

Course Objectives

Algorithms for NLP is an introductory graduate-level course on the computational properties of natural languages and the fundamental algorithms for processing natural languages.

The course is a recommended first-semester class for both the MLT and the PhD in Language Technologies programs.

The main objectives of the course are the following:

1. Develop a thorough understanding of the principles and formal methods used in the design and analysis of language processing algorithms.
2. Provide an in-depth presentation of the major algorithms used in NLP, including Lexical, Morphological, Syntactic and Semantic analysis, with the primary focus on parsing algorithms and their analysis.

Pre-requisites and Co-requisites

- Minimal exposure to syntax and structure of NL (English)
- College level course on algorithms
- College level programming skills in some imperative and/or functional programming language.
- The self-paced Laboratory in NLP (11-712) is designed to complement this course with programming assignments on relevant topics. Students are encouraged to take the lab in the semester immediately following this course.

Course Structure

- Class time will be mostly in the form of lectures interleaved with in-class discussion.
- Grades will be based on Midterm Exam (25%), Final Exam (50%), and Homework Assignments (25%).
- 5-6 Homework Assignments will be handed out (every 2-3 weeks). Assignments will include problem solving and short essay questions and minor program development tasks. Homework Assignments will usually be due two weeks after being handed out. Hand-in procedures will be announced later on.

Academic Integrity Expectations

- Students are expected to strictly follow standard rules of academic integrity
- Unless explicitly instructed otherwise, all handed in work must be performed individually.
- You may not copy the work of others, or any materials or solutions from previous years of the course, regardless of whether materials are publically available or not.
- **Always** cite the sources of any material that is not your own.
- In case of *any* doubt - ask the instructors for direction.
- Severe actions will be taken against students that violate the above policy, possibly resulting in course failure or dismissal from the program.

Major Topics to be Covered

1. **Introduction/Overview of NLP:**

NL Applications, Levels of Language Analysis, Knowledge Representation for NLP.

2. **Introduction to Formal Language Theory:**

Languages, recognition and decision algorithms; Regular Languages, Finite State Automata and their properties; CFLs, CFGs and their properties.

3. **Introduction to Algorithm Analysis:**

Time and Space Computational Complexity; Average case and worse case complexity analysis.

4. **Search Techniques and Algorithms:**

Search Spaces, DFS, BFS, A* and Beam Search.

5. **Morphological Processing and Lexical Analysis:**

Lexical Knowledge and its organization; Finite-State Morph.

6. Part-of-Speech Tagging:

Statistical and transformation-based part-of-speech tagging.

7. Parsing Algorithms for CFLs:

Top-down and bottom-up parsing; Chart Parsers, CYK and Earley's Parsing Algorithms; LR Parsing and Generalized LR Parsing.

8. Parsing Alternative Grammar Formalisms:

Features and Feature-structures; Feature Unification; Unification-based Grammar Formalisms; Augmented GLR Parsing; PROLOG and DCGs; Tree-Adjoining Grammars; Dependency Grammars.

9. Ambiguity Resolution:

Types and Levels of Ambiguity; Principle-based Methods: Right Association, Minimal Attachment; Statistical Methods: Probabilistic CFGs, probabilistic parsing; Interactive Ambiguity Resolution.

10. Introduction to Semantic Processing:

Semantic Knowledge Representation, Deep Structure and Logical Form; Compositional Semantic Interpretation; Semantic Grammars; Case Frames and Case Frame-based Parsing.

11. Natural Language Generation:

Problems in NL Generation; Basic Generation Techniques: GenKit.

12. Hard Problems in NLP:

Speech Understanding and Translation; Discourse Processing; Anaphora and Ellipsis Resolution.

Text Books and Readings

Text Books:

1. James Allen, “Natural Language Understanding”, 2nd edition.
2. Hopcroft and Ullman, “Introduction to Automata Theory, Languages and Computation”. Second Edition, Chapters 1-7.
3. Jurafsky and Martin, “SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition”.

Text Books and Readings

Additional Books and Readings:

- Rich and Knight, “Artificial Intelligence”.
- Manning and Schuetze, “Foundations of Statistical NLP”
- Aho, Sethi and Ullman, “Compilers: Principles, Techniques and Tools”.
- Cormen, Leiserson and Rivest, “Introduction to Algorithms”.
- Gazdar and Mellish, “Natural Language Processing in Lisp”.
- Terry Winograd, “Language as a Cognitive Process”, Vol I - Syntax.
- Barr, A., and Feigenbaum, E. (eds.), 1981, “The Handbook of Artificial Intelligence”. William Kaufmann, Inc. Vol. I, Chapter 2.

Class Web Page and Directory

Class Directory:

`/afs/cs.cmu.edu/project/cmt-55/lti/Courses/711/`

Class Web Page:

`http://www.cs.cmu.edu/afs/cs.cmu.edu/project/
cmt-55/lti/Courses/711/www/`

Will Contain:

- Slides from most of the lectures (.pdf)
- Other class notes and handouts
- Homework assignments and solutions
- Online versions of readings
- Other misc. course related documents
- Web page will also contain up-to-date class schedule

Introduction to NLP

Reading for today's class:

- James Allen: Chapter 1
- Rich & Knight:
Chapter 14 sections 1-2
Chapter 15 sections 1-2

Natural Language Processing and Understanding

What is NLP ?

The process of computer analysis of input provided in a human language, and conversion of this input into a useful form of representation (for immediate or delayed action).

Forms of Natural Language:

- **text/written language:** Newspaper articles, letters, manuals, prose, etc.
- **spoken language:** Read speech (i.e radio, TV, dictations), conversational/spontaneous speech, commands, etc.

Motivation for NLP/Understanding

Theoretical:

- Communication - One of the fundamentals of Human Intelligence
- Computer modeling of human language processing allows us to:
 1. Better understand language processing in humans
 2. Better understand other human cognitive processes
- Challenging AI task - Requires high levels of knowledge about the world and the ability to use this knowledge and reason with it.

Practical:

- Opens the door to natural communication with a variety of computer applications:
 - Human Computer Interaction (HCI), i.e. Information Retrieval
 - Computer assisted Human-Human communication, i.e - Machine Translation

Natural Language Understanding

What is Natural Language Understanding?

- Uncovering the mapping between the linear sequence of words (or phonemes) and the meaning that it encodes.
- Representing this uncovered meaning in a useful (usually symbolic) representation.
- By definition - heavily dependent on the target task:
 - Words and structures mean different things in different contexts
 - The required target representation is different for different tasks

Why is NL Understanding so hard?

- The mapping between words, their linguistic structure and the meaning that they encode is extremely complex and difficult to model and decompose
- Natural Language is extremely rich in form and structure, and *very* ambiguous.
- The goal of “understanding” is itself task dependent and complex.

Why is NL Understanding Hard?

1. *Complexity/abstractness* of the target representation :
In a database retrieval system - keywords for the search.
In a translation system - a symbolic representation of the meaning of the sentence (using frames, Conceptual Dependency Graphs, etc.)
2. *Ambiguity* - One input can mean many different things:
 - Lexical (word level) ambiguity (i.e. the words “can” “mean”) .
 - Syntactic ambiguity (different ways to parse the sentence).
 - Interpreting partial information, such as pronouns.
 - Contextual information.
3. Many inputs can mean *the same* thing
4. Level of interaction among components of the input.
5. Noisy input to the system (i.e. speech).

Separating Language Tasks

- Processing **written text** - using lexical, syntactic and semantic knowledge about the language, as well as the required real world information.
- Processing **spoken language** - involves all of the above, plus the challenges of speech recognition, and unique characteristics of human speech.
- Another dimension - understanding (analysis) vs. generation (synthesis)
 - For Machine Translation - both!
 - Our focus will be on understanding (analysis) of written text, looking mostly at English as a typical input natural language.

Levels in Language Analysis

- **Morphological Analysis** - analysis of words into their linguistic components.
- **Lexical Analysis** - Determine the meaning of individual words, and identifying non-word tokens (i.e. punctuation marks); POS tagging.
- **Syntactic Analysis** - Parsing : transforming linear sequences of words (sentences) into structures that show how they relate to each other.
- **Semantic Analysis** - assigning meanings to the structures created by the syntactic analysis. Mapping words and structures to particular domain objects in way consistent with our knowledge of the world.
- **Discourse Integration** - capturing the contextual effects that individual sentences have on each other in determining their joint meaning.
- **Pragmatic Analysis** - Using more general knowledge about the world and the task domain to modify the interpretation into it's "true" meaning.

Morphological Analysis

- Analyzing words into their linguistic components
- with English - a comparatively simple task.
- The analytic to synthetic spectrum of languages:
 - Analytic* - each linguistic component (such as time, person, gender, number etc.) represented by a separate word.
Examples : Chinese, English (to some extent).
 - Synthetic* - many linguistic components integrated into one word. Examples : Finnish, Turkish, Hebrew.
- Common analysis tools:
 - Rule based transformations, Finite-state transducers.

Example : Hebrew vs. English

Hebrew : TIFGOSH ET HA-YELED BA-GAN.

(4 words)

English : You will meet the boy in the park.

(8 words)

TIFGOSH = You will meet

Verb conjugation captures the tense, person, gender and number in one single word.

Lexical Processing

- Purpose - determine meanings of individual words
- Basic method - lookup in a database of meanings : a *lexicon*
- Identifying non-word tokens, such as punctuation marks.
- Problem - **word-level ambiguity** - words may have several meanings, and the “correct” one cannot be chosen based solely on the word itself
- Example : the word “bank”, the word “mean”
- Further problem - domain specialized meanings
- Solutions: resolve on the spot (i.e POS tagging), or pass on the ambiguity

Part-of-Speech Tagging

- Most syntactic grammars describe sentence structure from the level of POS (articles, nouns, verbs, prepositions, etc.)
- Many words have multiple possible POS (“*can*”: *VB*, *VBP*, *NN*)
- For a given input sentence, how do we determine the correct sequence of POS tags?
- The *meaning* of the sentence determines the correct tagging
⇒ a “chicken and egg” problem
 - Consider multiple POS for each word during parsing - may pass along the ambiguity
 - Select *a single* POS tagging prior to parsing, using various statistical methods

Syntactic Processing

- Parsing - converting a flat input sentence into a hierarchical structure that corresponds to the units of meaning in the sentence.
- Large variety of parsing formalisms and algorithms
- Most formalisms have two main components :
 - A *grammar* - a declarative representation describing the syntactic structure of sentences in the language in a succinct way.
 - a *parser* - an algorithm that analyzes the input and outputs a structural representation of it (a parse), consistent with the grammar specification.
- Context-free grammars serve as the nucleus of many of the parsing mechanisms, although in most systems, they are complemented by some additional features that make the formalism more suitable to handle NL.

Semantic Analysis

Semantic Analysis:

- Assigning meanings to the structures created by syntactic analysis.
- Standard Methodology - Symbolic Representation
- Mapping words and structures to particular domain objects in way consistent with our knowledge of the world.
- Semantic interpretation plays an important role in selecting among competing syntactic analyses and weeding out “illogical” analyses.

Knowledge Representation for NLP

- Depends on the task : Machine Translation vs. Database query system
- Requires the choice of a Representational Framework, as well as the specific meaning vocabulary (i.e. what are the concepts used and the relationship between them)
- Must enable the application to perform the desired task
- Common Representations:
 - Logical Forms (i.e first-order predicate logic)
 - Semantic Frame representations
 - Conceptual Dependency Graphs

Next Class Topic - Introduction to FLT

Reading for next class:

- Hopcroft & Ullman, “Introduction to Automata Theory, Languages, and Computation”, Chapters 1,2.1-2.2.