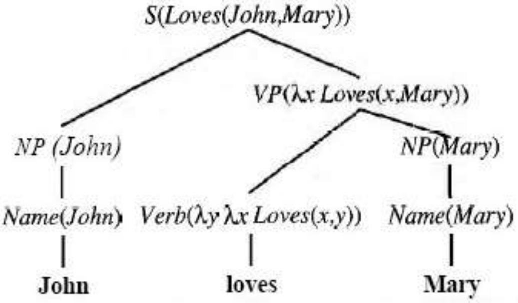


Figure: A parse-tree with semantic interpretations for the string: "John loves Mary"".

5.7. Ambiguity and Disambiguation

A statement, phrase or word is ambiguous, if it contains greater than 1 sense. Categories of ambiguity:

1. **Lexical ambiguity**: in which a statement contains greater than 1 sense. It is absolutely general;
2. **Syntactic ambiguity**: should happen with or without lexical-ambiguity.
3. **Semantic ambiguity**: The Syntactic-ambiguity directs to a semantic-ambiguity, for the reason that 1 translate signifies that the wumpus is in 2.2 and diversely that a stink is in 2.2. In this instance, receiving the incorrect interpretation should be a dangerous fault. It should happen still in phrases without syntactic or lexical ambiguity.

A **metonymy** is a form of talking in which one entity is utilized to set from one more.

A **metaphor** is a form of talking in which an expression with 1 truthful definition is utilized to propose s dissimilar meaning ay method of likeness.

Disambiguation

Disambiguation is a question of diagnosis. It is probable by joining proof, utilizing all the methods from uncertain reasoning and knowledge representation.

We should divide the knowledge down into **4** models:

1. Acoustic model
2. Mental model
3. World model
4. Language model

5.8. Discourse Understanding

A discourse is some statement of words – frequently one that is greater than a single statement extended. Text books, conversations, climate reports and novels are each discourse. We will look at two particular sub problems:

1. Reference Resolution

It is the explanation of a definite noun or a pronoun expression, which mention to an object in the universe. The resolution is the earlier parts of the discourse and based on knowledge of the universe.

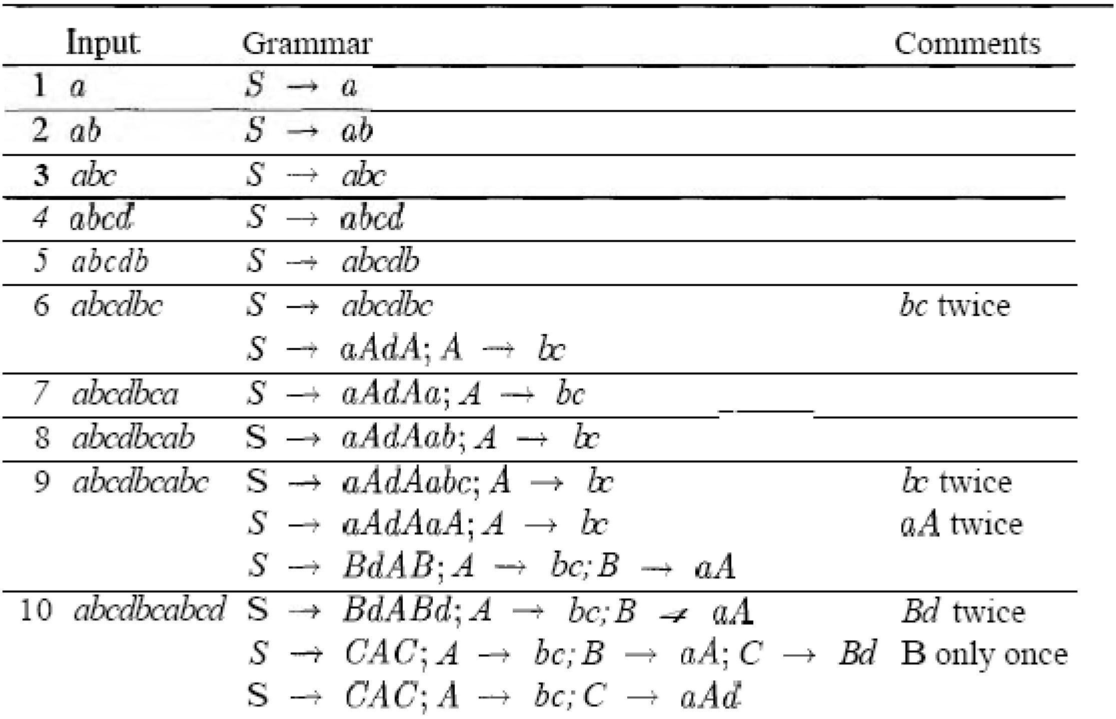
2. *Structure of Coherent Discourse*

A discourse contains structure over the level of statement. The grammar states that a discourse is collected of sections which every section is a collection of statements or either a statement and where sections are collected by coherence relations.

Grammar Induction

It is the work of studying a grammar from records. It makes a grammar of a particular from a grammar, which creates just a single statement namely the original text.

Figure displays the algorithm in process of the text:abcdbcabcd



Input		Grammar	Comments
1	*a*	$S \rightarrow a$	
2	*ab*	$S \rightarrow ab$	
3	*abc*	$S \rightarrow abc$	
4	*abcd*	$S \rightarrow abcd$	
5	*abcdb*	$S \rightarrow abcdb$	
6	*abcdbc*	$S \rightarrow abcdbc$	*bc* twice
		$S \rightarrow aAdA; A \rightarrow bc$	
7	*abcdbca*	$S \rightarrow aAdAa; A \rightarrow bc$	
8	*abcdbcab*	$S \rightarrow aAdAab; A \rightarrow bc$	
9	*abcdbcabc*	$S \rightarrow aAdAabc; A \rightarrow bc$	*bc* twice
		$S \rightarrow aAdAaA; A \rightarrow bc$	*aA* twice
		$S \rightarrow BdAB; A \rightarrow bc; B \rightarrow aA$	
10	*abcdbcabcd*	$S \rightarrow BdABd; A \rightarrow bc; B \rightarrow aA$	*Bd* twice
		$S \rightarrow CAC; A \rightarrow bc; B \rightarrow aA; C \rightarrow Bd$	B only once
		$S \rightarrow CAC; A \rightarrow bc; C \rightarrow aAd$	

Figure: The Algorithm in Operation in the Text “abcdbcabcd”

Probabilistic Language Processing

Introduction

Probabilistic language models deployed on ‘n’ grams improve an unexpected quantity of data regarding a language.

In probabilistic-language model, this should learn from data. It is simpler than **Define Clause Grammar** (**DCG**). It has three specific asks.

1. Information extraction
2. Information retrieval
3. Machine translation

5.9. Probabilistic Language Models

A Probabilistic-language model describes a probability division above a group of strings.

- It can be conveniently trained from records.
- Learning is only counting number of incidences.
- It is robust
- It is applied for disambiguation probability should be utilized to select the most probable understanding.

E.g. Trigram and Bigram language model applied in speech identification.

A **unigram method** allocates a probability "P(w)" to every word in the word list. This method imagines that words are selected in separately, so the probability of a string is only the creation of the probability of words.

A **bigram method** allocates a probability $P(w_i|w_{i-1})$ to every word, specified the earlier word.

In common an n-gram method states on the earlier "n-1" words, allocates a probability for $P(w_i|w_{i-(n-1)} \ldots w_{i-1})$. We need someway of rectifying above the nil counts. The easiest method to perform this is known as AddOne smoothing; we insert a single to the count of each probable bigram.

One more method is **linear-interpolation-smoothing**, that joins bigram, unigram, and trigram methods by limited interruption.

Evaluating a Language Model

Split the information into training information and test information. Verify the restrictions of the model from the training information. Next compute the probability allotted to test information by the model. If the probability is bigger, then it is improved.

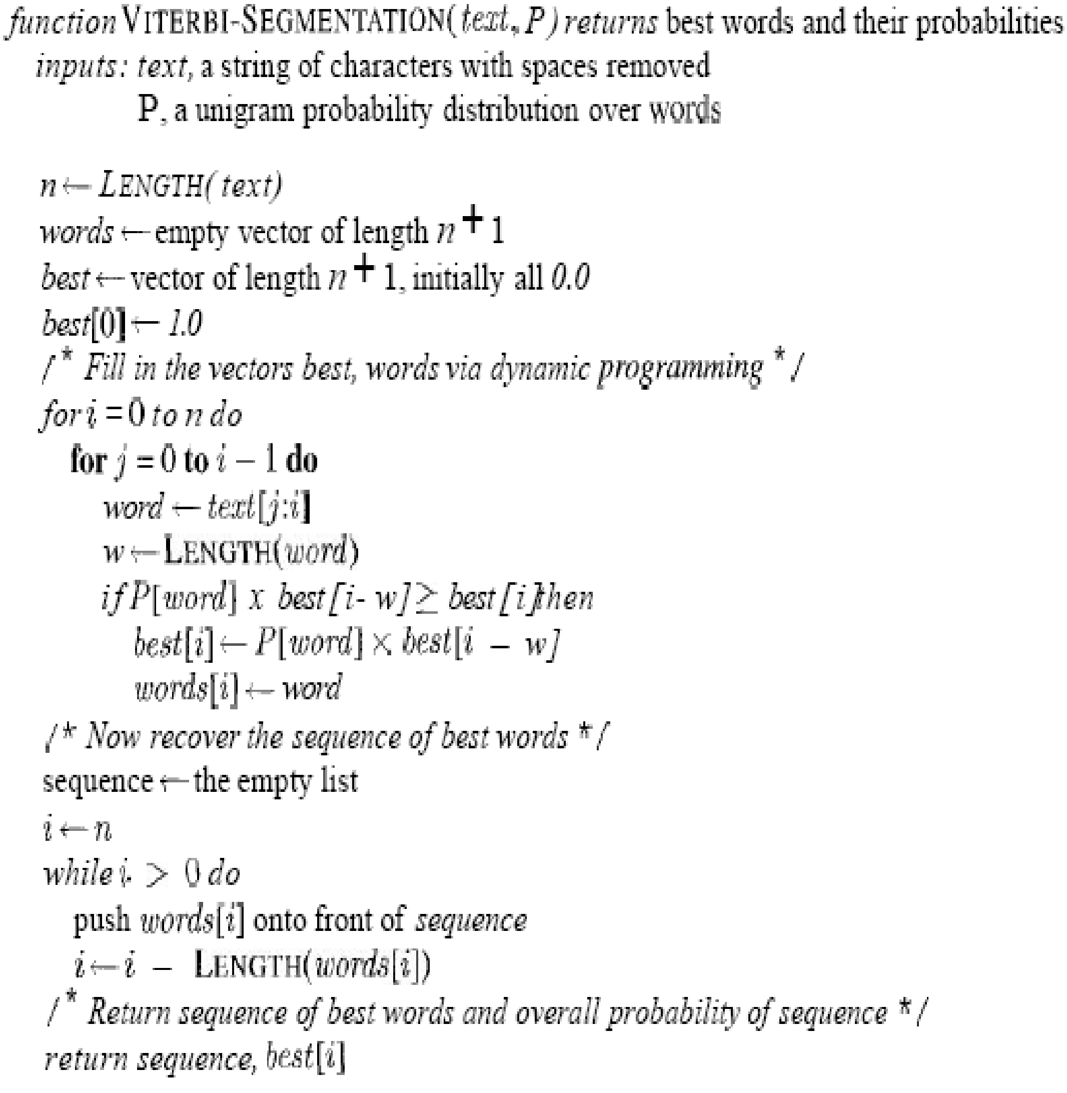
The second approach from evaluating a model is by computing the confusion of the model on test string of words.

Segmentation: It is used to find word margins in a text without spaces.

It is used to read words without spaces.

It is easy to read by human because they have full awareness of pragmatics, semantics and English sentences.

It is done by Viter bi algorithm specifically designed for segmentation problem.



```
function VITERBI-SEGMENTATION(text, P) returns best words and their probabilities
  inputs: text, a string of characters with spaces removed
          P, a unigram probability distribution over words

  n ← LENGTH(text)
  words ← empty vector of length n + 1
  best ← vector of length n + 1, initially all 0.0
  best[0] ← 1.0
  /* Fill in the vectors best, words via dynamic programming */
  for i = 0 to n do
    for j = 0 to i − 1 do
      word ← text[j:i]
      w ← LENGTH(word)
      if P[word] x best[i- w] ≥ best[i] then
        best[i] ← P[word] × best[i − w]
        words[i] ← word
  /* Now recover the sequence of best words */
  sequence ← the empty list
  i ← n
  while i > 0 do
    push words[i] onto front of sequence
    i ← i − LENGTH(words[i])
  /* Return sequence of best words and overall probability of sequence */
  return sequence, best[i]
```

Figure: A Viterbi-based Word Segmentation Algorithms

Probabilistic Context-free Grammars (PCFG)

Rams method get benefit of co-occurrences character in the corpra, but they does not contain notion of grammar at lengths larger than ‘n’. PCFG contains of a CFG where every rule contain a connected probability.

The addition of probability of all rules is 1.

Learning Probability from PCFG

To create a PCFG, we have to combine probability from each CFG rule.

This proposes that learning the grammar from records could be improved than data-engineering method.

Two types of data are given:

1. Parsed
2. Unparsed

The 'E' step approximates the probabilities, which every subsequence is created by all rules. The 'M' step next approximates the probability of all rules. It is done by Inside-Outside algorithm.

Inside-Outside Algorithm

It induces grammar from unparsed tree. It is slow.

Learning Rule Structure from PCFG

If the grammar rules structure is not recognized, then make use of **Chomsky Normal Form (CNF)**.

X → Yz
X → t
Where t is a terminal and X, Y, Z is non terminals.

5.10. Information Retrieval (IR)

It is a work of selecting documents that are related to user's requirement for data.

E.g. Google, Yahoo etc.

An IR is characterized as

A. A manuscript gathering

B. Query in a query language

C. A result set

D. A demonstration of the result set.

The initial IR systems performed on a Boolean word type. Every word in this manuscript gathering is considered as a Boolean attribute, which is true of a manuscript if the word take places in the manuscript and false if it is not.

Evaluating IR Model

Here two measures to assess whether the IR method is performing well or not. They are:

1. Recall
2. Precision

Recall: is the percentage of all the related manuscripts in the gathering that are in the outcome.

Precision: is the percentage of manuscripts in the outcome set that are truly related.

Other Measures to Evaluate IR

The other measures are

1. Reciprocal rank.
2. Time to answer

IR Refinements

The uni-gram method considers all words as totally autonomous, but we recognize that a few words are associated.

E.g. the word -Couch|| has two closely related words

Couches

Sofa

So IR systems do refinements from these types of word by the following methods.

1. Case folding
2. Streaming
3. Recognize synonyms
4. Metadata

Presentation of Result Sets

There are three mechanisms to attain performance development. They are

1. Document classification
2. Relevance feedback
3. Document clustering
    - Agglomerative clustering
    - K-means clustering

Implementing IR Systems

IR systems are made efficient by two data structures. They are

1. Lexicon:
   - It indexes each and every word in the manuscript gathering.
   - It supports one operation
   - It is implemented by hash table.
2. Inverted index:
   - It is similar to the index at back side of a book. It consists of a set of hit lists, which is the place where the word occurs.
   - In uni-gram method, it is an index of pairs.

5.11. Information Extraction (IE)

IE is the procedure of generating database entries by removing content and viewing for connections between those events and objects and for incidents of a specific class of event or object.

E.g. To extract addresses from WebPages with database fields as road, village, pin code and state.

Categories of IE System

1. Attribute based System

This system imagines that the total text as an individual object and attempt to remove attributes of that object.

If standard expression equals the text closely once, next that section of text is extracted. It is the value of the attribute.

If there is no equal, not anything will happen.

2. Relational based System

It observes the connections between them and greater than one object.

This system is made by applying cascaded limited state generators, i.e. it contains of a sequence of Finite State Automata (FSA).

An example of this system is FASTUS.

It contains of 5 phases. They are:

A. Tokenization
B. Complex word handling
C. Basic groups
D. Complex phrases
E. Merger structures

5.12. Machine Translation

It is the automated translation of text from one language (source) to another (target).

Types of Translation

1. Rough translation
2. Restricted source translation
3. Pre-edited translation
4. Literary translation

Machine Translation System

If translation is to be done fluently and perfectly then the interpreter (machine or human) should study the genuine text, be aware of the condition to which it is mentioned and discovers a better equivalent text in the target language reporting same or similar situation.

These systems differ in the stage to which they evaluate the text.

Statistical Machine Translation

It is a new approach proposed during last decade. Here the whole translation process is based on discovering the best possible translation of a statement, applying records collected from ‘s’ bilingual process.

The method for p(F/E) has 4 set of limitations.

1. Translation method
2. Language method
3. Fertility method
4. Word choice method
5. Offset method

Learning Probabilities for Machine-translation

- Align sentences
- Segment into sentence
- Approximate the French language model
- Approximate the primary fertility model
- Approximate the primary choice model
- Approximate the primary offset model
- Improve estimates

Question Bank

Unit - V

Part - A

1. What is meant by communication?
2. Define utterance.
3. Define parsing.
4. List the types of grammar with its rules.
5. Give example for open classes and closed classes.
6. Differentiate bottom up parsing with top-down parsing.
7. What is meant by chart parsers?
8. Give example for chart parsing systems.
9. What is meant by define clause grammar.
10. Give example for augmented grammar.
11. List the types of ambiguity with example.
12. What is meant by probabilistic language model?
13. Define smoothing? What are the types of smoothing?
14. Differentiate information retrieval with information extraction.
15. List the stages of FASTOS.
16. What is meant by language model and translation model.
17. Write the steps for K means clustering.
18. Differentiate document classification with document clustering.
19. What is meant by stemming.
20. Define interlingua.

Part - B

1. Explain the seven processes involved in communication with one example.
2. Explain the two kinds of parsing with suitable example.
3. Explain the chart parsing algorithm with suitable example.
4. How semantic interpretation is done for a sentence? Explain with an example.
5. Write short notes on:
   - (a) Discourse understanding.
   - (b) Grammar induction
6. Explain the segmentation algorithm in detail.
7. Explain the steps to build an IR system.
8. List the different methods to do machine translation. Explain in detail about the statistical machine translation.
9. How IE is done using FASTOS system, explain the stages of it.