Natural Language Processing 1

Connecting the dots

Katia Shutova

ILLC
University of Amsterdam

11 December 2020

Levels of language analysis

1. Morphology — the structure of words.
2. Syntax — the way words are used to form phrases.
3. Semantics
   - Lexical semantics — the meaning of individual words.
   - Compositional semantics — the construction of meaning of longer phrases and sentences (based on syntax).
4. Discourse and pragmatics — meaning in context.

Ambiguity

Ambiguity: *same strings can mean different things*

- Morphology: unionised (*un- ion -ise -ed* vs. *union -ise -ed*)
- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Syntactic structure: I saw a man with a telescope
- Discourse relations: Max fell. John pushed him.

Ambiguity

Ambiguity: *same strings can mean different things*

- Morphology: unionised (*un- ion -ise -ed* vs. *union -ise -ed*)
- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Syntactic structure: I saw a man with a telescope
- Discourse relations: Max fell. John pushed him.

Ambiguity

Ambiguity: *same strings can mean different things*

- Morphology: unionised (*un- ion -ise -ed* vs. *union -ise -ed*)
- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Syntactic structure: I saw a man with a telescope
- Discourse relations: Max fell. John pushed him.

Ambiguity

Ambiguity: *same strings can mean different things*

- Morphology: unionised (*un- ion -ise -ed* vs. *union -ise -ed*)
- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Syntactic structure: I saw a man with a telescope
- Discourse relations: Max fell. John pushed him.

Ambiguity

Ambiguity: *same strings can mean different things*

- Morphology: unionised (*un- ion -ise -ed* vs. *union -ise -ed*)
- Word senses: bank (finance or river?)
- Part of speech: chair (noun or verb?)
- Syntactic structure: I saw a man with a telescope
- Discourse relations: Max fell. John pushed him.

Modelling morphology

unionised: *un- ion -ise -ed* vs. *union -ise -ed*

- stemming, i.e. removing inflections
  unionise
- lemmatisation, i.e. full morphological analysis
  unionise PAST VERB

Modelling morphology

How?

1. Traditionally, finite state transducers
2. More recently, neural models: e.g. character LSTMs (advanced NLP courses)

Why is it useful?

- provides information about word structure, e.g. *shame -less*. Relevant to semantics.
- and grammatical properties, e.g. part of speech, tense, number. Informative for syntactic tasks.

Modelling syntax

How?

1. **n-gram** language models
   - compute probability of a sequence
2. **Part-of-speech** tagging
   - Sequence labelling task (assign a label to each word)
   - Hidden Markov Models (HMM)
   - more recently, neural sequence labelling (e.g. LSTMs)
3. Syntactic **parsing**
   - (Probabilistic) context-free grammars
   - Chart parsing
   - Dependency structure

Modelling syntax

What kind of information do they capture?

1. **n-gram** language models
   - word order
   - short-distance dependencies
2. **Part-of-speech** tagging
   - grammatical properties of words
   - coarse-grained word sense
3. Syntactic **parsing**
   - hierarchical structure of sentences
   - dependencies between words
   - types of phrases (e.g. NP, VP).

Modelling syntax

Why is this useful?

1. **n-gram** language models
   - language generation, e.g. fluency ranking
   - speech recognition, i.e. hypothesis ranking
   - as features in classification tasks
2. **Part-of-speech** tagging
   - precursor to parsing
   - lexical semantics
   - as features in classification tasks
3. Syntactic **parsing**
   - semantic composition
   - co-reference resolution (to identify NPs)
   - applications (e.g. summarisation).

Modelling semantics

How?

1. **Lexical** semantics
   - word sense disambiguation (supervised classification)
   - distributional semantics
   - skip-gram word embeddings
2. **Compositional** semantics
   - compositional distributional semantics
   - neural models: LSTMs and tree LSTMs

Which of the above models rely on syntax?

Modelling semantics

What kind of information do these models capture?

1. **Lexical** semantics
   - word meanings / senses
   - semantic similarity
   - semantic relations (e.g. hyponymy, synonymy)
2. **Compositional** semantics
   - meanings of phrases
   - sentence representation learning
     (general-purpose representations useful for many tasks – underlie SOTA models; discussed in ATCS course)

Modelling semantics

Why is this useful?

1. **Lexical** semantics
   - in applications (e.g. sentiment, summarisation)
   - in parsing (e.g. to resolve PP attachment ambiguity)
   - semantic similarity useful in co-reference resolution
   - input to neural models
2. **Compositional** semantics
   - paraphrasing
   - sentence similarity in applications (e.g. ordering in summarisation)
   - sentence representation learning underlies SOTA models

Modelling discourse

How?

1. **Discourse** relations
   - Classification over pairs of sentences
   - Tree-structured representations of documents
2. Learning **document representations**
   - Neural models: LSTMs, attention, HAN
   - Some later models incorporate discourse structure (ATCS)
3. **Co-reference** resolution
   - Linguistically-motivated features
   - Neural models: Lee et al (2017)

Modelling discourse

Why is this useful?

1. **Discourse** relations
   - in applications
   - e.g. summarisation: remove specific types of satellites
   - sentiment: identify contrasts in discourse
2. Learning **document representations**
   - Underlie all document classification tasks
3. **Co-reference** resolution
   - in semantics: pronouns need to be resolved
   - in applications (e.g. sentiment, summarisation)

Why does the course cover so much linguistics?

Why does the course cover so much linguistics, when all we use nowadays is machine learning anyway?

To be able to advance the state of the art you need to:

- understand the nature of the learning problem
- understand the structure of your data
- understand what patterns you might find in the data
- develop an appropriate learning algorithm for this

Understanding linguistic properties can lead to algorithmic advances in ML, e.g. the Transformer architecture. Word meaning variation in context motivated the design of self-attention.

Why does the course cover so much linguistics?

Why does the course cover so much linguistics, when all we use nowadays is machine learning anyway?

To be able to advance the state of the art you need to:

- understand the nature of the learning problem
- understand the structure of your data
- understand what patterns you might find in the data
- develop an appropriate learning algorithm for this

Understanding linguistic properties can lead to algorithmic advances in ML, e.g. the Transformer architecture. Word meaning variation in context motivated the design of self-attention.

Exam content

All lectures except guest lectures.

- Morphological processing
- n-gram language models
- Part-of-speech tagging
- Syntax, formal grammars and syntactic parsing
- Distributional semantics and word embeddings
- Compositional distributional semantics
- Neural sentence representations
- Discourse processing
- Summarisation

This is an **open-book** exam.

Types of questions

- Explain a particular linguistic phenomenon and why it is challenging for particular NLP methods / applications
- Explain the strengths and limitations of a particular method
- Apply a method to a given example
- Given examples of system errors, explain why these arise
- How can one apply a method from one NLP task to solve a particular problem in another NLP task

Exam logistics

Friday, 18 December, 9-11am

- Join a zoom meeting with your TA at 8:50am
- Conducted on ANS platform: latex or upload a picture
- PDF with questions also available on Canvas at 9am.
- If you have questions during the exam, notify your TA via chat.
- **No other communication is allowed!**
- **Strict plagiarism checks – do not share your answers!**
- Submit on ANS before 11am.

More details about the logistics will be posted on Canvas.