Lecture 1: Introduction

LTAT.01.001 - Natural Language Processing
Kairit Sirts (kairit.sirts@ut.ee)
14.02.2020
Plan for today

- High level introduction to NLP
- Some basic NLP concepts
- Course logistics
High Level Introduction to NLP
Do you know what time it is?

Yes, it's half past ten.
Do you know what time it is?

Yes, it's half past ten.

Speech synthesis

Automatic speech recognition

- Syntax
- Semantic
- Discourse
- Pragmatics

- Dialogue planning
- Response generation
Do you know what time it is?

Yes, it's half past ten.

Speech synthesis

ASR

NLP

- Syntax
- Semantic
- Discourse
- Pragmatics

- Dialogue planning
- Response generation
Do you know what time it is?

Yes, it's half past ten.
Extremely short (and biased) history of NLP

- **Rule-based systems**
  - “One naturally wonders if the problem of translation could conceivably be treated as a problem in cryptography. When I look at an article in Russian I say: ‘This is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.’” Warren Weaver, “Translation”, 1955

- **Statistical methods**
  - “Every time I fire a linguist, the performance of our speech recognition system goes up.” Fred Jelinek, 1980s.

- **Deep learning**
  - Do we need linguistic biases or not?
The problem of many languages

- There are ca 6000 languages in the world right now
- Most NLP research is done on English
- Some other more researched languages: Chinese, Japanese, Arabic, German, French, Spanish

- Data-driven methods (including deep learning) should work regardless of the language?
- Multilingual systems?
- Interlingua?
Tower of Babel:

According to the story, a united humanity in the generations following the Great Flood, speaking a single language and migrating eastward, comes to the land of Shinar. There they agree to build a city and a tower tall enough to reach heaven. God, observing their city and tower, confounds their speech so that they can no longer understand each other, and scatters them around the world.

Wikipedia
Interlingua

“Think, by analogy, of individuals living in a series of tall closed towers, all erected over a common foundation. When they try to communicate with one another, they shout back and forth, each from his own closed tower. It is difficult to make the sound penetrate even the nearest towers, and communication proceeds very poorly indeed. But, when an individual goes down his tower, he finds himself in a great open basement, common to all the towers. Here he establishes easy and useful communication with the persons who have also descended from their towers”.

Warren Weaver
What NLP tasks do you know or can imagine?

Information retrieval
Speech recognition
Text generation
Question answering
chatbot

Aspect based Sentiment analysis
Named entity recognition
Language modeling / topic modeling
Summarization
Machine translation
Grammatical error correction
Some basic NLP concepts
Some basic NLP concepts

- Text corpora and NLP datasets
- What is a word?
- Word tokens and word types
- What is a vocabulary?
- Segmentation and tokenization
- Stemming and lemmatization
- Preprocessing
Text corpora and NLP datasets

- In the simplest form a text corpus is simply a collection of texts.
- A text corpus can be large or small
  - Is one sentence a text corpus?
  - What about 100 sentences?

- A text corpus with annotations or labels is typically called a dataset
- What kind of annotations?
  - POS tags, named entities, syntactic structures, sentiments etc
What is a word?

● Surely everyone knows what is a word, right?
  ○ A word (at least in Western languages) is something that is separated from other words with whitespaces

● However, what about
  ○ Can’t or isn’t
  ○ 75%
  ○ <div id="123">
  ○ Avrupalılaştıramadıklarımızdanmışsınızcasına (in Turkish: as if you were reportedly of ours that were unable to Europeanise)

● What about Eastern languages like Chinese or Japanese?
  ○ '上海浦东开发与建设同步' → ['上海', '浦东', '开发', '与', '建设', '同步']
Word tokens and word types

- **Word token** is a word that occurs in a running text.
- *The, cat, sat, on, the, mat* are all **word tokens** occurring in the sentence: *The cat sat on the mat*.

- Word types are the set of distinct words in a text:
  - the
  - cat
  - sat
  - on
  - mat
What is a lexicon?

- Lexicon, vocabulary, dictionary
- The set of word types an NLP system knows
- Can include frequencies
- Every word type has an index
  - the: 0, cat: 1, sat: 2, on: 3, mat: 4

- How to construct the lexicon?
Segmentation and tokenization

● Sentence segmentation or splitting
  ○ Where does a sentence end?
  ○ Simplest way: a sentence ends with a full stop, question mark or an exclamation mark

● Word tokenization
  ○ The simplest way: split on white spaces
  ○ Many English tokenizers split don’t and isn’t into two tokens: do n’t and is n’t
  ○ What to do with compounds? Door bell vs doorhandle

● In many cases segmentation and tokenization can be done reasonably well, even if it’s not always 100% correct
Subword segmentation

- Morphologically rich and productive language can have huge vocabularies
- In order to cover more words in text and still have a reasonable lexicon size (up to 100K), it might be reasonable to split words into subword units

Early attempts: linguistically correct morphological segmentation
  - misunderstanding → mis + understand + ing
  - impignorate (a rare English word) → impignorate

Nowadays approaches inspired from data compression
  - misunderstanding → misunderstanding
  - impignorate → im + pig + no + rate
Stemming and lemmatization

- Both generate the root form of an inflected word
- Wait, what is an **inflected word**? And what is a root form?
  - In linguistic morphology, inflection is a process of word formation, in which a word is modified to express different grammatical categories such as tense, case, voice, aspect, person, number, gender, mood, animacy, and definiteness. [Wikipedia](https://en.wikipedia.org/wiki/Inflection)
  - **reads** is an inflected form of a verbal root **read**
  - **books** is an inflected form of a noun root **book**
  - **men** is an inflected form of a noun root **man**
- Stemming originates from information retrieval where the goal was to reduce words with similar meaning to the same stem: **read, reads, reading, readings**
- Lemmatization produces a linguistically valid root that is itself always a valid word in the language
Stemming and lemmatization

**Stemming**

- walking → walk
- cares → car
- I am meeting you → meet
- The meeting is now → meet
- better → better
- men → men

**Lemmatization**

- walking → walk
- cares → care
- I am meeting you → meet
- The meeting is now → meeting
- better → good
- men → man
Preprocessing

The goal of **preprocessing** is to transform the text into a format that can be fed into an NLP system. Preprocessing can involve:

- Segmentation, tokenization
- Lowercasing the text
- Stemming or lemmatization
- Stopword removal

- Original sentence: this is a sentence full of content and we need to clean it up
  - Stop words removed: S S S sentence full S content S S S S clean S S

- Normalization:
  - aaaaawesome → awesome
  - acheive → achieve
  - :), :-), :)) → smile

- Noise removal: punctuation, special characters, urls, hashtags etc
Course Logistics
Course Logistics

- Lectures every week on Fridays as 10:15, Delta, 2034
- Practice sessions on Fridays at 12:15, Delta, 2034
- Course web page: [https://courses.cs.ut.ee/2020/NLP/spring](https://courses.cs.ut.ee/2020/NLP/spring)
- Reading tests in moodle

- Lecturer: Kairit Sirts - kairit.sirts@ut.ee - no regular office hours, meetings can be arranged upon request
- Teaching Assistant: Kirill Milintsevich - kirill.milintsevich@ut.ee
Course components

- Lectures – will be recorded and available in panopto.ut.ee
  - But coming to class will give an opportunity for active participation that watching videos does not
- Lab sessions
  - The TA Kirill is there for you to help and assist
- Homework assignments
  - Go to lab sessions for homework support by Kirill
- Reading tests
  - In moodle
- Project
  - In groups of up to 3 people
- Final exam
# Lecture plan

<table>
<thead>
<tr>
<th>Lecture</th>
<th>Date</th>
<th>Topic</th>
<th>Guest(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L01</td>
<td>14.02</td>
<td>Introduction to NLP</td>
<td></td>
</tr>
<tr>
<td>L02</td>
<td>21.02</td>
<td>Language modeling</td>
<td></td>
</tr>
<tr>
<td>L03</td>
<td>28.02</td>
<td>Word embeddings</td>
<td></td>
</tr>
<tr>
<td>L04</td>
<td>06.03</td>
<td>Text classification I</td>
<td></td>
</tr>
<tr>
<td>L05</td>
<td>13.03</td>
<td>Text classification II</td>
<td></td>
</tr>
<tr>
<td>L06</td>
<td>20.03</td>
<td>Sequence modeling I</td>
<td></td>
</tr>
<tr>
<td>L07</td>
<td>27.03</td>
<td>Sequence modeling II</td>
<td></td>
</tr>
<tr>
<td>L08</td>
<td>03.04</td>
<td>Syntactic parsing</td>
<td></td>
</tr>
<tr>
<td>L09</td>
<td>17.04</td>
<td>Attention mechanism</td>
<td></td>
</tr>
<tr>
<td>L10</td>
<td>24.04</td>
<td>Contextual word embeddings</td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td>08.05</td>
<td>Machine translation</td>
<td>(Mark Fišel)</td>
</tr>
<tr>
<td>G2</td>
<td>15.05</td>
<td>Ontologies</td>
<td>(Eduard Barbu)</td>
</tr>
<tr>
<td>G3</td>
<td>21.05</td>
<td>Speech recognition</td>
<td>??</td>
</tr>
<tr>
<td>G4</td>
<td>28.05</td>
<td>NLP in Industry</td>
<td>(Raul Sirel, Texta)</td>
</tr>
</tbody>
</table>

- **10.04 - Good Friday**
- **01.05 - Spring Day**

Notes:
- **G1**: Machine translation by Mark Fišel
- **G2**: Ontologies by Eduard Barbu
- **G3**: Speech recognition
- **G4**: NLP in Industry by Raul Sirel and Texta
Lab sessions and homeworks

- The goal is to get experience with some tools and tasks
- This year the practicals will be based on pytorch
- There will be 6 labs (almost) aligned with 6 homeworks

<table>
<thead>
<tr>
<th>Labs</th>
<th>Homeworks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab 1: NLP basics</td>
<td>HW 1: Text processing</td>
</tr>
<tr>
<td>Lab 2: NLP pipelines</td>
<td>HW 2: Word embeddings</td>
</tr>
<tr>
<td>Lab 3: Word embeddings</td>
<td>HW 3: Text classification</td>
</tr>
<tr>
<td>Lab 4: Text classification</td>
<td>HW 4: Sequence tagging</td>
</tr>
<tr>
<td>Lab 5: Sequence tagging</td>
<td>HW 5: Data annotation</td>
</tr>
<tr>
<td>Lab 6: Contextual word embeddings</td>
<td>HW 6: Contextual word embeddings</td>
</tr>
</tbody>
</table>
Project

● The goal of the project is to get an experience in:
  ○ working on an NLP problem
  ○ technical writing
  ○ presenting your work

● The project can be done individually or in groups of up to three people

● Runs through the second part of the course

● Project has 5 milestones
Project

- A suitable project topic conforms to the following criteria:
  - It is an NLP task that can be tackled with machine learning
  - There is suitable data available for implementing the task (or it can be easily obtained)
  - It is possible to specify criteria for evaluating the results of the project

- What can make a project topic?
  - Replication of an existing paper
  - Mixing and matching model architectures and datasets
  - New task based on an existing dataset
  - Shared tasks and competitions
    - Sigmorphon
    - Offenseval
    - Codalab competitions
## Project

<table>
<thead>
<tr>
<th>Milestone</th>
<th>Task</th>
<th>Date</th>
<th>Max points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milestone I</td>
<td>Register the group and the topic</td>
<td>April 3rd</td>
<td>3</td>
</tr>
<tr>
<td>Milestone II</td>
<td>Prepare related work section</td>
<td>April 17th</td>
<td>5</td>
</tr>
<tr>
<td>Milestone III</td>
<td>Describe the planned implementation</td>
<td>May 1st</td>
<td>5</td>
</tr>
<tr>
<td>Milestone IV</td>
<td>Report experimental results</td>
<td>May 22nd</td>
<td>5</td>
</tr>
<tr>
<td>Milestone V</td>
<td>Submit final report and code</td>
<td>May 29th</td>
<td>12</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td><strong>30</strong></td>
</tr>
</tbody>
</table>
Reading tests

● The course is partially based on two books:
  ○ Yoav Goldberg’s “Neural Network Methods for Natural Language Processing”
  ○ Jurafsky and Martin’s “Speech and Language Processing” 3rd ed. draft

● There will be reading assignments based on these books

● Reading tests as are in moodle
Deadlines

- Homework and project deadlines are **Friday mornings at 10:00am**
- Reading test deadlines are **Wednesdays at 23:59**
- All deadlines are strict
- The course has 10 extra points (110 in total)
- This will give some buffer in case something unexpected happens and you cannot submit some assignment on time
# Grading

<table>
<thead>
<tr>
<th>Component</th>
<th>Max points per item</th>
<th>Total max</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 homeworks</td>
<td>6 points</td>
<td>36 points</td>
</tr>
<tr>
<td>10 reading tests</td>
<td>2 points</td>
<td>20 points</td>
</tr>
<tr>
<td>Project</td>
<td></td>
<td>30 points</td>
</tr>
<tr>
<td>4 guest lectures</td>
<td>1 point</td>
<td>4 points</td>
</tr>
<tr>
<td>Exam</td>
<td></td>
<td>20 points</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>110 points</strong></td>
</tr>
</tbody>
</table>
How to pass the course?

- To be eligible to take the exam, obtain at least half of the points from all assessment types:
  - At least 18 from homeworks
  - At least 10 from reading tests
  - At least 15 from the project (+ the final project report must be submitted)
  - Attend at least 2 guest lectures
- Take the exam and obtain at least 10 points (out of 20)
- Sit back and enjoy!

Questions?