

Devi Ahilya University, Indore, India Institute of Engineering & Technology				IV Year B.E. (Information Technology)			
Subject Code & Name	Instructions Hours per Week			Credits			
6ITRE3 Information Retrieval & Extraction	L	T	P	L	T	P	Total
	Duration of Theory Paper: 3 Hours	3	1	2	3	1	1

Learning objective:

- Understand the biological processes involved in speech processing, including place and manner of articulation, and word boundary detection.
- Learn about the fundamentals of morphology and morphological diversity in Indian languages, and apply finite state machine-based morphology and automatic morphology learning.
- Explore parsing algorithms and theories, and understand how to implement robust and scalable parsing on noisy text, including hybrid rule-based and probabilistic parsing, scope ambiguity, and attachment ambiguity resolution.

UNIT-I

Sound : Biology of Speech Processing; Place and Manner of Articulation; Word Boundary Detection; Aramex based computations; HMM and Speech Recognition. Words and Word Forms .

UNIT-II

Morphology fundamentals; Morphological Diversity o f Indian Languages; Morphology Paradigms; Finite State Machine Based Morphology; Automatic Morphology Learning; Shallow Parsing; Named Entities; Maximum Entropy Models;

UNIT-III

Random Fields. Structures : Theories of Parsing, Parsing Algorithms; Robust and Scalable Parsing on Noisy Text as in Web documents; Hybrid of Rule Based and Probabilistic Parsing; Scope Ambiguity and Attachment Ambiguity resolution.

UNIT-IV

Meaning: Lexical Knowledge Networks, Wordnet Theory; Indian Language Wordnets and Multilingual Dictionaries; Semantic Roles; Word Sense Disambiguation; WSD and Multilingualism; Metaphors; Coreferences. Web 2.0

UNIT-V

Applications : Sentiment Analysis; Text Entailment; Robust and Scalable Machine Translation; Question Answering in Multilingual Setting; Cross Lingual Information Retrieval (CLIR).

Learning Outcomes: Information retrieval and extraction, including speech processing, morphology, parsing, semantics, and applications of these techniques in different domains. The learning outcomes are geared towards developing the ability to implement these techniques in practical scenarios and to develop innovative solutions to information retrieval and extraction challenges.

List of practical:

- 1) Experiment with speech processing algorithms and tools such as HMM and Argmax-based computations.
- 2) Explore the morphology of different Indian languages and create finite-state machine-based models for morphological analysis.
- 3) Learn about the theories and algorithms of parsing and experiment with them on noisy text as in web documents.
- 4) Experiment with word sense disambiguation techniques, such as lexical knowledge networks and wordnet theory, and learn how to apply them to different languages.
- 5) Explore the applications of Web 2.0 in natural language processing, such as sentiment analysis, text entailment, and question-answering in a multilingual setting.
- 6) Learn how to collect and preprocess speech and text data using different tools and techniques.
- 7) Learn how to extract different features from speech and text data, such as MFCCs for speech and Bag of Words for text.
- 8) Learn how to train different models, such as HMMs and maximum entropy models, using different training techniques and datasets.
- 9) Evaluate the performance of different models using different metrics, such as precision, recall, and F1 score.