ABHYASMITRA.IN

COMPREHENSIVE UNIVERSITY STUDY NOTES

NATURAL LANGUAGE PROCESSING

Unit IV — Core NLP Tasks, Applications & Model Deployment

Topics Covered in this Unit

Sequence Labeling Tasks: POS Tagging (HMM, Brill Tagger TBL) & Named Entity Recognition (NER)
Parsing: Constituency vs. Dependency Parsing (Transition-Based Shift-Reduce vs. Graph-Based MST)
Comprehensive NLP Evaluation: Token Accuracy, Entity F1, Attachment Scores (UAS/LAS), BLEU & ROUGE
Core NLP Applications: Sentiment Analysis Pipelines, Document Similarity & Recommendation Systems
Information Retrieval (Vector Space Model & BM25) & Information Extraction (IE) Architecture
Word Sense Disambiguation (WSD): Polysemy Resolution & The Simplified Lesk Algorithm with WordNet
Speech Processing Fundamentals: MFCC Feature Extraction Pipeline & Automatic Speech Recognition (ASR)
Model Deployment & Production: REST APIs, Request-Response Architecture & Hugging Face Inference
Deployment Engineering: Quantization, ONNX, Docker Containerization, Cloud Serving & Ethical AI / GDPR
High-Yield University Exam Question Bank with Detailed Model Answers & Unit Summary

Compiled in a structured, easy-to-understand, textbook style format
Specially curated for B.Tech / B.E. / BCA / MCA / B.Sc Computer Science & Data Science Students

COURSE SYLLABUS & MAPPING

[UNIT IV SYLLABUS — PRESCRIBED UNIVERSITY CURRICULUM]

This section contains the official syllabus for Unit IV. Verify with your university curriculum mapping.

Unit IV Syllabus Breakdown:

Unit IV - NLP Tasks: Sequence Labeling Tasks: POS Tagging (HMM, Brill tagger), Named Entity Recognition (rulebased + statistical view). Parsing: Dependency parsing (transition-based and graph-based concepts), constituency vs. dependency, Parsing errors and evaluation. Evaluation Metrics: Confusion matrix, Precision, Recall, F1, evaluation for tagging and parsing, token-level accuracy, entity-level F1 (NER), Parse accuracy, Attachment score (basic idea), BLEU, ROUGE. Core NLP Applications: Sentiment analysis, Text classification pipeline, document similarity, Recommendation System, Information retrieval: Vector Space Model, Information Extraction using sequence labelling, sentiment analysis, Word Sense Disambiguation (WSD): concept, Lesk algorithm, and applications. Speech Processing: Mel-Frequency Cepstral Coefficients (MFCC), Automatic Speech Recognition (ASR) pipeline. Model Deployment: APIbased NLP Services, Introduction to Model Deployment and Inference, REST APIs, Request–Response Architecture, Overview of Hugging Face Inference APIs, Workflow of NLP Deployment, Introduction to Cloud Platforms, Basic Concepts of Containerization, and Performance Monitoring with Ethical Considerations.

• Course Code: _____ • Semester / Year: _____

TABLE OF CONTENTS

4.1 Sequence Labeling Tasks: POS Tagging & NER

Definition: Sequence Labeling

Sequence Labeling is a fundamental NLP task where a model receives an input sequence of tokens X = (x_1, x_2, ..., x_n) and outputs a corresponding aligned sequence of discrete categorical labels Y = (y_1, y_2, ..., y_n).

1. Part-of-Speech (POS) Tagging Approaches

- **1. Statistical HMM Taggers:** Generative model estimating joint probability P(W, T) using Viterbi decoding over transition P(t_i|t_{i-1}) and emission P(w_i|t_i) probabilities.
- **2. Brill's Tagger (Transformation-Based Learning - TBL):** Combines rule-based and statistical paradigms. (1) Initial State Annotator assigns every word its most frequent tag, (2) Iteratively applies learned Transformation Rules to fix errors (e.g., 'Change tag from NN to VB if preceding word is TO'). Generates human-interpretable rules with compact memory.

2. Named Entity Recognition (NER)

NER locates and classifies named entities in unstructured text into predefined semantic categories: Person (PER), Organization (ORG), Location (LOC), Temporal/Date (DATE), and Monetary values (MONEY).

- **Rule-Based vs. Statistical NER:**
  - Rule-Based: Uses extensive dictionary lookup lists (Gazetteers), regex patterns (e.g., capital letters, honorifics 'Dr.', 'Mr.', company suffixes 'Inc.', 'LLC').
  - Statistical / Neural View: Frames NER as BIO sequence labeling trained on annotated datasets (CoNLL) using Conditional Random Fields (CRFs) or BiLSTM-CRF / BERT models capturing bidirectional contextual dependencies.

4.2 Parsing: Constituency vs. Dependency Parsing

Comparison Dimension	Constituency Parsing (Phrase Structure)	Dependency Parsing (Grammatical Relations)
Core Representation	Nested hierarchical phrase structure constituents (S, NP, VP, PP) based on Context-Free Grammars.	Direct typed binary directed grammatical relations between words (Head -> Dependent).

Comparison Dimension	Constituency Parsing (Phrase Structure)	Dependency Parsing (Grammatical Relations)
Graph Nodes	Both Non-Terminal phrase symbols (NP, VP) and Terminal word tokens.	Word tokens ONLY (No phantom non-terminal phrasal nodes).
Edge Semantics	Parent-child structural containment.	Grammatical relation labels (e.g., nsubj, dobj, amod, prep, det).
Word Order Sensitivity	Rigid word order assumptions (struggles with free word-order languages).	Naturally handles free word-order languages (e.g., Latin, Russian, Hindi, Sanskrit).

Dependency Parsing Algorithms & Evaluation

- **1. Transition-Based Dependency Parsing (Shift-Reduce):** Uses a Stack, an input Buffer, and a set of transitions: SHIFT (push token from buffer to stack), LEFT-ARC (draw dependency edge from top of buffer to top of stack), RIGHT-ARC (draw dependency edge from top of stack to top of buffer). Linear time O(n).
- **2. Graph-Based Dependency Parsing:** Scores all possible directed edges between all word pairs and finds the Maximum Spanning Tree (MST) using the Chu-Liu-Edmonds algorithm. Time: O(n^2).
- **Evaluation Metrics:**
  - Unlabeled Attachment Score (UAS): Percentage of words assigned the correct syntactic head word.
  - Labeled Attachment Score (LAS): Percentage of words assigned both the correct syntactic head and the correct grammatical dependency label.

4.3 Comprehensive Evaluation Metrics in NLP

1. Classification & Sequence Labeling Metrics

- **Token-Level vs. Entity-Level F1 (NER):** Token-level accuracy measures per-word correctness. Entity-level F1 strictly measures exact multi-word span boundary and type matches (e.g., correctly predicting both 'Sundar' and 'Pichai' as [B-PER, I-PER]).

2. Machine Translation & Summarization Metrics: BLEU & ROUGE

- **1. BLEU Score (Bilingual Evaluation Understudy — Papineni et al., 2002):** Evaluates Machine Translation by comparing candidate output C against human reference translations R based on Modified N-Gram Precision penalized for brevity: BLEU = BP * exp(sum_{n=1}^N w_n * ln(p_n)) • Brevity Penalty (BP): BP = 1 if |c| > |r|, else exp(1 - |r| / |c|) (penalizes overly short translations). • p_n: Clipped n-gram precision (n = 1, 2, 3, 4 with equal weights w_n = 0.25).

- **2. ROUGE Score (Recall-Oriented Understudy for Gisting Evaluation — Lin, 2004):** Evaluates Text Summarization by measuring N-Gram Recall against human reference summaries: • ROUGE-N (ROUGE-1, ROUGE-2): Overlap of n-grams between candidate and reference. • ROUGE-L: Based on Longest Common Subsequence (LCS) capturing flexible sentence-level structural flow without requiring exact consecutive matches.

4.4 Core NLP Applications & Semantic Pipelines

1. Sentiment Analysis & Text Classification Pipeline

- **Standard Pipeline:** Raw Text -> Tokenization & Cleaning -> TF-IDF / Embedding Vectorization -> Classifier (Linear SVM / Logistic Regression / BERT) -> Sentiment Polarity (Positive, Negative, Neutral).
- **Lexicon-Based Sentiment (VADER / SentiWordNet):** Rule-based sentiment scoring using curated valence dictionaries with heuristics for punctuation ('!!!'), capitalization ('GREAT'), and negation flip ('not good').

2. Information Retrieval (IR) & Vector Space Model

- **Vector Space Model (Salton):** Documents and queries are represented as high-dimensional TF-IDF vectors in R^|V|. Relevance is computed using Cosine Similarity: sim(q, d) = (q . d) / (||q|| * ||d||).
- **BM25 (Best Matching 25):** Probabilistic ranking function incorporating non-linear term frequency saturation and document length normalization (prevents long verbose documents from dominating search results).

3. Word Sense Disambiguation (WSD) & The Lesk Algorithm

The Simplified Lesk Algorithm (Michael Lesk, 1986)

The Lesk algorithm resolves lexical polysemy (Word Sense Disambiguation) by selecting the dictionary sense whose definition gloss and example sentences share the MAXIMUM word overlap with the context words surrounding the target ambiguous word in the sentence.

Algorithm:

1. For each sense s_k of target word w in WordNet:

Signature(s_k) = Words in definition gloss and examples of s_k

2. OverlapScore(s_k) = | Signature(s_k) intersect Context(w) |

3. Return sense s = argmax_{s_k} OverlapScore(s_k).*

4.5 Speech Processing: MFCC & The ASR Pipeline

1. Mel-Frequency Cepstral Coefficients (MFCC) Extraction

Human auditory perception of sound frequency is non-linear (the ear resolves small pitch differences much better at low frequencies than high frequencies). MFCC transforms speech acoustic signals into perceptually relevant features:

- **1. Pre-emphasis:** High-pass filtering amplifying high frequencies: y[n] = x[n] - alpha * x[n-1] (alpha approx 0.97).
- **2. Framing & Windowing:** Segmenting continuous audio into short stationary frames (20–30 ms) multiplied by a Hamming Window to eliminate spectral leakage at boundary discontinuities.
- **3. Fast Fourier Transform (FFT):** Converts time-domain signal frame into frequency power spectrum.
- **4. Mel-Scale Triangular Filterbank:** Warps frequency scale into Mel scale: Mel(f) = 2595 * log_10(1 + f / 700).
- **5. Discrete Cosine Transform (DCT):** Computes DCT on log-filterbank energies, yielding 12–13 orthogonal cepstral coefficients.

2. Automatic Speech Recognition (ASR) Pipeline

CLASSICAL AUTOMATIC SPEECH RECOGNITION (ASR) ARCHITECTURE

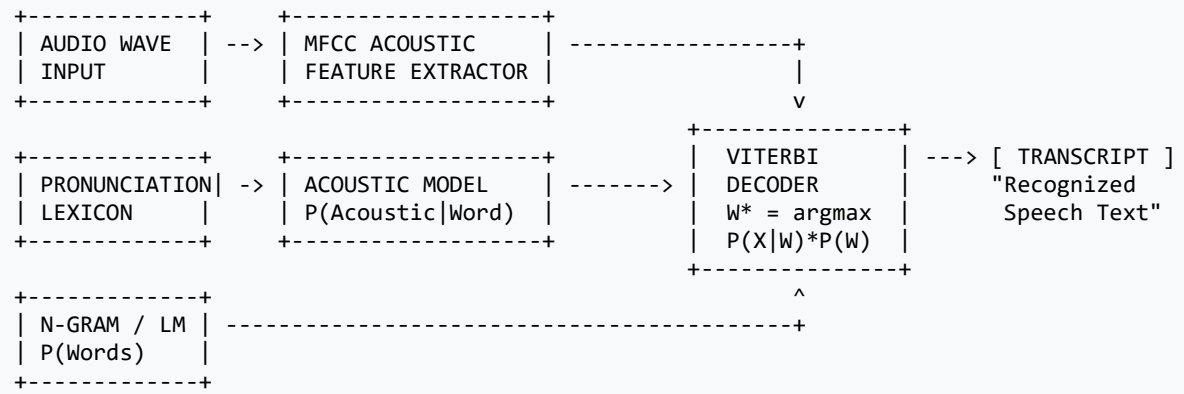

Figure: ASR combines Acoustic Models, Pronunciation Lexicons, and Language Models via Bayesian decoding

4.6 Model Deployment, APIs & Ethical Considerations

1. API Architecture & Hugging Face Ecosystem

- **REST APIs (FastAPI / Flask):** Lightweight, asynchronous web microservices serving NLP model predictions over HTTP POST endpoints using JSON payloads.

- **Hugging Face Inference APIs:** Serverless cloud inference pipeline allowing instant zero-configuration deployment of pretrained transformer checkpoints via standardized Python client APIs.

2. Production Deployment Engineering & Optimization

- **1. Model Quantization:** Converting 32-bit floating-point weights (FP32) into 8-bit integers (INT8), slashing memory consumption by 75% and speeding up CPU/GPU inference with negligible loss in accuracy.
- **2. ONNX Runtime:** Open Neural Network Exchange format for cross-platform model execution optimized across Intel, AMD, and NVIDIA hardware.
- **3. Containerization (Docker):** Packaging model code, Python dependencies, and runtime environment into immutable, reproducible Docker containers deployed to Kubernetes clusters.

3. Performance Monitoring & Ethical Considerations

- • **Performance Metrics:** Inference Latency (p95, p99 response times in ms), Throughput (requests/sec and generated tokens/sec), Data Drift & Concept Drift monitoring.
- • **Ethical AI & Privacy:** Mitigating demographic bias in training corpora, toxicity filtering guardrails, and automated Redaction of Personally Identifiable Information (PII) to comply with GDPR data protection laws.

4.7 High-Yield University Exam Question Bank

Part A: Short Answer Questions (2 to 5 Marks)

Q1. Differentiate between Constituency Parsing and Dependency Parsing. [3 Marks]

- **Answer:** Constituency parsing decomposes sentences into hierarchical phrase structure constituents (NP, VP) with non-terminal nodes. Dependency parsing establishes direct typed grammatical relations (head -> dependent, e.g., nsubj, dobj) exclusively between word tokens without non-terminal nodes.

Q2. Explain BLEU score and the role of the Brevity Penalty. [3 Marks]

- **Answer:** BLEU measures n-gram precision of machine translations against reference human texts. The Brevity Penalty BP = exp(1 - r/c) prevents the metric from assigning high scores to artificially short candidate outputs (e.g., generating only 'the' to get 100% precision).

Q3. Explain the Simplified Lesk Algorithm for Word Sense Disambiguation. [4 Marks]

- **Answer:** Given an ambiguous word in context, the Lesk algorithm compares the word's surrounding context words with the dictionary definitions (glosses) of all candidate WordNet senses. The sense with the highest word overlap count is selected.

Part B: Long Answer Questions (10 to 15 Marks Outline Guides)

- **Q4. Explain Sequence Labeling for Named Entity Recognition (NER) and POS tagging with Brill's Transformation-Based Learning. [10 Marks]**
- **Q5. Explain Transition-Based Dependency Parsing using Stack, Buffer, and Shift-Reduce Arc actions with an example parse trace. [12 Marks]**
- **Q6. Detail the complete Mel-Frequency Cepstral Coefficients (MFCC) extraction pipeline and Automatic Speech Recognition (ASR) Bayesian architecture. [12 Marks]**
- **Q7. Discuss the end-to-end NLP Model Deployment workflow covering REST APIs, Model Quantization, ONNX, Docker, and Ethical AI considerations. [10 Marks]**

4.8 Unit Summary & Key Takeaways

- Sequence labeling assigns aligned tags to tokens, powering POS tagging (HMM, Brill) and NER (BIO tagging).
- Dependency parsing maps head-dependent relationships via Transition-Based (Shift-Reduce) or Graph-Based (MST) algorithms, evaluated via UAS and LAS.
- Evaluation metrics include Token/Entity F1, BLEU (n-gram precision + Brevity Penalty for MT), and ROUGE (recall/LCS for summarization).
- Core applications encompass Sentiment Analysis, Vector Space Information Retrieval, and Word Sense Disambiguation (Lesk gloss overlap).
- Speech processing uses MFCC acoustic feature extraction (Framing, FFT, Mel filterbank, DCT) within the Bayesian ASR pipeline.
- NLP deployment leverages REST APIs (FastAPI), INT8 Quantization, ONNX acceleration, Docker containers, and GDPR/Ethical safeguards.