Faculty of Computer Science, Dalhousie University *4-Sep-2024*

CSCI 4152/6509 — Natural Language Processing

Lecture 1: Course Introduction

Location: Carleton Tupper Building Theatre C Instructor: Vlado Keselj
Time: 16:05 – 17:25

Part I

Introduction

1 Course Introduction

In this section we will go over basic course information, which is covered in more details in the course syllabus.

1.1 Logistics and Administrivia

CSCI 4152/6509
(Advanced Topics in) Natural Language Processing

Lectures:	Mon, Wed 16:05–17:25
	Carleton Tupper Bldg Theatre C
Labs:	Fri 10:05–11:25 (u/g-B01) Mona Camp.1108
	Fri 16:05–17:25 (g-B02) Goldberg CS 134
Instructor:	Vlado Keselj
	(Vlado Kešelj, pron.≈ 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

A short URL to access the course web site is: `https://vlado.ca/nlp`

1.2 Main References

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.

- "Syntactic Theory: A Formal Introduction" by Sag and Wasow, 1999.
- "Modern Information Retrieval" by Ricardo Baeza-Yates and Bethier Ribeiro-Neto, 1999.
- "Pattern Recognition and Machine Learning" by Christopher Bishop, 2006.
- "Statistical Language Learning" by Eugene Charniak, 1993.
- "Statistical Methods for Speech Recognition" by Fredrick Jelineck, 1997.
- "Artificial Intelligence: A Modern Approach" by Stuart Russell and Peter Norvig, 2003.

1.3 Evaluation

The following evaluation scheme will be used:

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

1.4 Tentative Course Schedule

Tentative Course Schedule

1. Core Material

   (a) Introduction to NLP
   (b) Stream-based Text Processing
   (c) Probabilistic Approach to NLP
   (d) Syntactic Processing
   (e) Unification-based NLP and Semantics
2. Course Review
3. Student Presentations
4. Final Exam

2 Introduction to Natural Language Processing

Reading: Chapter 1 of Jurafsky and Martin [JM]

Giving a basic definition of area of Natural Language Processing (NLP) is not straightforward because it changes over time and the understanding of the area is not uniform for different groups of people working in the area. We will try to approach this definition by describing the NLP in three different ways:

1. By analyzing meaning of the phrase "Natural Language Processing",
2. By describing the problems that NLP is trying to solve, and
3. By looking at what most current NLP research publications.

- What is a "natural" language?
  English, French, German, Russian, Chinese, Bambara, ...
- Other kinds of languages: artificial languages
  - music system
  - formal languages:
    - programming languages
    - markup languages
    - mathematical language (oldest)

2.1 Some NLP Applications

Slide notes:

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

2.2 NLP as a Research Area

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/`

2.3 Short History of NLP

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
- discourse modeling

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

2000–2012 "The rise of Machine Learning"

2012– Deep Learning approaches

2.4 Overview of NLP Methodology

For a general understanding of the NLP area, it is important to describe the main methodological approaches to solve NLP problem. These approaches can be roughly divided into two main groups:

1. Knowledge-driven or symbolic approach, and
2. Data-driven or stochastic approach.

Slide notes:

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.

2.5 Levels of NLP

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.

Phonetics and Phonology are the levels of processing related to speech. Phonetics is a lower level of processing and includes computer processing concerned with physical sounds of a language. It is mostly 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. It is a higher level than phonetics, mainly concerned with elementary sound units of a language called phonemes. For example, English has about 45 phonemes. The phonemes are usually denoted using a special alphabet of symbols, called the IPA alphabet (International Phonetic Alphabet).

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

Semantics is the NLP level at which we interpret literal meaning of language up to the sentence level. It includes lexical semantics, which is about word meaning, meaning of phrases, clauses, up to the full sentence. It is about literal meaning, which may not always match meaning understood by a human in a context, which is addressed at the pragmatic level

Slide notes:

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

Pragmatics is the NLP level concerned with intended and practical meaning of the language. For example, if we consider the sentence "Could you print this document?" the literal meaning is equivalent to "Do you have ability to print this document?", however the actual intended meaning is most likely the request: "Print this document."

Discourse is concerned with language structure beyond the sentence scope, such as inter-sentence relations, resolving references such as pronouns, and document structure. For example, we may want to resolve who is the person referred to by a specific pronoun "she" or "he", or which company is referred to by a simple phrase "the company" in a sentence.

2.6 Why is NLP Hard?

Since the start of NLP it has been known to be deceptive in term of difficulty of the major NLP tasks, such as machine translation or question answering. It is relatively easy to build toy systems that can perform these tasks successfully on small examples, and they give appearance that scaling up to a large domain will not be difficult. However, very often scaling up these systems to a usable coverage of examples has proven not only to require much more time but actually impossible due to emerging complexities. We will try to here to explain some main sources of this hidden difficulty of NLP.

NLP is Generally Hard

- NLP problems were tackled since 1950s
  - progress has been surprisingly slow and difficult
- Some external evidence of why NLP would be hard:
  - Turing test (imitation game)
  - Evidence from neuro-science:
    "A defining difference between man and non-human primates has been found in the circuitry of brain cells involved in language, according to researchers at the Medical College of Georgia."

`https://www.sciencedaily.com/releases/2001/09/010905071926.htm`

Some Computational Reasons that NLP is Hard

1. *highly ambiguous*
   - not easy to program disambiguation
2. *vague* (the principle of minimal effort)
   - not easy to program the context and a priori knowledge
3. *universal* (domain independent)
   - not easy to program general knowledge representation

   All of these require reasoning (inference)

Natural Language Processing (NLP) is an interesting area in the sense that many tasks appear to be solvable in a relatively straightforward way, but after solving a small number of examples and trying to scale up to a more general language scope they turn up to be much more difficult, and in practice even impossible to solve. It is useful to understand what is the source of this surprising and often hidden difficulty so that we can more easily distinguish very difficult tasks from more feasible ones, and so that we can modify a given task in a way that makes it more manageable.

It is generally recognized that ambiguities in any natural language are the main source of difficulty in NLP, and they are made even worse by minor grammatical mistakes that we freqently make. We will add two more properties of natural languages that make NLP difficult and identify that the three main sources of difficulty are:

- ambiguity,
- vagueness, and
- universality

of natural languages.

Ambiguity. Natural languages are ambiguous. They are ambiguous frequently and at different levels of NLP. When we use language as humans we resolve many literal ambiguities by using our complex understanding of the context, but frequently language is ambiguous even to us. Many language jokes are based on these inherent ambiguities. This means that if we write a program to process language and it operation depends on resolving these ambiguities, it may be a difficult or even impossible to solve some input cases. In further text, we will show many examples of ambiguities.

Vagueness. Another source of difficulty is vagueness of language in its typical use.

Universality. The third source of NLP difficulty is universality of language in the sense that the same natural language is typically used to describe very different areas of human interest: It is used for small talk, for communicating common sense knowledge, scientific knowledge, history, physics, abstract mathematics, legal rules, jokes, and so on.