DSE308 or HSS322/622: Computational Linguistics (4 credits)

(also cross listed as DSE308 for Data Science and Engineering students)

Learning Objectives

This course will introduce students to computational linguistics (CL), a field combining insights from linguistics and computer science.

The course is concerned with concepts, models, and algorithms to interpret, generate, and learn natural languages, as well as their applications in the field of Natural Language Processing (NLP). Subsequently, we will discuss how the aforementioned techniques can be utilized to explore questions related to language cognition.

Students will be provided exposure to standard NLP tools and datasets as well as software packages like Natural Language Toolkit (NLTK) so that they can analyze language data for scientific as well as engineering goals. The course will have practical exercises involving the aforementioned tools and techniques. Finally, students will do a course project using the knowledge and skills they acquired in the course.

Course Contents

- Mathematical foundations: Elementary probability theory (Random variables and conditional probability), Concepts from information theory (entropy and mutual information), Noisy channel model.

- Language models: Using text corpora to estimate entropy and conditional probability using statistical models, Standard smoothing techniques.

- Part of Speech tagging: Rule-based taggers and using Hidden Markov Models (HMMs)

- Formal grammars: Chomsky hierarchy and types of grammars, Regular languages and regular expressions. Formal definition of Context Free and Context Sensitive Grammars. Constituency and dependency grammars. Models of linguistic complexity.

- Parsing: Algorithms for parsing using context free grammars, Introduction to NLTK methods to recurse through trees.

- Statistical parsing: Probabilistic context free grammars. Using statistical parsing models to model human sentence processing using measures like surprisal. Computational cognitive models of language processing and evolution.

- Foundations of intelligence in humans and machines: Thinking machines and the Turing Test, Nature of intelligence, Computational theory of mind, Connectionism

- Machine translation systems: Challenges of Machine Translation. IBM model 1 and phrase-based translation using parallel corpora.

Selected Readings:

- Speech and Language Processing: An Introduction to Natural Language Processing,Computational Linguistics and Speech Recognition (Prentice Hall Series in Articial Intelligence). Daniel Jurafsky and James H. Martin. Prentice Hall, 2nd Edition 2013.

- Foundations of Statistical Natural Language Processing. Christopher D. Manning and Hinrich Schutze. MIT Press, Cambridge, MA, USA, 1999.

- Language and Computers, 1 st Edition, Markus Dickinson, Chris Brew and Detmar Meurers, Wiley-Blackwell, 2012.

- Introduction to Psycholinguistics: Understanding Language Science, Matthew J. Traxler. John Wiley and Sons Ltd., 2012

- Turing, Alan Computing Machinery and Intelligence. Mind 49: 433-460, 1950