Natural Language Processing and Machine Translation

Course Administration

Abhishek Koirala

M.Sc. in Informatics and
Intelligent Systems
Engineering

THAPATHALI CAMPUS
INSTITUTES OF ENGINEERING

Course Overview

1) Instructor: Abhishek Koirala
   Email: readeravskh@gmail.com

2) Join Piazza as class discussion forum
   Signup link: https://piazza.com/thapathali_campus/summer2022/msiiseelective
   (Code: msnlp)

3) Presentation and labs repo
   https://github.com/developeravsk/NLP-and-Machine-Translation

THAPATHALI CAMPUS
INSTITUTES OF ENGINEERING

Grading

Internal Evaluation

Type	Weightage
Minor tests	70%
Assignments	30%

External Evaluation

Units	Chapters	Marks *
1	1,2	12
2	3	12
3	4	12
4	5	12
5	6	12
Total		**60**
* There may be minor variation in distribution of marks		

THAPATHALI CAMPUS
INSTITUTES OF ENGINEERING

Grading

Internal Evaluation

- Test/Examination
- Assignments/Paper reviews/Case Study
- Final Project

Outline

Wk 1: Fundamentals of NLP
- **Content**: NLP basics, early NLP systems, Knowledge in NLP, Phases of NLP
- **Lab sessions**: None

Wk 2: Fundamentals of NLP
- **Content**: Evaluation of NLP systems, Programming languages, Applications of NLP
- **Lab sessions**: Python basics

Wk 3: Morphology, Computational Phonology
- **Content**: Regular expressions and Automata, Text Extraction, Tokenization, Derivational Morphology, Rules, Morphological Parsing , Stemming, Lemmatization
- **Lab sessions**: Raw text tokenization, stemming and lemmatization

Wk 4: Computational Phonology and Speech Processing
- **Content**: FST Lexicons and Rules, Articulatory Phonetics, Acoustic Phonetics, MFCC Features, Mel Filter Bank, Cepstrum, Duphone Waveform Synthesis, Triphones
- **Lab sessions**: None

Wk 5: Language Models
- **Content**: N-gram model, Good Turing discounting, Laplace Smoothing, Kneser-Ney Smoothing, Huge language models and backoff, Perplexity Relation to Entropy, Feed forward Neural Language Models, Word Embeddings and OOV words
- **Lab sessions**: None

Outline

Wk 6: Language Models
- **Content**: RNN, LSTM, GRU language Models, Cross-Entropy for comparing models
- **Lab sessions**: ANN, Word Embeddings, RNN, LSTM and GRU language models

Wk 7: Parsing
- **Content**: Parsing, Earley Algorithm, Chart Parsing, CCG Parsing, PCFG, CKY, Collins Parser
- **Lab sessions**: None

Wk 8: Sequence Labelling, POS Tagging
- **Content**: Dependency Parsing, Sequence Labelling, EM algorithm, Rule based POS tagging, Viterbi Algorithm
- **Lab sessions**: None

Wk 9: POS Tagging
- **Content**: Transformation based Tagging, Tag Indeterminacy, Tagger Combination
- **Lab sessions**: Parsing, POS tags

Wk 10: Information Retrieval
- **Content**: Entities, Relations, Vector Space Model, Weighting, Models of IT, Relation Matching, Conceptual Graphs, Cross Lingual IR, Evaluating IR systems
- **Lab sessions**: None

Outline

Wk 11: Question Answering
- **Content**: Factoid QA, Question Processing, Passage Retrieval, Answer Processing, Evaluation of Factoid Answers, Summarization, Basic Dialogue systems
- **Lab sessions**: None

Wk 12: Question Answering And Conversational Agents
- **Content**: Interpreting Dialogue Acts, Detecting Correction Acts, Evaluating dialogue Systems
- **Lab sessions**: Information Retrieval, Retrieval based Conversational Agents

Wk 13: Discourse Processing and Machine Translation
- **Content**: Cohesion, Coreference Resolution, Discourse Coherence and Structure, Rule Based MT, Corpus Based MT
- **Lab sessions**: None

Wk 14: Discourse Processing and Machine Translation
- **Content**: Statistical MT, Neural Translation model, Transformers
- **Lab sessions**: NLP application using transformers

Wk 15: Discourse Processing and Machine Translation
- **Content**: Beam Searching, Attention Mechanism, MT evaluation(NIST, METEOR, ROGUE, Word Error Rate, BLEU
- **Lab sessions**: Attention technique implementation

THAPATHALI CAMPUS
INSTITUTES OF ENGINEERING

Project

- Proposal **(Wk 11 / Wk 12)**
  - Background
  - Solution
    - Discuss why this solution
    - Key characteristics of your solution
    - Advantages and limitations of your solution

- Final Presentation **(Wk 15)**
  - Problem Introduction
  - Related work
  - Experiments
  - Results
  - Discuss potential future problems, limitations and actual usage
  - Discuss what have you done well, what have you learnt

THAPATHALI CAMPUS
INSTITUTES OF ENGINEERING

Project Deliverables

- Github link
  - Members and their contribution
    - README.md - who did what, and if possible, how much percentage
  - Codebase
    - README.md - detailing how it works
- Presentation file (PDF/PPT)
- Final report in NEURIPS format
  https://media.neurips.cc/Conferences/NeurIPS2020/Styles/neurips_2020.pdf

THAPATHALI CAMPUS
INSTITUTES OF ENGINEERING