**Course Outline**

* Introduction (1 Lecture)
* Estimation Techniques, and Language Modeling (1 Lecture)
* Parsing and Syntax (5 Lectures)
* The EM Algorithm in NLP (1 Lecture)
* Stochastic Tagging, and Log-Linear Models (2 Lectures)
* Probabilistic Similarity Measures and Clustering (2 Lectures)
* Machine Translation (2 Lectures)
* Discourse Processing: Segmentation, Anaphora Resolution (3 Lectures)
* Dialogue Systems (1 Lecture)
* Natural Language Generation/Summarization (1 Lecture)
* Unsupervised Methods in NLP (1 Lecture)

**Objectives**

Upon completion of 6.864, students will be able to explain and apply fundamental algorithms and techniques in the area of natural language processing (NLP). In particular, students will:

* Understand approaches to syntax and semantics in NLP.
* Understand approaches to discourse, generation, dialogue and summarization within NLP.
* Understand current methods for statistical approaches to machine translation.
* Understand machine learning techniques used in NLP, including hidden Markov models and probabilistic context-free grammars, clustering and unsupervised methods, log-linear and discriminative models, and the EM algorithm as applied within NLP.

**Textbooks**

Suggested textbooks for the course are:

Jurafsky, David, and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. Upper Saddle River, NJ: Prentice-Hall, 2000. ISBN: 0130950696.

### CS8084 - NATURAL LANGUAGE PROCESSING (Syllabus) 2017-regulation Anna University

|  |  |  |
| --- | --- | --- |
| **CS8084** | **NATURAL LANGUAGE PROCESSING** | **LPTC** |

**3003**

**OBJECTIVES:**

• To learn the fundamentals of natural language processing  
• To understand the use of CFG and PCFG in NLP  
• To understand the role of semantics of sentences and pragmatics  
• To apply the NLP techniques to IR applications

|  |  |  |
| --- | --- | --- |
| **UNIT I** | **INTRODUCTION** | **9** |

Origins and challenges of NLP – Language Modeling: Grammar-based LM, Statistical LM - Regular Expressions, Finite-State Automata – English Morphology, Transducers for lexicon and rules, Tokenization, Detecting and Correcting Spelling Errors, Minimum Edit Distance

|  |  |  |
| --- | --- | --- |
| **UNIT II** | **WORD LEVEL ANALYSIS** | **9** |

Unsmoothed N-grams, Evaluating N-grams, Smoothing, Interpolation and Backoff – Word Classes, Part-of-Speech Tagging, Rule-based, Stochastic and Transformation-based tagging, Issues in PoS tagging – Hidden Markov and Maximum Entropy models.

|  |  |  |
| --- | --- | --- |
| **UNIT III** | **SYNTACTIC ANALYSIS** | **9** |

Context-Free Grammars, Grammar rules for English, Treebanks, Normal Forms for grammar – Dependency Grammar – Syntactic Parsing, Ambiguity, Dynamic Programming parsing – Shallow parsing – Probabilistic CFG, Probabilistic CYK, Probabilistic Lexicalized CFGs - Feature structures, Unification of feature structures.

|  |  |  |
| --- | --- | --- |
| **UNIT IV** | **SEMANTICS AND PRAGMATICS** | **9** |

Requirements for representation, First-Order Logic, Description Logics – Syntax-Driven Semantic analysis, Semantic attachments – Word Senses, Relations between Senses, Thematic Roles, selectional restrictions – Word Sense Disambiguation, WSD using Supervised, Dictionary & Thesaurus, Bootstrapping methods – Word Similarity using Thesaurus and Distributional methods.

|  |  |  |
| --- | --- | --- |
| **UNIT V** | **DISCOURSE ANALYSIS AND LEXICAL RESOURCES** | **9** |

Discourse segmentation, Coherence – Reference Phenomena, Anaphora Resolution using Hobbs and Centering Algorithm – Coreference Resolution – Resources: Porter Stemmer, Lemmatizer, Penn Treebank, Brill's Tagger, WordNet, PropBank, FrameNet, Brown Corpus, British National Corpus (BNC).

**TOTAL : 45 PERIODS**

**OUTCOMES:Upon completion of the course, the students will be able to:**

• To tag a given text with basic Language features  
• To design an innovative application using NLP components  
• To implement a rule based system to tackle morphology/syntax of a language  
• To design a tag set to be used for statistical processing for real-time applications  
• To compare and contrast the use of different statistical approaches for different types of NLP applications.

**TEXT BOOKS:**

1. Daniel Jurafsky, James H. Martin―Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech, Pearson Publication, 2014.  
2. Steven Bird, Ewan Klein and Edward Loper, ―Natural Language Processing with Python‖, First Edition, O‗Reilly Media, 2009.

**REFERENCES**

1. Breck Baldwin, ―Language Processing with Java and LingPipe Cookbook, Atlantic Publisher, 2015.  
2. Richard M Reese, ―Natural Language Processing with Java‖, O‗Reilly Media, 2015.  
3. Nitin Indurkhya and Fred J. Damerau, ―Handbook of Natural Language Processing, Second Edition, Chapman and Hall/CRC Press, 2010.  
4. Tanveer Siddiqui, U.S. Tiwary, ―Natural Language Processing and Information Retrieval‖, Oxford University Press, 2008.

**Course Syllabus for  
CS 388: Natural Language Processing**

Chapter numbers refer to the text: [*SPEECH and LANGUAGE PROCESSING*](http://www.cs.colorado.edu/~martin/slp2.html)

1. **Introduction**  
   Chapter 1. NLP tasks in syntax, semantics, and pragmatics. Applications such as information extraction, question answering, and machine translation. The problem of ambiguity. The role of machine learning. Brief history of the field.
2. **N-gram Language Models**  
   Chapter 4. The role of language models. Simple N-gram models. Estimating parameters and smoothing. Evaluating language models.
3. **Part Of Speech Tagging and Sequence Labeling**  
   Chapters 5-6. Lexical syntax. Hidden Markov Models (Forward and Viterbi algorithms and EM training).
4. **Basic Neural Networks**  
   Any basic introduction to perceptron and backpropagation such as section 18.7 in [Artificial Intelligence: A Modern Approach (3rd ed)](http://aima.cs.berkeley.edu/), Chapter 4 of [*Machine Learning*](http://www.cs.cmu.edu/afs/cs.cmu.edu/user/mitchell/ftp/mlbook.html), or sections 5.0 - 5.3.3 of [Pattern Recognition and Machine Learning](http://www.springer.com/us/book/9780387310732).
5. **LSTM Recurrent Neural Networks**  
   ["Understanding LSTM Networks" blog post](http://colah.github.io/posts/2015-08-Understanding-LSTMs/), optionally the original paper [Long Short Term Memory](http://dl.acm.org/citation.cfm?id=1246450).
6. **Syntactic parsing**  
   Chapters 12-14. Grammar formalisms and treebanks. Efficient parsing for context-free grammars (CFGs). Statistical parsing and probabilistic CFGs (PCFGs). Lexicalized PCFGs. Neural shift-reduce dependency parsing (see [this paper](http://cs.stanford.edu/~danqi/papers/emnlp2014.pdf)).
7. **Semantic Analysis**  
   Chapters 18-20. Lexical semantics and word-sense disambiguation. Compositional semantics. Semantic Role Labeling and Semantic Parsing.
8. **Information Extraction (IE)**  
   Chapter 22. Named entity recognition and relation extraction. IE using sequence labeling.
9. **Machine Translation (MT)**  
   Chapter 25. Basic issues in MT. Statistical translation, word alignment, phrase-based translation, and synchronous grammars.