

Natural Language Processing 1

Summary of the course

Katia Shutova and Wilker Aziz

ILLC
University of Amsterdam

12 December 2024

Outline.

Summary of the course

What to expect at the exam

NLP and LLMs

NLP group

Levels of language analysis

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

Ambiguity

Ambiguity: *same strings can mean different things*

- ▶ 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*

- ▶ 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*

- ▶ 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*

- ▶ 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 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**
 - ▶ CKY parsing

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

- ▶ **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)

NLP and linguistics

Does linguistics play any role in today's NLP?

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 word meaning variation in context motivated the design of **self-attention**.*

Linguistics plays a big role in interpretability and explainable AI.

NLP and linguistics

Does linguistics play any role in today's NLP?

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 word meaning variation in context motivated the design of **self-attention**.*

Linguistics plays a big role in interpretability and explainable AI.

NLP and linguistics

Does linguistics play any role in today's NLP?

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 word meaning variation in context motivated the design of **self-attention**.*

Linguistics plays a big role in interpretability and explainable AI.

Outline.

Summary of the course

What to expect at the exam

NLP and LLMs

NLP group

Exam content

All lectures including guest lectures.

- ▶ Language models and conditional independence (e.g., NGram LMs, HMMs, PCFGs)
- ▶ Sequence labelling (e.g., POS tagging, NER, SRL)
- ▶ Syntax, formal grammars and syntactic parsing
- ▶ Distributional semantics and word embeddings
- ▶ Compositional distributional semantics
- ▶ Neural sequence processing and sentence representations
- ▶ Discourse processing
- ▶ Summarisation, dialogue modelling, machine translation

You are allowed to bring a **cheat sheet** (A4) and a **calculator**.

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

Outline.

Summary of the course

What to expect at the exam

NLP and LLMs

NLP group

Paradigm shift

LLMs have ignited a paradigm shift

- ▶ Design a large language model: a deep stack of Transformer layers parameterising an autoregressive language model (conditional or not) or a masked language model (or some other form of denoising auto-encoder).
- ▶ Estimate the parameters of this deep model using large amounts of (mostly) web-crawled, possibly multilingual, possibly multimodal data.
- ▶ Adapt the model to for specific tasks (e.g., translation, summarisation, question answering, etc.). Adaptation techniques include: prompt engineering, prompt tuning, fine tuning, instruction tuning, reinforcement learning, etc.

These are a big changes in the landscape of NLP research and practice.

Some of the major changes

- ▶ Limited opportunities for allowing expert knowledge to affect model design choices.
- ▶ Evaluation is difficult if the goals aren't clear.
- ▶ Role of data collection is unclear: we can adapt our models with techniques like fine tuning (adapt pretrained parameters using some task-specific supervision), but the model still has various unscoped 'functionalities' acquired in pretraining (let alone biases and other unwanted 'features').
- ▶ Models are sometimes open-weights, but too often not open-source, nor open-knowledge. This is *unlike* NLP research. The ACL Anthology is a shining example of our long-held open-access values.

LLMs for the win?

You probably have heard more than I could tell you in 20 minutes about how amazingly flexible LLMs are and how they are disrupting everything we know. If not, don't worry, you will hear more about that as you go on in AI.

Instead, I want to talk about limitations, so when you engage with this technology, and you likely will, you will engage with it thoughtfully—as good researchers do.

Limitations - Data

Large-scale, low quality bar, biased points of view,
under-representation of marginalised groups and viewpoints,
magnifies the world views of privileged groups.

Limitations - Environment

Pretraining extremely large models demands tremendous amounts of electricity, produces a lot of heat.

Limitations - Power imbalance

Only rather privileged groups can develop this technology.

- ▶ They are mostly not transparent about training and data details.
- ▶ To the extent that they guide a narrative where there's no future without this technology, they are guiding us all towards a future where they control a new type of fundamental resource.

Even using this technology requires more resources than most researchers have easy access to.

Limitations - Ethics

Exploitation of human labour in data annotation.

Dual use: these models are deployed in contexts that have obvious detrimental impact to society.

Propagation of toxicity and bias: as automatically generated context spreads through the web, they propagate and exacerbate ethical issues with the original data.

Change of incentives: the perception of an amazing technology becoming available diverts funding from other/genuine initiatives (example: startups working on MT for African languages lost investment upon the release of Meta's NLLB paper).¹

¹See what Paul Azunre from Ghana NLP had to say about this at ICLR2023.

Limitations - Scope or purpose

The technology is unscoped, presented as general purpose, and left to others to figure out what they want to do with it (often presented as ‘what *it* can do’).

NLP researchers have developed careful task-specific evaluation protocols (data collection and methodology for analysis of results) for years. These are now often bundled together and results are aggregated with little scrutiny. That’s because they are not being used to drive progress along a well-defined research programme, they are being used to show off some larger LLM.

Limitations - Risks

Lowering the bar from idea to deployment means ideas are more likely to get tested without careful risk analysis.

Examples of clearly risky applications: surveillance, medical assistants, high-stake decisions (finance, law enforcement, public investment, policy making).

These are *very risky*, often the stakes are higher for less privileged groups (consider expensive quality professional medical advice vs. cheap and unqualified output by a chat bot).

Limitations - Existential risks?

Yes, if you are thinking about how serious the limitations I outlined are and that they are possibly existential to the groups that will be most affected by deployment of LLM-powered technology.

For a thorough discussion of the dangers behind and surrounding large scale transfer learning in NLP (and beyond), read this **excellent** paper: [On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?](#)

Are there opportunities?

New and old research questions are more approachable with models that store more data, process longer contexts, operate across languages and modalities (e.g., dialogue modelling, computational semantics and pragmatics, document-level and multi-document translation and summarisation, closed-book QA, visual QA).

We can also mitigate limitations:

- ▶ Controlled generation, debiasing and detoxification
- ▶ Representing, maintaining and accessing factual knowledge
- ▶ Efficient computation and training
- ▶ Data cleaning and curation, robust evaluation protocols
- ▶ Interdisciplinary and human-in-the-centre (rather than in-the-loop) approaches

Outline.

Summary of the course

What to expect at the exam

NLP and LLMs

NLP group

Computation Arts & Humanities

Research directions

- ▶ ML and DL to improve culture and society
- ▶ Improved decision-making and problem-solving in areas like climate science and humanitarianism

People

- ▶ Tobias Blanke, t.blanke@uva.nl
- ▶ Jaap Kamps, j.kamps@uva.nl

Computational Linguistics

Research directions

- ▶ Language Emergence in Agent-based Systems
- ▶ Cognitive Modeling of Language Processing

People

- ▶ Raquel G. Alhama, `rgalhama@uva.nl`

Cognition, Language & Computation Lab

Research directions

- ▶ Posthoc Interpretability & Explainability (“Opening the blackbox of Large Language Models”)
- ▶ Cognitive & Neural Relevance of LLMS

People

- ▶ Jelle Zuidema, `zuidema@uva.nl`

Natural Language Processing and Multimodality

Research directions

- ▶ Cognition-Aware Representation Learning
- ▶ Linguistic, Reasoning, and Pragmatic Abilities of Multimodal NLP Models

People

- ▶ Sandro Pezzelle, `s.pezzelle@uva.nl`

Dialogue Modelling Group

Research directions

- ▶ Computational semantics and pragmatics
- ▶ Language use in interaction
- ▶ Dialogue models
- ▶ Conversational AI
- ▶ Language variability and change
- ▶ Visually grounded language

People

- ▶ Raquel Fernández, R.FernandezRovira@uva.nl

Natural Language Understanding Lab

Research directions

- ▶ Multilingual NLP
- ▶ Few-shot learning,
- ▶ Interpretability
- ▶ AI for social good

People

- ▶ Katia Shutova, `e.shutova@uva.nl`

Probabilistic Language Learning Lab

Research directions

- ▶ Formal approaches to uncertainty representation
- ▶ Uncertainty-aware NLP and NLG
- ▶ Explainable models
- ▶ Model interpretability for DL

People

- ▶ Wilker Aziz, `w.aziz@uva.nl`
- ▶ Bryan Eikema, `b.eikema@uva.nl`

And that's all!

Enjoy the rest of the programme

Check out other NLP courses

- ▶ Advanced topics in computational semantics
- ▶ NLP2
- ▶ DL4NLP

as well as related topics: Foundation Models, Interpretability & Explainability in AI

See you around!