Introduction to natural language processing

Herman Kamper

2024-01, CC BY-SA 4.0

What is natural language processing?

Course information

Course goals and philosophy

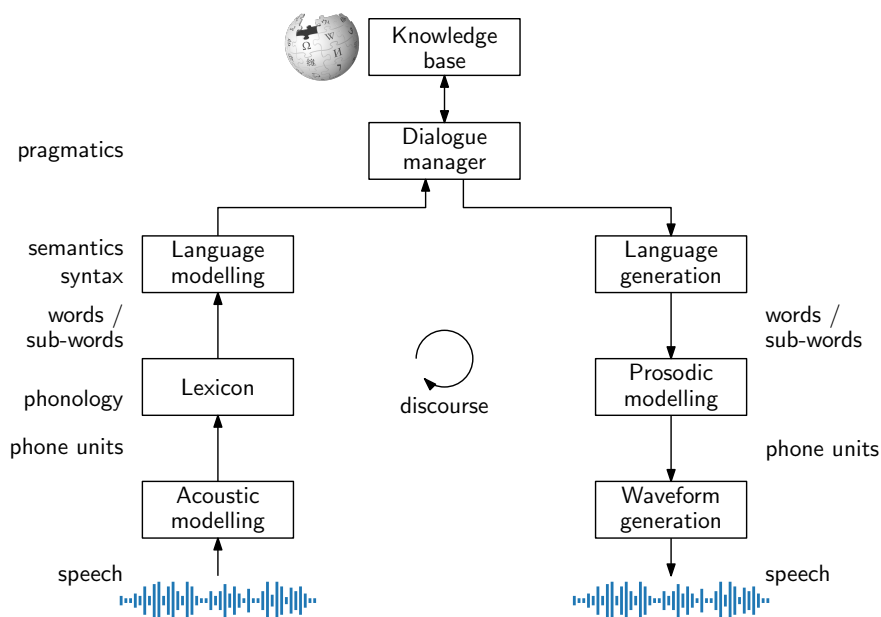# Google assistant

```
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]

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 a 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 or synthesis). This corresponds to the lower part of the figure above.

All these areas overlap. NLP often refers specifically to processing symbolic input in the form of text: the upper part of the figure.

More NLP applications

- Spam detection

- Text classification: Grammarly's tone detection predicting whether text is friendly or formal

- Machine translation: Google Translate

- Autocomplete and smart compose: Gmail

- Virtual assistants: Siri, Cortana, Google Assistant

- General dialogue systems: ChatGPT

- Coding assistants: Copilot, CodeT5

A note on ChatGPT

ChatGPT is a large language model that combines several of the modules at the top of the figure above. Dialogue management, queries to a knowledge base (e.g. a search engine or Python interpreter) and language generation are all solved using a next-word-predicting language model.

Nevertheless, it is still useful to understand the different sub-modules if we want to understand how GPT accomplishes these amazing feats and to determine its limitations.

NLP817 course information

Instructor

Prof Herman Kamper (kamperh@sun.ac.za)

Textbook

> D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. draft, 2024.

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.

Flipped classroom

The course will follow a flipped classroom structure. Every week you will watch the videos for a set of units. There will then be a Q&A lecture, followed by a short test. All the material will be posted on https://www.kamperh.com/nlp817/. The list of units will be posted weekly on SUNLearn.

Potential contact sessions

- Monday 14:00 to 16:00 (S4025)
- Wednesday 14:00 to 16:00 (S4025)
- Friday 14:00 to 16:00 (S4025)

We will normally only meet in the Wednesday session, but I will announce the schedule on SUNLearn at the start of each week.

Assessments

- Class tests (25%)
- Assignment 1 (25%)
- Assignment 2 (25%)
- Assignment 3 (25%)

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

Course websites

- SUNLearn: https://learn.sun.ac.za/course/view.php?id=89172
- Material: https://www.kamperh.com/nlp817/

NLP817 goals and philosophy

Goals of course

- Introduce the basic tasks in NLP and discuss why these 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 current NLP research papers and do your research assignment in language or speech processing.

Philosophical stuff that I hope you will take away

- Start simple, e.g. don't start with a VAE with a fancy prior

- Often we don't have a "correct answer"

- Writing is important

Course will be self-contained

Some of you might have extensive machine learning experience already. I will aim to make the course self-contained, which means I will explain a number of models from scratch. Even if you have seen these models before, 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.

Feedback

- 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. There are still tons of mistakes.
- If you love or hate the notes or lecturing style, please let me know. I want to get better.

Videos covered in this note

- What is natural language processing? (15 min)