**CSC 417 Unit 3 Day 1 Outline**

1. Natural Language Processing
   1. Introduction
      1. Many modern computer systems perform language-related tasks
         1. Virtual assistants
         2. Search engines
         3. Text analysis
      2. Language seems to be a uniquely human construct
         1. Animals communicate, but do not use language (e.g. write books)
         2. God communicates through language
            1. Created by speaking
            2. Jesus is the Word
            3. Human use of language is a dim reflection of God
      3. Problems
         1. Human language is full of ambiguity
            1. Offers us a range of ways to express our thoughts
            2. Language can be used creatively rather than formulaically
         2. Computers do not handle ambiguity well
         3. Miscommunication is possible
            1. Sounds could be unclear
            2. Handwriting may be illegible; wrong symbols may have been written
            3. Context and intent can be difficult to determine
      4. Ambiguity in Human Language
         1. Lexical ambiguity
            1. Word with multiple meanings
         2. Structural ambiguity
            1. Vagueness in syntax/semantics
            2. “Outside of a dog, a book is a man’s best friend. Inside of a dog, it’s too dark to read.”
         3. Anaphoric ambiguity
            1. Relies on context for meaning
            2. “Suresh kicked the ball. **It** went out of the stadium.”
         4. Idiom
            1. Phrase with non-literal meaning
         5. Vagueness
            1. Non-specific language
         6. Puffery
            1. Inaccurate language in advertising that no reasonable person would accept as literal
         7. And many more examples…
   2. Brief History of NLP
      1. Foundations (1940s-1950s)
         1. Finite automata and regular expressions
            1. Emphasis on patterns and structures in language
         2. Formal grammars
         3. Probability/channel model
            1. “Noisy” input
      2. Symbolic vs. Stochastic (1950s-1970s)
         1. Symbolic
            1. Views language as sequence of symbols following syntactical rules
            2. Emphasis on parsing
         2. Stochastic
            1. Use statistical methods to address ambiguity
            2. Bayesian methods
      3. Four Paradigms (1970s-1980s)
         1. Stochastic
            1. E.g. Hidden Markov Models
         2. Symbolic approach
            1. Emphasis on patterns
         3. Logic-Based Systems
            1. Emphasis on reasoning/conversing
            2. E.g. SHRDLU
         4. Discourse Modeling
            1. Emphasis on language as collocated, structured, and coherent
            2. Analysis of language requires consideration of the entire “discourse,” not simply isolated parts
      4. Empiricism/Finite-State Models (1980s-1990s)
         1. Symbolic approaches revived
            1. Finite state models
         2. Empiricism
      5. The Field Comes Together (1990s)
         1. Emphasis on probabilistic, data-driven approaches
         2. Tied to increases in computational speed and memory
            1. More data can be stored
            2. Data can be processed faster
         3. Speech recognition becomes more practical
      6. The Rise of Machine Learning
         1. Large, annotated (labeled) datasets facilitate machine learning methods (pattern analysis)
         2. Supervised machine learning plays a large role in NLP
      7. Hard Questions
         1. In human language, is all the information in the *words* themselves?
            1. Is information found in the *arrangement* of words?
            2. Is information found in the choice of words that are avoided?
         2. Can everything that people know be formalized (written down and represented symbolically)?
            1. Epistemological Assumption