Lecture 1: Introduction

LTAT.01.001 - Natural Language Processing
Kairit Sirts (kairit.sirts@ut.ee)
10.02.2021
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.
Automatic speech recognition

- Syntax
- Semantic
- Discourse
- Pragmatics

Do you know what time it is?

Yes, it's half past ten.

Speech synthesis

- 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.
Added by you in the lecture

Corpus linguistics

Psychology

Cultural studies

Neuroscience

Affective computing

Computational linguistics

Speech technology

Signal processing

ARTIFICIAL INTELLIGENCE

MACHINE LEARNING

DEEP LEARNING

Linear algebra and calculus

Software engineering

Probability theory

Statistics

Optimization

Text mining

Language technology
Extremely short (and biased) history of NLP
Extremely short (and biased) history of NLP

  - “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
Extremely short (and biased) history of NLP

  - “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: 1980 – 2010**
  - “Every time I fire a linguist, the performance of our speech recognition system goes up.” Fred Jelinek, 1980s.
Extremely short (and biased) history of NLP

  - “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: 1980 – 2010**
  - Movement from rules to feature representations
  - “Every time I fire a linguist, the performance of our speech recognition system goes up.” Fred Jelinek, 1980s.

- **Deep learning: 2010 - …**
  - Movement from feature representations to distributed representations
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
“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
Interlingual translation

Semantic translation

Syntactic translation

Direct translation

Source language analysis

Pragmatic analysis

Semantic analysis

Syntactic analysis

Morphological analysis

Target language analysis
Multilingual language models and interlingua?

- Large multilingual language models seem to capture some constructs universal across languages
- We will talk more about the large language models in a later lecture.
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?
  ○ Word categories, named entities, syntactic structures, sentiments, translations, summaries, answers to questions 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”>`
  - We were going there—it was beautiful
  - 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: 2
  - cat: 1
  - sat: 1
  - on: 1
  - mat: 1
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?
  - Do we lowercase the words?
  - Do we somehow restrict 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

● Earlier attempts: linguistically correct morphological segmentation
  ○ misunderstanding $\rightarrow$ mis + understand + ing
  ○ impignorate (a rare English word) $\rightarrow$ impignorate

● Nowadays approaches inspired from data compression
  ○ misunderstanding $\rightarrow$ misunderstanding
  ○ impignorate $\rightarrow$ 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*
  ○ **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:
  - *aaaawsome → awesome*
  - *acheive → achieve*
  - *(:, :-), :) → smile*
- Noise removal: punctuation, special characters, urls, hashtags etc
How much preprocessing is necessary?

- It depends …
- Neural models need less preprocessing than statistical models (like SVM)
- Proper sentence segmentation and tokenization might not be so important for text classification as opposed to token classification, e.g. named entity recognition
- Stop-word removal is not a good idea if the word ordering is important in the model
  - For instance in the sequential models
- Usually, restricting the vocabulary is a good idea
  - Observing a word once does not give much evidence about its typical usage.
Course Logistics
Course Logistics

● Lectures every week on Wednesdays as 10:15
● Practice sessions on Tuesdays at 16:15
● Lectures and practice sessions in February and March in zoom
● Course web page: https://courses.cs.ut.ee/2020/NLP/spring
● Assessments in moodle

● Lecturer: Kairit Sirts - kairit.sirts@ut.ee
  ● Office hours: Wednesdays at 12-13, in lecture Zoom
● Teaching Assistant: Claudia Kittask - claudia.kittask@ut.ee
● Extra TA: Kirill Milintsevich - kirill.milintsevich@unicaen.fr
Course components

- Lectures – will be recorded and available in panopto.ut.ee
  - Synchronous participation is not mandatory but will provide an opportunity for active participation that watching videos does not

- Practice sessions
  - The TA Claucia is there for you to help and assist

- Homework assignments, submission in moodle
  - Go to practicum sessions for homework support by Claudia

- Reading tests and reading questions, in moodle

- Project, in moodle
  - In groups of 3 people
<table>
<thead>
<tr>
<th>Lecture</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>L01</td>
<td>10.02</td>
<td><strong>Introduction to NLP</strong></td>
</tr>
<tr>
<td>L02</td>
<td>17.02</td>
<td>Language modeling</td>
</tr>
<tr>
<td><strong>24.02:</strong> Estonian Independence Day</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L03</td>
<td>03.03</td>
<td>Word embeddings</td>
</tr>
<tr>
<td>L04</td>
<td>10.03</td>
<td>Text classification</td>
</tr>
<tr>
<td>L05</td>
<td>17.03</td>
<td>Convolutional neural networks</td>
</tr>
<tr>
<td>L06</td>
<td>24.03</td>
<td>Sequence modeling</td>
</tr>
<tr>
<td>L07</td>
<td>31.03</td>
<td>Recurrent neural networks</td>
</tr>
<tr>
<td>L08</td>
<td>07.04</td>
<td>Attention mechanism</td>
</tr>
<tr>
<td>L09</td>
<td>14.04</td>
<td>Contextual word embeddings</td>
</tr>
<tr>
<td>L10</td>
<td>21.04</td>
<td>Syntactic parsing</td>
</tr>
<tr>
<td>L11</td>
<td>28.04</td>
<td>Machine translation?</td>
</tr>
<tr>
<td>L12</td>
<td>05.05</td>
<td>Speech synthesis (Liisa Rätsep)</td>
</tr>
<tr>
<td>L13</td>
<td>12.05</td>
<td>Guest lecture about chatbots?</td>
</tr>
</tbody>
</table>
Practicum sessions and homeworks

- The goal is to get experience with some tools and tasks
- The practicums are based on pytorch
- There will be 8 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 2.1: Data annotation – part I</td>
</tr>
<tr>
<td>Lab 3: Data annotation</td>
<td>HW 2.2: Data annotation – part II</td>
</tr>
<tr>
<td>Lab 4: Word embeddings</td>
<td>HW 3: Word embeddings</td>
</tr>
<tr>
<td>Lab 5: Text classification</td>
<td>HW 4: Text classification</td>
</tr>
<tr>
<td>Lab 6: Sequence tagging</td>
<td>HW 5: Sequence tagging</td>
</tr>
<tr>
<td>Lab 7: Contextual word embeddings</td>
<td>HW 6: Contextual word embeddings</td>
</tr>
<tr>
<td>Lab 8: Data annotation evaluation</td>
<td></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 is done in groups of three people
- Runs through the second part of the course
- Project has 6 milestones
## 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>30.03</td>
<td>3</td>
</tr>
<tr>
<td>Milestone II</td>
<td>Prepare related work section</td>
<td>13.04</td>
<td>5</td>
</tr>
<tr>
<td>Milestone III</td>
<td>Describe the planned implementation</td>
<td>27.04</td>
<td>5</td>
</tr>
<tr>
<td>Milestone IV</td>
<td>Report experimental results</td>
<td>11.05</td>
<td>5</td>
</tr>
<tr>
<td>Milestone V</td>
<td>Project presentation</td>
<td>18.05</td>
<td>5</td>
</tr>
<tr>
<td>Milestone VI</td>
<td>Submit final report</td>
<td>25.05</td>
<td>12</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>35</td>
</tr>
</tbody>
</table>
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
  ○ A new task based on an existing dataset
  ○ A task relevant for a company
  ○ Shared tasks and competitions available online
Reading tests

- The course is partially based on two books:
  - Jurafsky and Martin’s “Speech and Language Processing” 3rd ed. draft
  - Yoav Goldberg’s “Neural Network Methods for Natural Language Processing”
- There are 10 reading tests based on these books, in moodle
- Additionally, you also have to create questions based on reading materials
- One questions per reading topic, 5 questions in total
- Both open questions and multiple choice questions are allowed
- Try to create the questions for the aspects that you feel are relevant to understand the topic but that have not been covered in the reading test.
Deadlines

- Homework and project deadlines are **Tuesdays at 23:59**
- Reading test and questions deadlines are **Wednesdays at 23:59**
- All deadlines are strict
- The course has 8 extra points (108 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>7 points</td>
<td>42 points</td>
</tr>
<tr>
<td>Homework extra</td>
<td>1 points</td>
<td>6 points</td>
</tr>
<tr>
<td>10 reading tests</td>
<td>2 points</td>
<td>20 points</td>
</tr>
<tr>
<td>5 reading questions</td>
<td>1 point</td>
<td>5 points</td>
</tr>
<tr>
<td>Project</td>
<td></td>
<td>35 points</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>108 points</strong></td>
</tr>
</tbody>
</table>
How to pass the course?

● To receive a grade, obtain at least half of the points from all assessment types:
  ○ At least 21 from homeworks
  ○ At least 10 from reading tests
  ○ At least 2 from creating the reading questions
  ○ At least 18 from the project

● At least 5 homeworks have to be submitted (and obtained at least 1 point)
  ● The HW2.1 Data annotation – part I is mandatory to everyone!

● All project milestones have to be submitted.

Questions?