Natural Language Processing

Lecture 1: Introduction

7.09.2017

Kairit Sirts
Previous and other courses

• Undergraduate level courses (recommended prerequisites)
  • Language Technology MTAT.06.045 (Keeletehnoloogia)
  • Artificial Intelligence I MTAT.06.008 (Tehisintellekt I)

• Related graduate level courses
  • Machine Translation MTAT.06.055
  • Language and Image Processing LTAT.01.005
  • Syntactic Theories and Models MTAT.06.031 (in Estonian)
  • Neural Networks LTAT.02.001
  • Data Mining MTAT.03.183
  • Special Course in Machine Learning MTAT.03.317
Spring semester courses

• Machine Learning MTAT.03.227
• Natural Language Processing in Python MTAT.03.311
• Intelligent Systems MTAT.06.048
• Computational Semantics MTAT.06.057 (in Estonian)
• Seminar on Language Technology MTAT.06.046
The goal of this lecture

Explain what this course is about, how it is organized and what you will learn during this course

After this lecture you should:
1. Have an idea what NLP is about
2. Know what you have to do and when in order to pass this course
3. Have an idea about what you will learn during this course
What is Natural Language Processing?
“Open the Pod bay doors, HAL” scene

• https://www.youtube.com/watch?v=dSIKBlibolo

Stanley Kubrick and Arthur C. Clarke, screenplay of “2001: A Space Odyssey”
Open the pod bay doors, HAL

Dave Bowman

I'm sorry Dave, I'm afraid I can't do that

HAL

• Automatic speech recognition

• Speech synthesis

• Dialogue planning
  • Syntax
  • Semantics
  • Discourse
  • Pragmatics
  • Response generation
Open the pod bay doors, HAL

Dave Bowman

• Automatic speech recognition

• Speech synthesis

I’m sorry Dave, I’m afraid I can’t do that

HAL

Natural Language Understanding

Natural Language Generation

NLP
Natural Language Processing

- Also computational linguistics, language technology
- Process human languages with computers
- Human-computer interaction
- Deals with natural language understanding and generation
Natural Language Processing Pyramid
Figure 1: Dorr et al., 2010, Natural Language Processing and Machine Translation Encyclopedia of Language and Linguistics, 2nd ed. (ELL2). Machine Translation: Interlingual Methods
Interlingua

Babel Tower

Noam Chomsky:
Deep Structure, Universal Grammar

By Pieter Brueghel the Elder,
https://commons.wikimedia.org/w/index.php?curid=22178101

By Duncan Rawlinson,
http://www.flickr.com/photos/thelastminute/97182354/in/set-72057594061270615/
Course Logistics
Course logistics

• Lectures every week on Thursday as 16:00, Ülikooli 17-218
• Practical sessions scheduled every Friday at 10:00, Ülikooli 17-220
  • No practical session this week
• Course web page: https://courses.cs.ut.ee/2017/NLP/fall

• Lecturers:
  • Kairit Sirts (that would be me): kairit.sirts@ut.ee - no regular office hours, meetings can be arranged upon request
  • Featuring Mark Fišel: fishel@ut.ee
Course components

• Lectures – will be recorded and available in panopto.ut.ee
• Lab sessions
• Homework assignments
• Seminars
• Project
Lab Sessions

• The goal is to get experience with some of the tools frequently used for NLP

• NLTK (general text processing)
• Gensim (word embeddings, topic models)
• Log-linear models/CRF
• Keras/tensorflow (for neural networks)
Homework assignments

• The goal is to get some more practical experience

• Extensions to the lectures and labs
• Four assignments
• More information in coming weeks
• **The goal is to get some experience in scientific reading and presenting**

• Read a scientific paper on NLP and present it in class

• You can choose any paper you like (assuming it is related to NLP)
  • Confirm your choice with me (in person or by email)

• You can ask for recommendation based on your interests

• Look at the papers of the latest (and also older) NLP conferences and journals in ACL Anthology ([http://aclweb.org/anthology/](http://aclweb.org/anthology/))

• The presentation will be max 30 minutes
  • possibly shorter, depending on how many people we have to accommodate into a single seminar

• More details will be available in the course web page
Project

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

• You can choose any problem you like related to NLP

• The project submission consists of three parts:
  • code/data/scripts
  • Report
  • Presentation
Grading

• 4 practical homework assignments 10% each
• 1 seminar presentation 10%
• Final project 50%
Caution

• The course is given the first time this fall, thus it has not been “debugged”
• Thus, you are in some sense the alpha testers of this course
• Apologies for any “bugs” that we discover in the course structure, materials and logistics during this semester
Course topics
Course topics

• Language modeling
• Word embeddings
• POS tagging
• Morphology
• Syntactic parsing
• Lexical semantics
• Information extraction

• Text classification
• Sentiment analysis
• Information retrieval
• Machine translation
• Natural language generation
• Text summarization
• Dialog systems
The general plan

• A matrix of tasks and methods

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Classical</th>
<th>Feature-based</th>
<th>Neural networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language modeling</td>
<td>Ngram model</td>
<td>Maximum entropy model</td>
<td>Recurrent neural networks</td>
</tr>
<tr>
<td>Parsing</td>
<td>Probabilistic CFG</td>
<td>Log-linear model</td>
<td>Recurrent/recursive neural networks</td>
</tr>
<tr>
<td>Information extraction</td>
<td>Regular expression patterns</td>
<td>Conditional random fields</td>
<td>Some new fancy neural network based method</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Language modeling

\[ P(\text{The cat sat on the mat}) = ? \]

\[ P(\text{The mat sat on the cat}) = ? \]

\[ P(\text{The cat mat the on sat}) = ? \]
Word embeddings
Part-Of-Speech tagging

PRP  VBZ  NNS  IN  DT  NN
She  sells  seashells  on  the  seashore
Morphology

Inflectional morphology

Derivational morphology
Syntactic parsing

• Constituency parsing

ROOT
  
  S
    
    NP
      
      DT NN VBD
        
        The cat sat
          
          IN
            
            PP
              
              NP
                
                on
                  
                  DT NN
                    
                    the mat

• Dependency parsing

The cat sat on the mat
root
  
  nmod
det nsubj
case
det
Lexical semantics – the meaning of words

• WordNet – a large database describing the relations between words (synonyms, antonyms, hyponyms, meronyms etc)

• Word sense disambiguation
  • *serve*: help with food or drink; hold an office; put ball into play
  • *dish*: plate; a course of a meal; communications device

• Semantic role labeling – who did what to whom?
Information extraction

• Named entity recognition

At the W party Thursday night at Chateau Marmont, Cate Blanchet barely made it up in the elevator.

• Relation extraction

Bill Gates was born in Seattle, Washington.  

Antibiotics are the first line treatment for indications of typhus.
Text classification

• Document classification
  • Spam detection (binary classification: spam/not spam)
  • Topic classification (politics/sports/finance/travel/etc)
  • Genre classification (fiction/news/etc)
  • Language identification

• Author classification:
  • Authorship attribution
  • Native language identification
  • Diagnosis of a disease (psychiatric or cognitive impairments)
  • Identifying gender or educational background
Sentiment analysis 😊 or 😞 ?

• Discover people’s opinions, emotions and feelings about product, services, events, topics etc

• Essentially document classification

• Detect the sentiment of a document

• Detect the sentiment of an aspect/feature
Information retrieval

• Information Retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers).

• There are many other use cases beyond web search:
  • E-mail search
  • Searching your laptop
  • Corporate knowledge bases
  • Legal information retrieval
Machine Translation

- Automatically translate text in one language to another language
- One of the end tasks of NLP
- Most large technology companies offer web-based MT services:
  - Google, Microsoft, Baidu, Yandex

<table>
<thead>
<tr>
<th>System</th>
<th>Source</th>
<th>the new protocol now put to us by the Commission represents very rigorous cuts compared with the old one</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>Ukrainian</td>
<td>новый доповідь, який зараз надається Комісією, представляє дуже суворі скорочення в порівнянні з попереднім</td>
</tr>
<tr>
<td>Yandex</td>
<td>Ukrainian</td>
<td>новий протокол тепер до нас комісія представляє собою дуже строгий скорочень у порівнянні зі старим</td>
</tr>
<tr>
<td>Google</td>
<td>Estonian</td>
<td>komisjoni poolt praegu esitatud uus protokoll kujutab endast väga ranget kärpeid võrreldes eelmisega</td>
</tr>
<tr>
<td>Yandex</td>
<td>Estonian</td>
<td>uue protokolli pange nüüd meile, mida Komisjon esindab väga ranged kärped võrreldes vana</td>
</tr>
</tbody>
</table>
Natural language generation

NLP = NLU + NLG
Text summarization

• Reduce a text with a computer program in order to create a summary that retains the most important points of the original text.
• To simplify the process of finding relevant information
• An application of natural language generation

• Domains:
  • Books, films, news articles
  • Text simplification
  • Summaries of scientific papers
Dialog systems
Recap

1. What is natural language processing?
2. What do you have to do in order to pass the course?
3. Name 5 topics that will be covered in the course.