Introduction to natural language processing

Herman Kamper

Google assistant

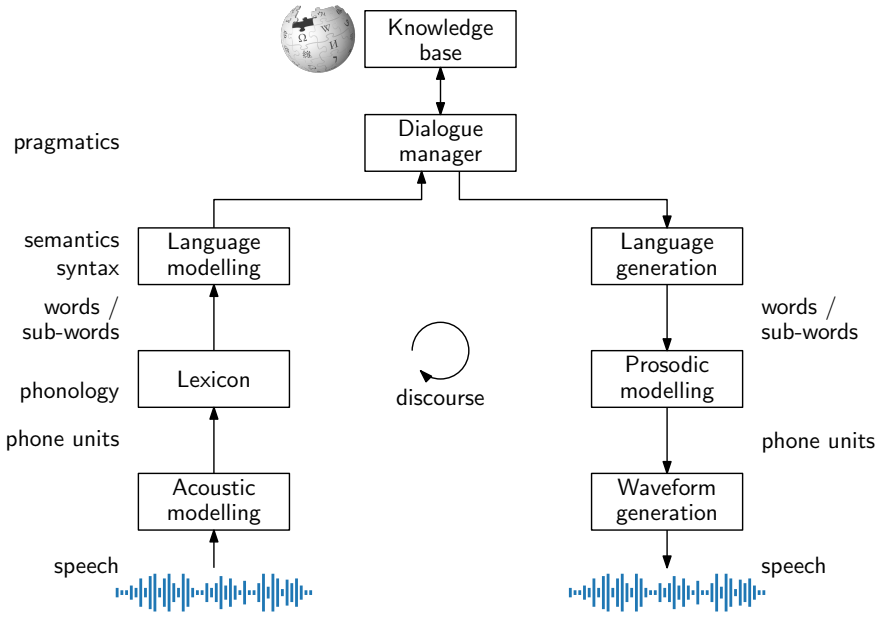
```
User:   Okay Google, what's the weather like in
        Stellenbosch today?

Google: Today's forecast for Stellenbosch is
        twenty-two degrees and sunny.

User:   What about tomorrow?

Google: Tomorrow's forecast for Stellenbosch is
        nineteen degrees and partly cloudy.
```

Think about all the components needed in the system for this brief conversation:[1]



[1] Figure adapted from https://zerospeech.com/.

What is natural language processing?

Natural language processing (NLP) aims to enable computers to process human language in order to perform useful tasks.

Computational linguistics uses computers to discover and better understand the principles of human language. In practice, the term is often synonymous with NLP (as is evident in the names of the big NLP conferences). But there is a somewhat more scientific rather than engineering (task) focus.

Spoken language processing deals specifically with continuous speech signals. In most cases this involves either mapping a speech waveform to categorical units (recognition) or converting units into a waveform (generation/synthesis). This corresponds to the lower parts of the figure above.

All these areas overlap, but often NLP refers specifically to processing symbolic inputs (text). This corresponds to the upper parts of the figure above.

More examples of NLP applications

- Spam detection.
- Text classification: Grammarly's tone detection which predicts whether text is friendly or formal.
- Machine translation: Google translate.
- Autocomplete and smart compose: Gmail.
- Virtual assistants: Siri, Cortana, Google Assistant.

NLP817 module information

Instructors

- Lecturer: Prof. Herman Kamper (kamperh@sun.ac.za)
- Teaching assistant: Leanne Nortje (18977138@sun.ac.za)

Textbook

> D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. draft, 2021.

A free draft is available online. I will refer to this as J&M3 in the notes. Older editions will be denoted as J&M1 and J&M2.

Lectures

- Lecture: Tuesday 10:00 to 12:00 (A409)
- Lecture: Wednesday 10:00 to 12:00 (A409)
- Q&A lecture: Friday 10:00 to 12:00 (A409)

If anything changes: Top of SUNLearn page.

Assessments

- Assignment 1 (23%)
- Assignment 2 (23%)
- Assignment 3 (23%)
- Assignment 4 (31%)

For each assignment you will write a report (in a paper format) and upload your report and code on SUNLearn.

Module websites

- SUNLearn: https://learn.sun.ac.za/
- After the module: https://www.kamperh.com/nlp817/

NLP817 goals and philosophy

Goals of module

- Introduce the basic tasks in NLP and discuss why they are challenging.
  - Be able to outline the processing pipeline for a task.
  - Datasets, models, algorithms and evaluation methods.
- Introduce the algorithms and models used to solve these tasks.
  - Simulate these algorithms step-by-step with pen and paper.
  - Implement some of these algorithms and models in code.
- Give you enough background to be able to read (some) current NLP research papers and do your research assignment in language or speech processing.

Module will be self-contained

Some of you might have (extensive) machine learning experience. I will aim to make the model self-contained, which means I will explain a number of models from scratch. Even though you might be experienced, hopefully these explanations in the context of a real problem will help you better understand the challenges in NLP, and maybe even help you understand the models themself better. Also, help those around you.

First time this module is offered

- If you feel the pace is too fast or slow, please let me know.
- If you spot *any* mistakes in the notes, please let me know (I am considering an additional participation mark). There will be tons of mistakes.
- If you love/hate the notes or lecturing style, please let me know.