

Natural Language Processing

CSCI 4152/6509 — Lecture 1

Course Introduction

Instructors: Vlado Keselj

Time and date: 14:35 – 15:55, 23-Sep-2025

Location: Studley LSC-Psychology P5260

CSCI 4152/6509

(Advanced Topics in) Natural Language Processing

Lectures: Tue, Thu 14:35–15:55

Studley LSC-Psychology P5260

Labs: Mon 08:35–09:55 (g-B02) Mona Camp.1201

Mon 14:35–15:55 (u/g-B01) Goldberg CS 1201

Instructor: Vlado Keselj

(Vlado Kešelj, pron. \approx Vlado Keshel)

e-mail: vlado@cs.dal.ca or vlado@dnlp.ca

URL: <http://web.cs.dal.ca/~vlado/csci6509>

E-mail list: nlp-course@lists.dnlp.ca

Main References

- Required Textbook: “Speech and Language Processing” by Daniel Jurafsky and James Martin, 2013.
- Recommended Textbooks
 - ▶ “Introduction to Natural Language Processing” by Jacob Eisenstein, 2019.
 - ▶ “Natural Language Processing with Python” by Steven Bird, Ewan Klein, Edward Loper, O’Reilly, 2009 (on-line version free)
 - ▶ “Learning Perl, 6th Edition” by Randal L. Schwartz, *et al.*, 2011.
- and more Related Books listed on the web site:
 - ▶ “Foundations of Statistical Natural Language Processing” by Manning and Schuetze, 1999.
 - ▶ . . . , more listed on the web site

Evaluation

- 32% Assignments
(theory and programming)
- 32% Final exam
on core material
- 10% Class Presentation
and Participation
- 26% Project Report

Academic Integrity Policy

- Please read the given handout (also available at the course web site)
- Suspected cases of plagiarism are referred to Academic Integrity Officers, and may lead to serious consequences
- Plagiarism is defined as “the presentation of the work of another author in such a way as to give one’s reader reason to think it to be one’s own”
- Fully reference sources in your assignments and reports
- Write in your own words
- You can look at other code, but do not cut-and-paste!
- Discussing assignments verbally is likely not an issue, but do not discuss it in writing or typing

Dalhousie Culture of Respect

- We believe that inclusiveness is fundamental to education and learning.
- Every person has a right to be respected and safe.
- Misogyny and disrespectful behaviour on campus, wider community, and social media is not acceptable. We stand for equality and hold ourselves to a higher standard.
- Take an active role:
 - ▶ Be ready: do not remain silent
 - ▶ Identify the behaviour, avoid labeling or name-calling
 - ▶ Appeal to principles, particularly with friends, co-workers or similar
 - ▶ Set limits
 - ▶ Find an ally and be an ally, lead by example
 - ▶ Be vigilant

Tentative Course Schedule

- ① Core Material
 - ① Introduction to NLP
 - ② Stream-based Text Processing
 - ③ Probabilistic Approach to NLP
 - ④ Syntactic Processing
 - ⑤ Unification-based NLP and Semantics
- ② Course Review
- ③ Student Presentations
- ④ Final Exam

Introduction to Natural Language Processing

- Reading: Chapter 1 of Jurafsky and Martin [JM]
- How to define NLP?
 1. Direct definition
 - ▶ What is a **natural** language?
 - ▶ What are other kinds of languages?
 2. NLP applications
 3. NLP as a research area

Some NLP Applications

- machine translation
- speech analysis and generation systems
- spell checking and grammatical correction
- conversational agents (e.g., chat bots)
- document generation (or computer support in document writing)
- text classification, summarization, mining
- information retrieval and information extraction
- question answering
- support applications, such as: stemming, POS tagging, semantic tagging, and partial parsing
- natural language programming code generators, query generators

NLP as a Research Area

- relatively old (as old as CS), but still very active
- can be seen as a part of AI
- related to several other areas, such as:
 - ▶ Programming and Formal Languages
 - ▶ Information Retrieval
 - ▶ Machine Learning
 - ▶ Text Mining
- Some important conferences and journals:
 - ▶ ACL — Association of Computational Linguistics, NAACL, EACL, HLT, AAAI, ...
 - ▶ Computational linguistics, Natural Language Engineering, ...
- Check “NLP Research Links” on the course web site
- Useful research site:

<http://aclweb.org/anthology-new/>

Short History of NLP

before computers

1947–54 pioneers and foundational insights

1954–66 decade of optimism (“look ma no hands”), two camps: symbolic and stochastic

1966 ALPAC report in US (negative report on MT research)

1980 emergence of various systems and approaches:

- stochastic paradigm, logic-based, NLU...

1990–2000 stochastic NLP, Web, unification-based NLP

2000–2012 “The rise of Machine Learning”

2012– Deep Learning approaches

NLP Methodology Overview

- 1 Knowledge-driven and symbolic approaches using crafted rules
 - ▶ older methodology, scalability issues, appropriate for more controlled language formats
 - ▶ example applications: information extraction
 - ▶ methodology: rules and direct coding, regular expressions, unification-based methods, etc.
- 2 Data-driven and stochastic approaches using machine learning
 - ▶ newer, scalable, for open-ended applications
 - ▶ example applications: classification, clustering
 - ▶ methodology: probabilistic models, Bayesian classifiers, neural networks, deep learning, fuzzy methods, etc.

Levels of NLP

- 1 **phonetics:** physical sounds
- 2 **phonology:** sound system (phonemes) of a spoken language
- 3 **morphology:** word structure
- 4 **syntax:** inter-word structure up to sentence structure
- 5 **semantics:** meaning up to the sentence level
- 6 **pragmatics:** “speaker’s meaning” — extended from the literal sentence meaning
- 7 **discourse:** units larger than an utterance (e.g., inter-sentence meaning, references)

Phonetics and Phonology

- Levels of processing related to speech
- **Phonetics:** is computer processing concerned with physical sounds of language; performed using signal processing methodology. It can be divided into speech generation and speech analysis.
- **Phonology:** is linguistic processing of the sounds of spoken language; higher level than phonetics, mainly concerned with elementary sound units of a language called *phonemes*.

Morphology

- **Morphology:** is level of processing concerned with the structure of words in a language.
- Morphological process — word transformation
- Main morphological processes
 1. Inflection
 2. Derivation
 3. Compounding
- Example of morphological processing: stemming

Syntax

- **Syntax:** is concerned with the sentence structure, i.e., the rules for arranging words within a sentence. One of the main tasks is *parsing*, which is the task of producing a parse tree given a sentence as the input.
- Grammar — set of rules for deriving syntactic structure
- Different types of parse trees: Context-free parse trees and dependency parse trees

Semantics

- **Semantics:** is interpreting literal meaning of language up to the sentence level.
- Lexical semantics: semantics of words
- Building semantic representation of larger structures
- Methodology: neural networks, FOPC (first-order logic), unification
- Example resources: WordNet, SentiNet

Pragmatics and Discourse

- **Pragmatics:** is concerned with intended, practical meaning of language.
 - ▶ Example: “Could you print this document?”
- **Discourse:** is concerned with language structure beyond sentence level; such as inter-sentence relations, references, and document structure.
 - ▶ Examples: turn taking, speech acts