PAGE NO. 22nd April, 202
DATE:

Chapter 9 Context-Free Grammars for English

~~Syntax~~ S → NP VP

Syntax: (NP, VP — constituents)

English - SVO
Nepali - SOV

Transitive verbs - direct objects. eg I played football.
Intransitive verbs: eg I cried.
(no direct objects) I laughed

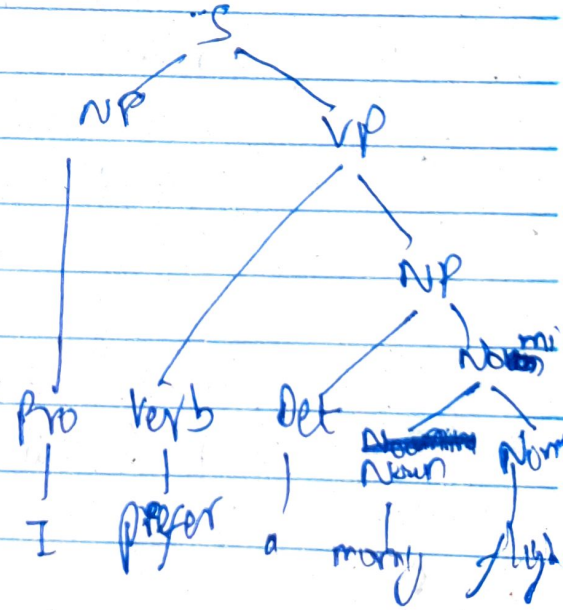
Nepali Language - Dependency grammar followed instead of CFG.
Panini Grammar

CFG also called PSG (Phase Structured Grammar)

→ Grammar for Lo

S → NP VP
NP → Pronoun | Proper-Noun | Det Nominal
Nominal → Noun Nominal | Noun
VP → Verb | ~~Verb NP~~
| Verb NP
| Verb NP PP
| Verb PP
PP → Preposition NP

I prefer a morning flight



6:45
7:30

PAGE NO.:
DATE:

- Bracket notation of parse tree

Sentence level construction

1. Imperative structure → S → VP (No subject) eg Show the lowest fare
2. Declarative structure - Subject NP + VP S → NP VP (Subject-NP)
3. Yes-n-o question structure : S → Aux NP VP (Subject-NP)
4. Wh- question structure : S → Wh-NP VP eg What airlines fly from Nepal to Thailand?

eg Do any of these flights have stops? - Yes-no ques

Noun Phrase

- head - central noun in NP
  eg a boring lecture
  ↓ head
- prenominal and postnominal modifiers
- determiners (optional)
- Mass nouns / Material nouns - no determiners
- predeterminers and postdeterminers
  all the flights

Nepali - no preposition, only postposition

Agreement - not confined in CFG anymore
- adds context

English - number agreement only
French / German / Nepali - Gender and number agreement
(Honorifics as well)

1st till POS tagging

PAGE NO.
DATE: 5th April

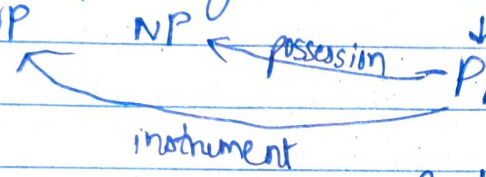
- Parsing

Bottom-up parsing
- CKY algorithm

Top-down
- Earley algorithm

Chart parser
→ mix of top-down and bottom-up

→ I saw the boy with the telescope



Ambiguity

→ I shot the elephant in my pajamas.

Problems with top-down parser - 1) left recursion
2) 1 Attachment ambiguity
   2 Coordination ambiguity
eg [old [men and women]]
   [old men [and women]]

PAGE NO: 28th Apr
DATE:

Earley algorithm
- Predictor
- Scanner
- Completor

$O(n^3)$ n - no of words in the input

→ CKY Parsing (Bottom up)

CNF grammar used i.e.
$A \rightarrow BC$ or 2 non terminals or
$A \rightarrow w$ single terminal

Conversion to CNF
$S \rightarrow ABC$
$S \rightarrow XBC$
$X \rightarrow AB$

Operation (CKY)
1. Recognition
2. Parsing

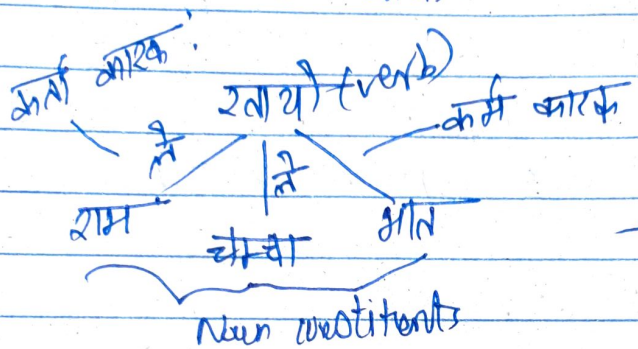
→ POS tagging in Nepali texts
43 tags

Penn Treebank tagset - popular
- 112 forms of single verb

→ Chunker, Grammar analyzer, POS tagger - Nepali

Grammar analyzer
- Dependency grammar formalism
- free word order language - Nepali

रामले भात खायो
भात रामले खायो



→ Karaka relations

PAGE NO. 2nd May
DATE:

Statistical Parsing
Probabilistic (Stochastic CFG (PCFG)

- 4 parameters (N, Σ, R, S)

N - non terminal symbols
Σ - set of terminal symbols disjint from N
R - set of rules or productions, each of the form
$A \to \beta$ [p]; — diff from CFG
A. A - non terminal
β - string of symbols from a set of strings $(\Sigma \cup N)^*$
p no bet^n 0 & 1 expressing $P(\beta|A)$
S - start symbol

$$p = P(A \to \beta)$$
$$= P(A \to \beta | A)$$
$$= P(RHS | LHS)$$

Consider all possible expansions of a non-terminal

$$\sum_{\beta} P(A \to \beta) = 1$$

Prob of parse tree T is

$$P(T, S) = \prod_{i=1}^{n} P(RHS_i | LHS_i)$$

$$P(T, S) = P(T) P(S|T) = P(T) \quad (\because P(S|T) = 1)$$

PAGE NO.:
DATE:

Prob of sentence = sum of prob of parse trees

→ Most probable derivation (parse tree) for a sentence

Yield
Consider all possible parse trees for a sentence S.
The string of words S - yield of parse tree over S.

Disambiguation algo picks the parse tree which is most probable given S

$$\hat{T}(S) = \underset{T \text{ s.t. } S = \text{yield}(T)}{\text{argmax}} P(T|S)$$

$$\hat{T}(S) = \underset{T \text{ s.t. } S = \text{yield}(T)}{\text{argmax}} \frac{P(T,S)}{P(S)}$$

$\hat{}$ - approximate

$$\hat{T}(S) = \underset{T \text{ s.t. } S = \text{yield}(T)}{\text{argmax}} P(T)$$

Prob of next word w_i in a sentence given all words seen so far

$$P(w_i | w_1, w_2 \ldots w_{i-1}) =$$

9-10 12-3 10-

PAGE NO.:
DATE:

→ Probabilistic CKY
→ learning PCFG Rule Prob

- Head words

ground truth
gold standard

Unification grammar (Feature - Structure Grammar)

PAGE NO.:
DATE:

Semantics level

Chapter 14 Representing Meaning

Lexical semantics - Slide

meaning conveyed by a word

art
- wound
- share
- ...
- :

Natural Language Understanding (NLU) - current focus

Natural language generation (NLG)

~~गाड़ीने~~ ~~गीत~~ ~~गाओ~~

world knowledge - KU is located in Dhulikhel

C

PAGE NO.
DATE: 6th May, 20

Lexical Semantics

WordNet - largest lexical database
- contains synonyms set (synset)
Relation among lexemes & their senses

1. Homonym - lexemes that share a form. (phonological, orthograph
eg Bat - wooden same word diff meaning
Bat - animal

Homophones - sound same Homograph
It
Eat

2. Synonym

3. Polysemy - different senses of same word
has floppy ears
has a good ear for jokes.

4. Hyponym / Hypernym

Lemma (stem / root - interchangeable

PAGE NO.:
DATE:

ENGINE

Vector Space (Distributional) Lexical Semantics

Information Retrieval System

The Vector Space Model

Graphic Representation

Term Weights : Term frequency

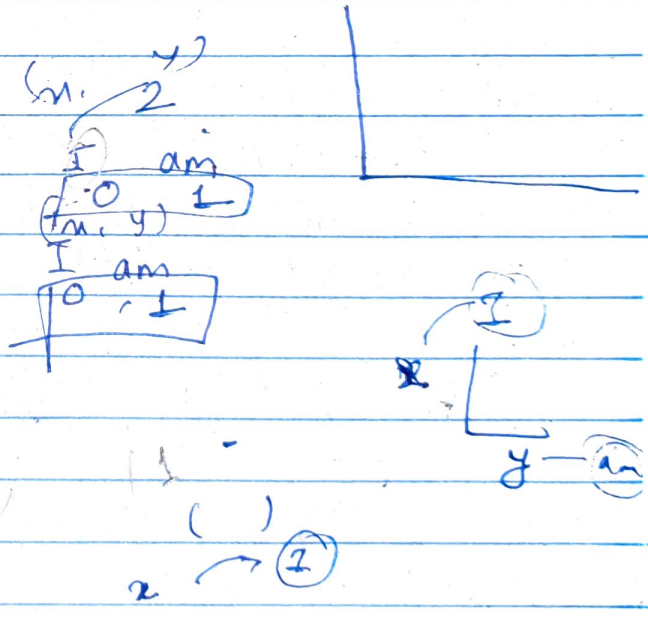
TF-IDF Weighting
Term frequency
Inverse Document frequency

Similarity Measure

Cosine similarity

Gold standard
Ground truth data
Bench-marked datasets

SemEval competition - NLP popular



PAGE NO.:
DATE:

Discourse (highest level of text analysis)
↑
Pragmatics (contextual meaning)
↑
Semantics (word level)
↑
Syntax
↑
Morphology

} document

→ Pragmatics
context
i) Linguistic context (cotext)
ii) Physical context

acts - locutionary, illocutionary, ~~perlocu~~ perlocutionary

13th May

Discourse and Discourse analysis

Doctors delivering diagnosis

Sequence by Sequence

Discourse

Conversation

LAA

PAGE NO.:
DATE: 13th May

Discourse and Discourse Analysis

Te bhako topic ko bachcha chai

OK
REGULT

Text linguistics and discourse analysis

Discourse - both text and speech

Information Retrieval - return a complete document based on query.

Information Extraction - extract part of information about a document eg model name.

Discourse wala course ma ko book chapter

Reference Resolution
Anaphora Resolution
Co-reference Resolution

~~Re~~ Referents ⟷ Named-entity

Referring expressions.

Cataphora vs Anaphora

He . . . John . . . - Cataphora
(John → referent)

* - not well formed sentence. (wrong sentences)

\# Recency

PAGE NO.:
DATE:

Salient → more important / pertinent

Sentiment Analysis

16th May

- Opinion mining
- Polarity :- +ve, -ve, neutral

Product Reviews
- Movie
- Book
- Any other product

Opinion extraction / mining
Sentiment mining
Subjectivity analysis

Sentiment analysis

subject to change

Google trends

- Polarity detection / Classification
- Degree / Intensity of

Baseline Algorithm
- Tokenization
- Feature extraction - (Adjective, Adverb) - Potential features
- Classification using (Naive Bayes, MaxEnt, SVM)

→ Sentiment bearing words may be misleading

So, all words need to be considered
eg " didn't like this movie more
liked this movie

Naive Bayes (Boolean Multinomial)
- for sentiment analysis, word occurrence more important than word frequency.
eg worst movie - -ve
good .. good .. worst - -ve

So, duplicate words removed

Why reviews hard to classify?

- Implicit opinions (opinions - not expressed explicitly)

MPQA Subjectivity Cues Lexicon
Bing Liu Opinion lexicon
Sentiwordnet

too [adj]
too bad - -ve
too good - +ve
too good to be true - -ve (didn't work)

Target

PAGE NO:
DATE: AKS

1 I bought a mobile phone last week.
2 It fits in my pocket. - Size : small
3 It is affordable. : Price : affordable
4 It is cool. : Texture : cool
5 Battery life is just a few hrs. Battery life : short

finding sentiment of a sent
- important for finding aspects or attributes.
eg: size in sent 2 : size = small
~~sent~~ 3: price

→ Detection of friendliness

PAGE NO. 20th May 2
DATE:

Information Retrieval
- search engine (google)

Precision: fraction of retrieved documents that are relevant to user's information need

$$= \frac{TP}{TP + FP}$$

Recall: fraction of relevant documents in collection that are retrieved.

$$\frac{TP}{TP + FN} \qquad TP + FN$$

Rare words more important in collection of documents
~~By~~ Frequent words more important inside a document.

- Boolean search
- Rank retrieval search

PAGE NO. 23rd May
DATE:

Information extraction and Named Entity Recognition

- POS tagging - 1st step
- Phrase level (Named entity tagging)

Unstructured text → structured info

Question Answering

Gold leveled - true
What is the capital of Nepal?
Kathmandu - gold leveled answer.

2nd Text summarization