92586
COMPUTATIONAL LINGUISTICS

Alberto Barrón-Cedeño
Alma Mater Studiorum–Università di Bologna

February 10, 2020

Contents

0.1 Status 5
0.2 Requirements 5
0.3 Materials 5

1 Introduction 7
1.1 Computational Linguistics 7

2 Online Resources 11
2.1 Free Text Collections 11
2.2 Code 11

Preliminaries

0.1 Status

These notes are been produced for the 2020 Computational Linguistics course, held at the Department of Interpreting and Translation, Alma Mater Studiorum–Università di Bologna. The notes and the course cover Computational Linguistics/Natural Language Processing dealing with text (and not speech).

As of February 2020 it is in an early stage and it will be developed as the course advances.

0.2 Requirements

Understanding and acting in the field of computational linguistics require some preliminary knowledge and skills (part of which are intended to be acquired during the course):

Required

- Basic linguistics
- Basic algebra
- Basic knowledge of the Python programming **language**[1]

Desirable

- Intermediate programming (e.g., object-oriented programming, testing)
- Version control (git)
- High-performance computing (e.g., slurm)

0.3 Materials

These notes are being developed by considering different materials. Among them:

1. The book Natural Language Processing in Action Lane et al. (2019).

[1] Right: this is just a (formal) language. Its vocabulary is tiny when compared to any natural language and its grammar is extremely simple.

2. Numerous Wikipedia articles on relevant topics.[2]

Some other materials will be considered, including

1. The book Neural network methods for natural language processing.
2. The book Linguistic fundamentals for natural language processing : 100 essentials from morphology and syntax.

[2] Over the years, numerous scholars have challenged the value of the Wikipedia as an academic resource. I argue in favour, as fas as it is used as a departing point to deepen into the consulted concepts.

Chapter 1

Introduction

Add Turing machine

1.1 Computational Linguistics

Computational linguistics (aka natural language processing; NLP) is multidisciplinary by nature. As the Wikipedia article about it defines it

Definition 1. *"Natural Language processing is a subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data."*[1]

Notice that the definition by Lane et al. (2019, p. 4) is fairly different:

Definition 2. *Natural language processing is an area of research in computer science and artificial intelligence (AI) concerned with processing natural languages such as English or Mandarin. This processing generally involves translating natural language into data (numbers) that a computer can use to learn about the world. And this understanding of the world is sometimes used to generate natural language text that reflects that understanding.*

In the old days, NLP was rule-based. Models were based on a number of hand-crafted rules or grammars. By the 1990s, NLP became "statistical" by the creation of techniques that learned the rules by statistical inference from corpora.

NLP ranges from the simple counting of tokens in a text to dig into the use of language to sophisticated models aiming at understanding and producing human language. No long ago, search engines were only able to process a number of keywords and would roughly combine them to perform a better search. Nowadays, natural-language queries can be processed accurately and information needs are fulfilled better. Table 1.1

As stressed Hobson Lane stresses in (Lane et al., 2019, p. xvi), often multiple names are used when referring to the same concept or idea in our field. A Markov

[1] https://en.wikipedia.org/wiki/Natural_language_processing

Table 1.1: Non-exhaustive list of NLP applications (partially derived from (Lane et al., 2019, p. 8)).

Search	**Web**. Given a query, retrieve the most relevant documents from the Web.
	Autocomplete. Searching for the most likely next item given a sequence of text.
Editing	**Grammar**. Identifying potential grammar issues in a text.
Dialog	**Chatbot**. A system that can interact (assist) a user in a conversation.
Email	**Spam filter**. Identifying commercial, phishing and other undesired email messages.
	Classification. Organising email messages according to their nature (e.g., trips, finance, entertainment).
Text mining	**Summarization**. Automatic creation of summaries given one or multiple documents.
News	**Event detection**. Grouping of the coverage of a specific event by different media.
	Fact checking. *Machine-assisted* verification of the factuality of a claim given certain evidence.
Attribution	**Plagiarism detection** Determining whether a text has been borrowed from another document (without giving proper credit).
	Literary forensics Determining whether a document has been actually written by its claimed author.
Sentiment analysis	**Product review triage**. Ranking reviews on a product/service according to their quality.
	Customer care. Understanding the level of satisfaction/stress of a client to prioritise response.
Creative writing	Generation of movie scripts, narrative, poetry, lyrics on specific topics and with a given style.

chain, which defines the likelihood of a sequence of elements (e.g., words), is a table with probabilities. Such probabilities could be computed by counting in a large corpus —by maximum likelihood estimation. Once again, these probabilities compose a probability distribution which determines the probability of a given word conditioned to the previous one. This is no other than a language model.

Chapter 2

Online Resources

2.1 Free Text Collections

1. The Wikipedia[1] is an excellent collection with crowdsourced text in multiple languages. Its encyclopedic contents represent a large scale comparable corpus and its metadata allows for analysing multiple characteristics of writing and collaborative creation. Sibling Wikimedia projects, such as the Wiktionary and Wikinews are worth considering as well.

2. Project Gutenberg [2] contains $60k+$ free ebooks, also available in plain text format.

3.

2.2 Code

1. The code from Lane et al. (2019) is available at `https://github.com/totalgood/nlpia`

[1] `http://www.wikipedia.org`

[2] `https://www.gutenberg.org/wiki/Main_Page`

Bibliography

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