Lecture 1: Introduction

LTAT.01.001 - Natural Language Processing
Kairit Sirts (kairit.sirts@ut.ee)
08.02.2022
Plan for today

● What is NLP?
● Some basic NLP concepts
● Course logistics
What is NLP?
Do you know what time it is?

Yes, it’s half past ten.

SPEECH

TEXT

MEANING

RESPONSE
Do you know what time it is?

Yes, it's half past ten.

Automatic speech recognition

Syntax
Semantic
Discourse
Pragmatics

Dialogue planning
Response generation

Speech synthesis
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.
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 discrete feature representations to distributed representations
  - Jan LeCun vs Chris Manning (2018): How much innate priors and biases are necessary in neural models?
Some basic NLP concepts
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
The unit of analysis in an NLP system

- document
- paragraph/passage
- sentence
- word
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%
  - 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']

1https://github.com/sebastianruder/NLP-progress/blob/master/chinese/chinese_word_segmentation.md
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
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
What are the tokens and types of this sentence?

This is a sentence, this isn’t a document.
What is a vocabulary?

- 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?
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 → 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*
  ○ **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 always itself 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 input to an NLP system. Preprocessing can involve:

- Tokenization, sentence segmentation
- 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:
  - aaaaawosome → 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
  ○ Usually stop-word removal and stemming/lemmatization do not make sense in neural models
● Proper sentence segmentation and tokenization might not be so important for text classification as opposed to token classification, e.g. named entity recognition
● Usually, restricting the vocabulary is a good idea,
  ○ Observing a word once does not give much evidence about its typical usage.
● Rule of thumb: start with as little preprocessing as possible and add only if necessary.
Regular expressions

- Regular expressions are simple useful tool for basic rule-based NLP
- Regular expressions are a reasonable choice for:
  - tokenization
  - sentence segmentation
  - simple pattern extraction
    - money, percentages, dates, abbreviations
    - specific topical words from a fixed list (good words, bad words, hate words etc)
    - etc

- If you are not familiar with regular expressions then you have the opportunity to learn about them for the first reading test.
Course Logistics
Course Logistics

- Lectures every week on Tuesdays as 12:15
- Practice sessions on Tuesdays at 16:15
- Lectures and practice sessions are in classroom, both will be recorded.
- Course web page: https://courses.cs.ut.ee/2022/NLP/spring
- Assessments in moodle
## Course components

<table>
<thead>
<tr>
<th>Assessment type</th>
<th>Points</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading tests</td>
<td>40 points</td>
<td>10 reading tests, 4 points each</td>
</tr>
<tr>
<td>Practical homeworks</td>
<td>40 points</td>
<td>10 practical homeworks, 4 points each</td>
</tr>
<tr>
<td>Project</td>
<td>25 points</td>
<td>5 milestones, 5 points each</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>105 points</strong></td>
<td></td>
</tr>
</tbody>
</table>
Reading tests

- The reading tests are based on the book
  - Jurafsky and Martin’s “Speech and Language Processing” 3rd ed. draft
- Multiple choice questions, time-limited.
- Tests are graded automatically, the review will be available after the test has closed.
- Deadlines are on Monday night at 23:59
- **No late submissions possible.**
Practical sessions and homeworks

- The goal is for you to get experience with some tools and tasks
- The neural network practicums are based on pytorch
- Deadlines are Monday night at 23:59
  - Submitting on time will guarantee quick grading within a week
- Late submission is possible (up to two weeks after the deadline), but:
  - there will be some late penalty
  - quick grading cannot be guaranteed
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

● Starting from late April
Project

- The plan is to make a competition
- You will have an opportunity to choose between couple of tasks
- There will be starter code
- Your goal is to try out an/some idea(s) on top of the starter code
- We will have leaderboards for all competition tasks
How to pass the course?

To receive a grade, obtain at least half of the points from all assessment types:

- At least 20 from reading tests
- At least 20 from the practical homeworks
- At least 11 from the project, all project milestones have to be submitted.

Questions?