DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
KATHMANDU UNIVERSITY

Subject: Speech and Language Processing **Course Code: COMP 473**
Credit: 3 **F.M: 100**
Type: Elective

Course Description:

This course introduces students to the basic concepts of Natural Language Processing or Computational Linguistics namely, Morphology, Syntax, Semantics and Discourse. In addition, the advanced concepts and applications along with the state-of-the-art in the domain are also covered.

Course Objectives:

To provide the students a general overview of the basics as well as the advanced concepts of Natural Language Processing (NLP). Upon the completion of the course, students are expected to be able to apply the different concepts of NLP both theoretically and practically.

Prerequisites:

It is expected that students have taken prior courses like Statistics and Probability, Discrete Mathematics, Artificial Intelligence, Programming and Data Structures. For the understanding and implementation of the algorithms, it is essential that the students have a fairly good command of some of the high level programming languages like C, C++, Python or Java.

EVALUATION:
Internal: 50
External: 50

Contents:

Unit 1: Introduction to NLP [2 hrs]

1.1. Introduction to NLP
1.2. Origins and importance of NLP
1.3. Challenges in NLP (Difficulties, Ambiguities and Evolution)
1.4. Language and Knowledge (Syntax, Semantics, Pragmatics and Discourse)
1.5. A multi-disciplinary field (Psychology, Information Retrieval), Applications of NLP.

Unit 2: Words and Morphology [6 hrs]

2.1. Finite State Machines (FSM) and Morphology
2.2. Introduction to FSM and FST
2.3. Morphological Processes
2.4. Principles of Word Construction (Suffix, Prefix, Stem, Affixes)
2.5. Morphological Representation and FSM
2.6. Lexicon
2.7. Morphotactic and Orthographic rules
2.8. Morphological Parsing and FST
2.9. Mealy machines
2.10. FST operations

Unit 3: Part of Speech Tagging [7 hrs]

3.1. Parts of Speech (PoS) Tagging and Hidden Markov Models (HMM)
3.2. PoS Tagsets
3.3. Rule-based PoS Tagging,
3.4. Stochastic PoS Tagging
3.5. Transformation based tagging

Unit 4: Syntax [8 hrs]

4.1. Syntactic Analysis
4.2. Context Free Grammar (CFG) & Probabilistic CFG
4.3. Word's Constituency (Phrase level, Sentence level)
4.4. Parsing (Top-Down and Bottom-Up)
4.5. CYK Parser, Probabilistic Parsing

Unit 5: Lexical Semantics [7 hrs]

5.1. Lexical Semantics
5.2. Lexeme
5.3. Lexicon
5.4. Senses
5.5. Lexical relations
5.6. WordNet (Lexical Database)
5.7. Word Sense Disambiguation (WSD)
5.8. Word Similarity
5.9. Vector Semantics
5.10. Distributional Models
5.11. Word Embedding

5.12. Topic Modeling

Unit 6: Discourse [5 hrs]

6.1. Pragmatic & Discourse Analysis
6.2. Monologue and Dialogue
6.3. Reference Resolution
6.4. Coherence and Cohesion
6.5. Discourse Structure

Unit 7: Advanced NLP [7 hrs]

7.1. Deep Learning for NLP
7.2. Text Sequence Modeling and Deep Learning
7.3. Statistical Language Models
7.4. Kernel Methods
7.5. Neural Language Models
7.6. Recurrent Neural Network and its variants
7.7. Attention Mechanism
7.8. Reinforcement Learning
7.9. Unsupervised Learning

Unit 8: Applications [3 hrs]

Text Book:

Speech and Language Processing – Jurafsky and Martin, Second Edition, Pearson Education.

Reference Book:

Natural Language Processing with Python. Stephen Bird, Ewan Klein, and Edward Loper. O'Reilly Media, 2009. http://www.nltk.org/book/