23CS801 Natural language processing for Indian languages using deep learningmethods 3 0 1 4

Summary
The course is aimed at giving an introduction to the principles and working of natural language processing using deep learning and other neural network based methods, with a focus on Indian languages. The goal of the course is to introduce the various neural network models to students, and show how they can be employed to solve the various tasks of machine translation with a special focus on Indian languages.

There will be two classes of 90 minutes per week and a practical class of 180 minutes per week. The practical sessions will be devoted to teaching software and analysing existing software that exists for NLP in Indian languages.

Credits: 4
Lectures: 3
Tutorial: 0
Practicals: 1

Prerequisites
A course in linear algebra and probability (undergraduate level is sufficient), and some calculus (especially continuity and differentiability). Knowledge of a programming language, either Python or C++, is essential. Knowledge of basic topology is desirable, but not essential. What is needed will be taught in class.

Topics to be covered in the course
1) Introduction to neural networks
What is deep learning and what does it do?
Basics of neural networks
How do they work?

2) The translation problems
Goals of translations
Ambiguity
Linguistic and data view
Practical issues
Uses

3) Syntax and semantic analysis
Syntactic analysis
Semantics analysis
Discourse analysis
Language generation

4) Words and Morphology in machine translation
Problems caused by rich morphology
Combinatorial explosion
Application to Indian languages

5) Computational graphs
Neural networks as computational graphs
Gradient computations
Hands on frameworks

6) Neural language models
Feed forward networks
Word embeddings
Noise contrastive estimation
Recurrent neural language models
Long short term memory models

7) Neural translation models
Encoding-decoding approach
Alignment model
Deep models

7) Decoding in neural translations
Searches
Ensemble decoding
Reranking
Optimising decoding
Directing decoding

8) Machine learning tricks
Machine learning failures
Hyperparameters
Sentence level optimisation

9) Adaptation
Monolingual and multilingual word embeddings
Large vocabularies
Domains and mixture models
Subsampling
Fine-tuning

10) Beyond parallel data
Using monolingual data
Multiple language pairs
Training on related tasks

11) Alternate architectures
Attention models
Convolutional models with attention
Transformers

12) Challenges in the Indian language context
Linguistic structure

Guided alignment training
Modelling coverage
Linguistic annotations

13) Corpus acquisition
Large scale parallel document mining
Web as a parallel corpus
N-gram counts and language models from common crawl
Shared task on parallel corpus filtering

14) Analysis and visualisation
Error analysis
Probing representations
Visualisation
Tracing decisions to inputs

Reference books
Phillip Koehn, ``Neural machine translation''
Ralph Grishman, ``Computational linguistics – an introduction''
Thushan Ganegedara, ``Natural language processing with Tensorflow''

Main objectives of the course
Students will be conversant with the following aspects at the end of the course
1) Basics of deep-learning based machine translation
2) Metrics of machine translation
3) Create and evaluate deep-learning based NLP models for various aspects of machine translation in Indian languages
4) Acquire training and testing corpus from the net

Course outcomes

CO1	Students can evaluate the utility of the various deep-learning based NLP models for a task
CO2	Students can deep-learning based NLP models for specific machine translation tasks
CO3	Students can design deep learning based NLP models for machine translation in Indian languages and evaluate the processes
CO4	Students can learn to acquire corpus from the net and evaluate their utility for the machine translation task at hand.

Evaluation pattern
The course carries four credits. The evaluation pattern is given below:
4 assignments – 40 points. The assignments are designed to test the student's understanding of the materials. These are both theoretical and problem oriented so that the student can assess his own abilities in handling the different aspects of the course.
1 project – 30 points. The project is designed to produce some software that would be of use in Indian industry and also provide Amrita with useful IP in NLP in Indian languages. The project will end up producing software that will be useful to the public and can be incorporated in larger projects so that the industry can directly benefit and the students become employable in the industry.
1 mid term – 10 points

1 final exam – 20 points [theory+viva]. The exams and viva are used to test the understanding of the student. They are meant to discern the ability of the student to think on the spot and complete tasks within a specific time frame.

Importance of the course:
The course focusses on the underlying essential skills needed both for research and industry in NLP area. Students need to be trained in NLP skills – especially in deep-learning based NLP, which is the backbone of NLP with large databases. In this course, we focus on the deep learning based NLP so that the students acquire skills that are actively sought by companies across India. The material focusses on applied research that prepares students with deep-learning based NLP techniques that are not only vital for further research, but also sought in the industry.

Signature with Date:

Nagesh Subbanna,
Assistant Professor,
AmritaWNA

13/02/2023