L101: Machine Learning for Language Processing
Practicalities

Lecturers:
- Ryan Cotterell
- Andreas Vlachos

Materials: [https://www.cl.cam.ac.uk/teaching/1920/L101/materials.html](https://www.cl.cam.ac.uk/teaching/1920/L101/materials.html)

Any questions, email both of us:
- rdc42@cam.ac.uk
- av308@cam.ac.uk
Assessment

5% for attendance at lecture sessions, reading of assigned material, and satisfactory contribution during lectures.

95% for a small project to be agreed with the lecturers and write a project report of not more than 5000 words:

- Pick a dataset/task
- Literature survey
- Implement a system motivated by the survey
- Compare against previous work

This needs to be agreed with us by the 10/11/2019. Proposals, questions, suggestions by the 1/11/2019. Deadline to submit: 14/1/2020, 4PM (moodle)
Project ideas

- Dependency Parsing or Morphological Tagging (https://universaldependencies.org/)
- Morphological Inflection Generation (https://sigmorphon.github.io/sharedtasks/2019/task2/)
- Fact Checking against Wikipedia (http://fever.ai/)
- Natural Language Generation (http://www.macs.hw.ac.uk/InteractionLab/E2E/)
- Your choice! Please clear it with us. Need to ensure it is interesting and feasible within time/resource constraints
L101 Objectives

- Learn how to develop machine learning-based systems to perform natural language processing tasks
- Understand the algorithms powering modern NLP systems
- See some important applications in the process
L101 Prerequisites

The module has two prerequisites:

- L90: Overview of Natural Language Processing
- L95: Introduction to Natural Language Syntax and Parsing

Both are needed! A lot of topics will not be covered here, but you need to know, e.g.:

- Knowledge of some linguistics (L95)
- Distributional semantics, a.k.a. Embeddings (L90)

Also advised to look at MLRD: Machine Learning for Real Data
Why Natural Language Processing (NLP)?
What are the challenges?

Natural languages (unlike programming languages) are not designed; they evolve!
- new words appear constantly
- the parsing rules are flexible
- ambiguity is inherent

No known/agreed universal representation
- most are application-specific

World knowledge is necessary for interpretation

Many languages, dialects, styles, etc.
Why ML for NLP?

Learning from data (a.k.a. machine learning) adapts:
- to evolution: just learn from new data
- to different applications: just learn with the appropriate target representation

Compared to rule-based approaches, statistical ones:
- offer wider coverage
- can capture more complex patterns:
  - weighted features
  - continuous representations (a.k.a. neural networks)
Why ML for NLP?
Is NLP a sub-field ML?

Short answer: No, really

- Useful ML-based NLP captures linguistic intuition
- The target representations come from linguistics
Words of caution

When exploring a task, it is often useful to experiment with some simple rules to test our assumptions.

In fact, for some tasks rule-based approaches rule, especially in the industry:

- coreference resolution
- natural language generation

If we don't know how to perform a task, unlikely that an ML algorithm will find it out for us.
Which languages do we study?

Source: http://sjmielke.com/acl-language-diversity.htm
What is a word?

Writing conventions (e.g. whitespace) are not universal:

<table>
<thead>
<tr>
<th>A sentence in Chinese</th>
<th>我喜欢新西兰兰花</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interpretation 1</td>
<td>我 喜欢 新西兰 花</td>
</tr>
<tr>
<td>Interpretation 2</td>
<td>我 喜欢 新 西兰花</td>
</tr>
</tbody>
</table>

“I like New Zealand Flowers” or “I like fresh broccoli?”
(http://www.cs.waikato.ac.nz/~ihw/papers/00WT-YW-RMN-IHW-Comprsbased.pdf)

Even in English: “don't” or “do n't”?
The digital divide

• If we don’t even acknowledge that we’re working (mostly) only on English, other languages get left in the dust

• If English gets to go unnamed, then work on other languages looks “language-specific” while work on English is “NLP”

• If we only value results on English, work on other languages isn’t incentivized

https://twitter.com/emilymbender/status/1135907994678562817
Related fields

Obvious:
- machine learning
- linguistics

Kind of obvious:
- cognitive science
- statistics

Any field that involves human language and its processing:
- literature, history, etc. (a.k.a. digital humanities)
- biology
- journalism
- psychology ...
Course overview

- Introduction to machine learning for natural language processing

- Classification
  - Perceptron and friends
  - Probabilistic methods
  - Optimization fundamentals
  - Feed forward neural networks
Course overview

- Structured Prediction
  - Language models
  - Sequence tagging
  - Constituency parsing
  - Dependency parsing
  - Neural models
  - Decoding strategies
Course overview

- Sequence to Sequence models
  - Recurrent neural networks
  - Encoder-decoder architectures
  - Weighted finite-state transducers
- Applications
  - Information extraction
  - Dialogue agents
Bibliography

Jurafsky and Martin, *Speech and Language Processing 3rd edition*

and other materials referenced in the end of each lecture

Today's reading:


Jochen Leidner and Vassilis Plachouras, *Ethical by Design: Ethics Best Practices for Natural Language Processing*