Natural Language Processing

Michael Noukhovitch

Fall 2020, McGill

Notes written from Jackie Cheung's lectures

Contents

1 Introduction

1.1 Overview

language is a form of communication

- *arbitrary* pairing between form and meaning
- very expressive and productive
- nearly universal
- uniquely human*

computational linguistics modelling natural language with computational models

- acoustic signals
- NL understanding (comprehension)
- NL generation (production)

goals of the field

- practical technologies (NLP)
- understanding how language works (CL)

models and techniques

- gathering data
- evaluation
- statistical methods (ML)
- rule-based systems

some example problems

- is language an instinct? (Chomsky)
- language processing to understand meaning of sentence
- can we learn mathematical properties of language

types of language

- **text** an idealization of spoken language
  - luckily English is similar between writing and speaking, and there is lots of data on it
  - older work used "clean" language but recent work ventures into messy data (e.g. Twitter)
- **speech** is much messier
  - automatic speech recognition (ASR)
  - text-to-speech generation (TTS)

1.2 Domains of Language

phonetics study of speech sounds

- articulation, transmission
- how each sound is made in the mouth

phonology rules that govern sound patterns

- how the sounds are organized
- "p" in peach and speach are the same phoneme but phonetically distinct (aspiration)

morphology word formation and meaning

- anti-dis-establish-ment-arian-ism

syntax structure of language

- "I a woman saw park in the" is **ungrammatical**
- **ambiguity** different possible meaning for the same phrase

semantics meaning of language

- "Ross wants to marry **a** Swedish woman"

pragmatics meaning of language in context

- different from literal meaning
- **deixis** interpretation that relies on extra-linguistic context
- "dessert would be delicious"

discourse structure of larger spans of language

- do large spans of text form a coherent story

1.3 Technology

combination of hand-crafted knowledge and ML on data

- rule-based systems
- machine learning
- knowledge representation