92586 Computational Linguistics

Lesson 1. Introduction

Alberto Barrón-Cedeño

Alma Mater Studiorum-Università di Bologna
a.barron@unibo.it   @_albarron_

25/02/2021

Table of Contents

Materials

Introduction

Requirements

Materials

Core Bibliography

1. Lane et al. (2019)'s 📕 **Natural Language Processing in Action**[1]
2. Numerous **Wikipedia articles** on relevant topics
3. Lecture notes ▉ (*under development*)

---

[1] https://www.manning.com/books/natural-language-processing-in-action

Complementary Bibliography

1. Hovy (2021)'s 📘 **Text Analysis in Python for Social Scientists**[*][2]
2. Bender (2013)'s 📘 **Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax**[3]
3. Goldberg (2017)'s 📘 **Neural Network Methods for NLP**[4]
4. Kenneth W. Church's 📄 **Unix for poets**[5]

---
[2] https://doi.org/10.1017/9781108873352
[3] https://doi.org/10.2200/S00493ED1V01Y201303HLT020
[4] https://doi.org/10.2200/S00762ED1V01Y201703HLT037
[5] https://web.stanford.edu/class/cs124/kwc-unix-for-poets.pdf

Lesson coordinates

Slides, code, and more are all available here:

🦊

`albarron.github.io/teaching/computational-linguistics`

Tools

Essential
Python 3 development framework on any modern OS

1. Command line **or**
2. Integrated development Environment; e.g., Pycharm[6], Eclipse[7] **or**
3. Jupyter notebook; e.g., Google's colab[8], Jupyter itself[9]

Desirable

1. Git Version control system; e.g., ⭘ Github[10] **or** 🦊 Gitlab[11]
2. LATEX system for document preparation

---
[6] https://www.jetbrains.com/pycharm/
[7] https://www.eclipse.org/
[8] https://colab.research.google.com/
[9] https://jupyter.org/
[10] https://github.com
[11] https://gitlab.com

Introduction

Introduction
Natural Language as a Measure of Intelligence

📄 Turing (1950). "Computing machinery and intelligence". Mind. 59(236)

```
upload.wikimedia.org/wikipedia/commons/e/e4/Turing_Test_version_
3.png
```
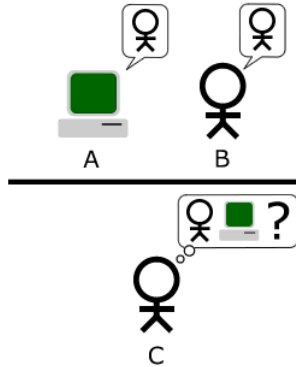

Introduction
CL vs NLP

Computational linguistics[12]
- ▶ **Interdisciplinary** field concerned with the **statistical or rule-based modeling of natural language** from a computational perspective
- ▶ Study of computational approaches to **linguistic questions**

Natural Language Processing[13]
- ▶ Subfield of **linguistics**, **computer science**, **information engineering**, and **artificial intelligence** concerned with the interactions between computers and human (natural) languages
- ▶ How to program computers to process and **analyze large amounts of natural language data**

[12]https://en.wikipedia.org/wiki/Computational_linguistics
[13]https://en.wikipedia.org/wiki/Natural_language_processing

---

Introduction
CL vs NLP

Natural Language Processing (Lane et al., 2019, p. 4)
- ▶ Area of research in computer science and artificial intelligence concerned with **processing natural languages**
- ▶ This processing generally involves **translating natural language into data** (numbers) that a computer can use to learn about the world

---

Introduction
Who can have fun with NLP

Back to the $\mathbb{W}$ikipedia article[14]

Often [. . .] performed by **computing scientists** specialized in the application of computers to the processing of a natural language

Also [. . .] interdisciplinary teams [. . .] **linguists**, **experts in the target language**, and computing scientists

[14]https://en.wikipedia.org/wiki/Computational_linguistics

Introduction

Rule-based vs Statistical NLP

Introduction
Rule-based NLP

Models are based on a number of hand-crafted rules or grammars

Diagram borrowed from L. Moroney's Introduction to TensorFlow for Artificial Intelligence, Machine Learning, and Deep Learning

---

Introduction
Rule-based NLP

Models are based on a number of hand-crafted rules or grammars

```
greeting_inputs = ("hey", "morning", "evening", "hi",
                "whatsup", "hello")
greeting_responses = ["hey", "hey hows you?", "*nods*",
                "hello, how you doing", "hello",
                "Welcome, I am good and you"]

def generate_greeting_response(input):
    for token in input.split():
        if token.lower() in greeting_inputs:
            return random.choice(greeting_responses)
```

Derived from `https:`
`//stackabuse.com/python-for-nlp-creating-a-rule-based-chatbot/`

Introduction
Statistical NLP

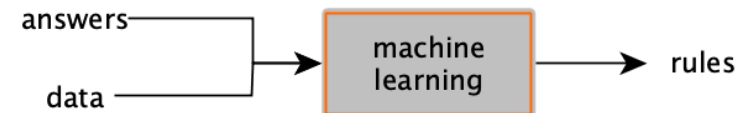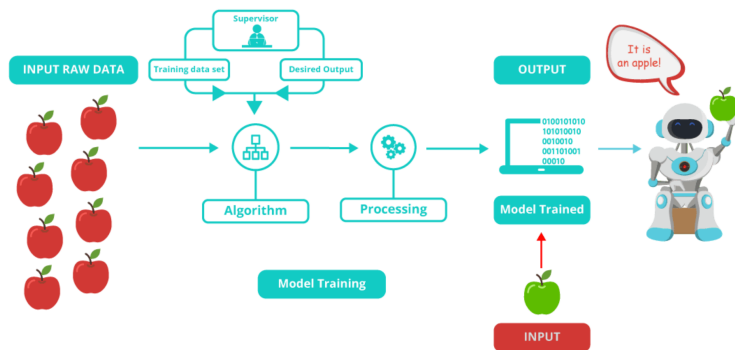Models are tuned on annotated data

Diagram borrowed from L. Moroney's Introduction to TensorFlow for Artificial Intelligence, Machine Learning, and Deep Learning
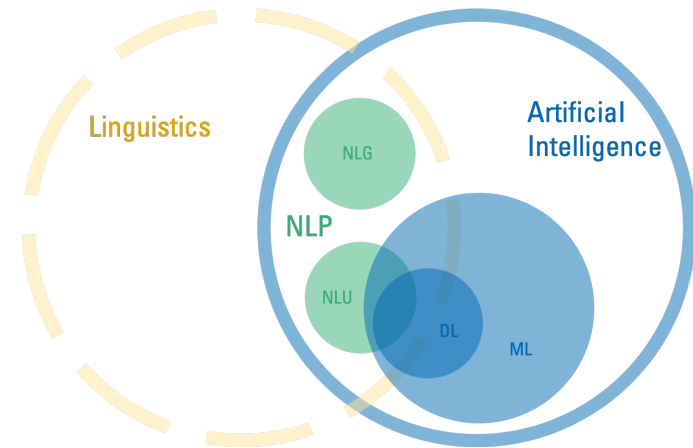
Introduction
Statistical NLP

Models are tuned on annotated data

Borrowed from
https://www.edureka.co/blog/machine-learning-tutorial

Introduction
The NLP neighborhood

Borrowed from https:
//www.retresco.de/en/how-to-ai-natural-language-processing/

Introduction
Non-exhaustive list of NLP applications with examples

🔍	**Search**	web search engines · text autocompletion
✏	**Editing**	grammar issues identification
💬	**Dialog**	chatbot creation
✉	**Email**	spam filtering · message classification
📄	**Text mining**	(multi-)document summarisation
📰	**News**	event detection · fact checking
👤	**Attribution**	plagiarism detection · authorship attribution
👍	**Sentiment analysis**	product review ranking · opinion mining
✎	**Creative writing**	text generation with a narrative and style
🈯	**Translation**	

Partially derived from (Lane et al., 2019, p. 8)

Requirements & Evaluation

Requirements

Mandatory
- ▶ Basic linguistics
- ▶ Basic algebra
- ▶ **Python**

Desirable
- ▶ Intermediate programming (e.g., object-oriented, testing)
- ▶ High-performance computing (e.g., slurm)

Evaluation: **One final project**

You will address a relevant problem...
- ▶ within the range of your own (research) interests
- ▶ participating (formally) in a shared task
- ▶ proposed by me, if you prefer

Evaluation: **One final project**

Approximate pipeline

1. You propose a topic/problem. We assess if it is reasonable, doable...
2. You compile data, study the problem, design experiments, code...

IF you plan for a publication[15]
- ▶ We meet regularly to see the advances and shape the experiments, submissions, and/or paper towards the submission deadline

ELSE
- ▶ We could meet sporadically, if you need it

3. You submit a written report ($\sim$ 7 pages) **1 week before the appello**
4. We meet on the appello date to discuss about your project, in the context of the lecture

---
[15]Talk to me well in advance, as it would require my heavy involvement and a high quality will be necessary

Evaluation: **Final mark**

Poster presentation
If the conditions (covid) are favourable, I would like you to have a poster presentation to show your work to the rest of students and to the whole department

The ingredients of the mark
Combination of the quality of the experiments, report, and oral discussion

Targetting 30L?
If I let you submit a paper, it is very likely. In summary...

$$p(30L \mid \text{paper submited} == True) \approx 0.90 \quad (1)$$
$$p(30L \mid \text{paper submited} == False) \approx 0.10 \quad (2)$$

Evaluation: Final project examples

- ♥ AriEmozione: Identifying Emotions in Opera Verses

- 🐦 UniBO@AMI: A Multi-Class Approach to Misogyny and Aggressiveness Identification on Twitter Posts Using AlBERTo

- ☰ Identifying Characters' Lines in Original and Translated Plays. The case of Golden and Horan's Class

- 🐦 Classifying An Imbalanced Dataset with CNN, RNN, and LSTM

References

Bender, E. M.
   2013. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. Morgan & Claypool Publishers.

Goldberg, Y.
   2017. *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers.

Hovy, D.
   2021. *Text Analysis in Python for Social Scientists: Discovery and Exploration*, Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

Lane, H., C. Howard, and H. Hapkem
   2019. *Natural Language Processing in Action*. Shelter Island, NY: Manning Publication Co.