CS831 **Statistical Machine Translation** **2-0-0 2**

Year of Introduction : 2022

Summary:

The course is aimed at giving an introduction to the principles and working of statistical machine translation, with a focus on Indian languages.

Course syllabus

1) Introduction to machine translationWhat is machine translation?
Tasks of machine translation
Brief outline of the concepts employed How
machine translation has evolved.

2) Basics in linguistics
The structure of the language
The grammar
The alignment of phrases and words

3) Basics of probability
The basics of probability theory
Different distributions and models Bayesian
learning

4) Introduction to automata theory
Regular expressions
Different grammars

5) Metrics of machine translation quality
Methods of manual evaluation Methods of
automatic evaluation empirical confidence
bounds Bootstrapping

6) overview of the approaches –statistical machine translation phrase-based machine translation neural machine translation

7) Alignment
Parallel data acquisition
Document alignment
Sentence alignment
Word alignment
Linguistic adequacy of word alignment

8) Phrase based machine translation
Phrase based machine translation – overview
Phrase extraction
Log linear model
Features used
Traditional pipelines
Decoding
Pruning and feature cost estimation
Local and non-local features

9) Morphology in statistical machine translation
Problems caused by rich morphology
Combinatorial explosion
Application to Indian languages

10) Syntax in statistical machine translation
Motivation of grammars
Hierarchical model
Proper syntax
Dependency syntax

11) Word and sentence representationsDoes machine understand?
Introducing semiotics
Continuous representations
Aspects of meaning
Evaluating sentence representations

12) Machine translation in Indian languages
Peculiarities of Indian languages,
Specific problems of Indian languages,
Use of statistical methods for Indian languages Limitations of
statistical machine translation

Reference books:
Philipp Koehn, ``Statistical Machine Translation''
Emily Bender, ``Linguistic Fundamentals for natural language processing'' Ralph
Grishman, ``Computational linguistics – an introduction''

Other references will be given during the lectures.

Main objectives of the course
Students will be conversant with the following aspects at the end of the course
1) Different types of machine translation
2) Metrics of machine translation
3) Create and evaluate statistical models for various aspects of machine translation in Indian languages

Course outcomes
CO1 Students can evaluate the utility of different types of machine translation for a task
CO2 Students can create statistical models for specific machine translation tasks
CO3 Students can design machine translation models for Indian languages
CO4 Students can evaluate the accuracy of translation using different parameters **Evaluation pattern**
The course carries two credits. The evaluation pattern is given below:
3 assignments – 30 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 be something that would be directly deployable as part of the software that will benefit people at large – something like a word guesser based on part of a word, based on the context, or a poetry analyser for a particular Indian language. Indian language NLP has been actively engaging people and translation across languages and software that enhances the capabilities of software used for Indian languages is vital. 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 – 30 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 – both in statistical NLP and deep-learning based NLP. In this course, we focus on the statistical 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 statistical machine learning techniques that are not only vital for further research, but also sought in the industry.