Machine Learning for Natural Language Processing

The *Why* and *What* of NLP

Lecture 1
February 9th 2021, ENSAE Paris

Benjamin Muller

INRIA Paris - ALMANACH
`benjamin.muller@inria.fr`

This course

- We will cover techniques used in industry (Facebook, Google, Apple, Twitter...)
- Introduce core ideas at the basis of modern NLP algorithms
- Focus on machine learning applied to NLP

Goal: Provide a toolkit of concepts and methods to describe and tackle NLP problems in real-life.

Course Logistics

- 6 sessions
- 1h30 lecture followed by 1h30 applied *lab* session
- Online session interacting on Teams and Wooclap [1]
- Lecture and Lab Material nlp-ensae.github.io

[1] https://app.wooclap.com/auth/register/ENSAE

Course Evaluation

- **Final Project**: Solve a NLP problem and present your experiments synthetically (4/5 of the final grade)
  - Outcome: Self-contained **notebook** uploaded to **github** and **google colab** + 2 pages report
- **Quizz** happening during the lectures on wooclap (1/5 of the final grade)

Course Outline

1. The Why and What of Natural Language Processing - B. Muller
2. Representing text with vectors - G. Guibon
3. Task specific Modeling of Text - G. Guibon
4. Neural Natural Language Processing - G. Guibon
5. Language Modeling - B. Muller
6. Transfer Learning with Neural Modeling for NLP - B. Muller

Today session outline

- Why doing NLP?
- Why is language hard to model? The 4 challenges of NLP.
- What is Natural Language Processing?
  - A non-exhaustive definition of NLP
  - A brief history of NLP
  - NLP in three pipelines

Why Natural Language Processing?

Survival Guide

- Always asks *why ?*
- Be focused: Focus means being active (ask questions, take notes, ...)
- Practice (code) often

Why Natural Language Processing ?

What do we use language for?

- We communicate using language
- We think (mostly) with language
- We tell stories in language
- We describe our theories in language

Why NLP ?

- Information Retrieval (search, recommendation, aggregation)
- Better interfaces (human-computer, human-human interface)
- Better understanding of our thinking process and of language itself

Why Natural Language Processing ?

Amount of online textual data...[2]

- 70 billion web-pages online (1.8 billion websites)
- 55 million Wikipedia articles (open source)

...growing at a fast pace

- 8000 tweets/second
- 3 million mail / second (60% spam)

[2] internet live stats

Why Natural Language Processing ?

Potential Users of Natural Language Processing

- 7.8 billion people use some sort of language (January 2021)
- 4.8 billion people connected (January 2019)

Why Natural Language Processing ?

What products ?

- Search: +2 billion people use Google, 700 millions people use Baidu
- Social Media: +3 billion users of Social media (Facebook, Instagram, WeChat, Twitter...)
- Voice assistant: +100 million users (Alexa, Siri, Google Assistant)

Why Natural Language Processing ?

Myth or Reality of ""Artificial General Intelligence"" ?

- Billions $ invested in research in AI
- Fast adoption paced : Incremental progress in research is quickly spreading to users
- Myth or Reality of AGI ?

Why Natural Language Processing ?

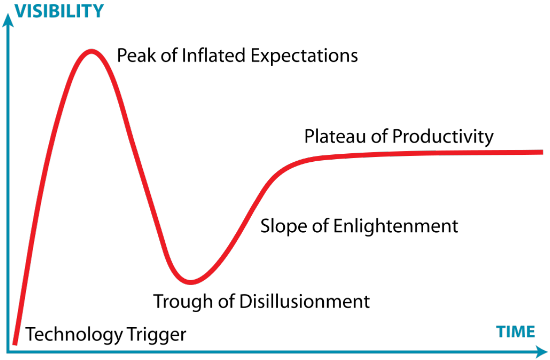


Figure: The Hype Cycle

Objective of the course

- Toolkit for how to approach any NLP problem
- Get a theoretical understanding of most recent NLP models
- Grasp the challenges (model, data, computation, time...) of NLP projects

Why is language hard to model?

A Definition of Language

Definition 1: *Language is a means to communicate, it is a* ***semiotic*** *system. By that we simply mean that it is a set of signs. A sign is a pair consisting in [...] a signifier (or exponent) and a signified (or meaning).*

Definition 2: *A sign consists in a* ***phonological*** *structure, a* ***morphological*** *structure, a* ***syntactic*** *structure and a* ***semantic*** *structure*[3]

[3] (Kracht)

Quick introduction to linguistics

Analysis in context

Extra-linguistic context

*Found **him** in the street inside a bag. I think **he** is happy with his new life*

Linguistic context

— *You know what? John gave Peter a Christmas present yesterday*
— *Wow, was **he** surprised? What was **it** like?*
— *Surprisingly good. **He** spent quite a bit on **it**.*

Sentence-level analysis

Semantic level

The landlord$_{\text{SPEAKER}}$ has not yet **REPLIED**$^{\text{Communication\_response}}$ in writing$_{\text{MEDIUM}}$ to the tenant$_{\text{ADRESSEE}}$ objecting the proposed alterations$_{\text{MESSAGE}}$. *DNI*$_{\text{TRIGGER}}$

Syntactic level

John saw a dog yesterday which was a Yorkshire Terrier

Morphological level

brav+itude, bio+terror-isme/-iste, skype+(e)r

mang-er-i-ons = MANGER+cond+1pl

Phonological level

International Phonetic Alphabet

[aɪ pʰiː eɪ]

Graphemic level

enough, cough, draught, although, brought, through, thorough, hiccough

6 Levels of analysis

- Phonological level
- Morphological Level
- Syntactic level
- Semantic Level
- Linguistic Context
- Extra-linguistic level

$\rightarrow$ All NLP problems can be split between one or several of these levels of analysis

Why is language hard ?

- Language **diversity**
- Language **variation**
- Language **ambiguity**
- Language **sparsity**

Phonological Diversity

- Syllables are formed of phoneme sequences
- In most languages, some syllables are valid, some are not

E.g : Japanese has only one *liquid* phoneme /r/

Phonological Diversity

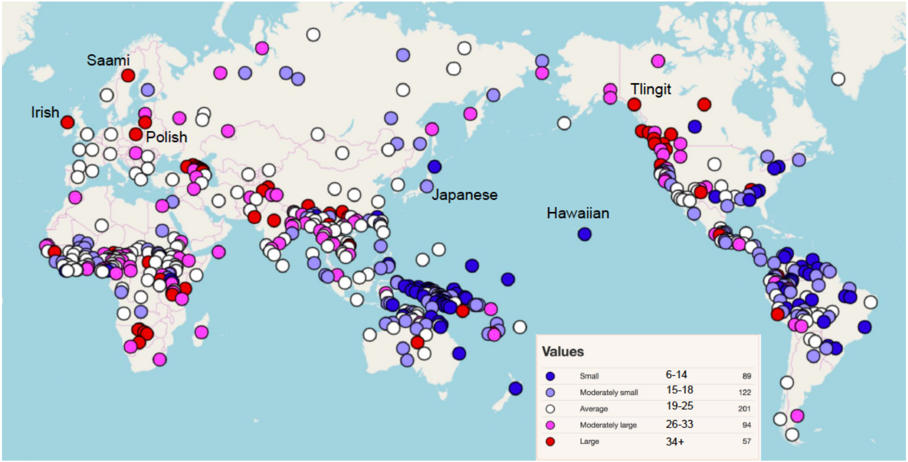


Figure: Consonant Inventory (size of the set of consonants) Source: The World Atlas of Language Structures

Morphological Diversity

- Analytic and isolating languages

  Each word carries exactly one meaning (e.g Chinese)
- Synthetic languages
  - Agglutinative

    Each word can have several morphs, each carrying one meaning
    e.g : Turkish el-ler-imiz-in (HAND-pl-poss1pl-genitive) 'of our hands'
  - Fusional :

    Each word can have several morphs, each carrying one or more meanings, of which (generally) only one lexical morph
  - Polysynthetic :

    Each word can have several lexical or grammatical morphs

Morphological Diversity

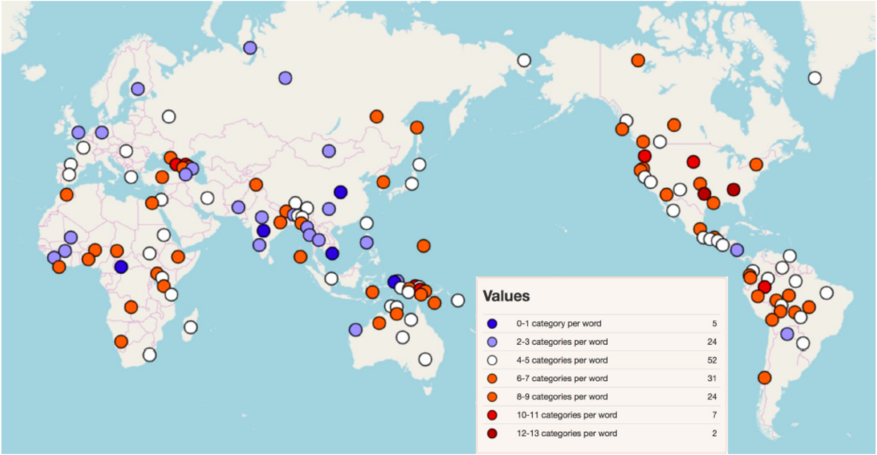


Figure: Number of Category per Word
Source: The World Atlas of Language Structures

Syntactic Diversity

- Word order differs across languages
- Word order degree of freedom also differs across languages
- We characterize word orders with : Subject - Verb - Object order

Syntactic Diversity

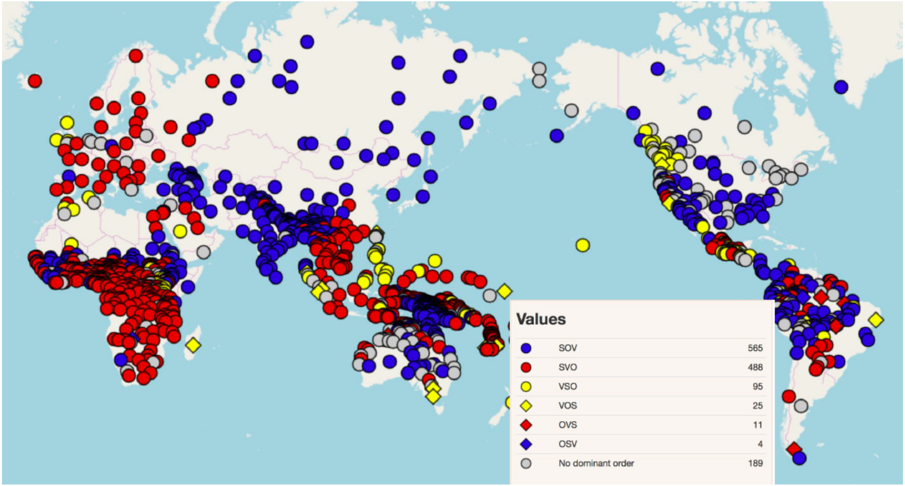


Figure: Order of Subject (S) Object (O) and Verb (V)
Source: The World Atlas of Language Structures

Morphology and Syntax

- Word orders freedom and morphology are usually related
- The more freedom in word orders
  $\rightarrow$ the less information is conveyed by word positions
  $\rightarrow$ the more information should be included in the "symbols"
  $\rightarrow$ the richer the morphology
- e.g English vs. Russian (object indicated with -ей):

cats eat mice

Кошки едят мышей

Мышей едят кошки.

Едят кошки мышей.

Едят мышей кошки.

Constrained word order
Limited or no morphological marking

(Relatively) free word order
Rich morphology

Semantic Diversity

- Words partition the semantic space
- This partition is very diverse across language

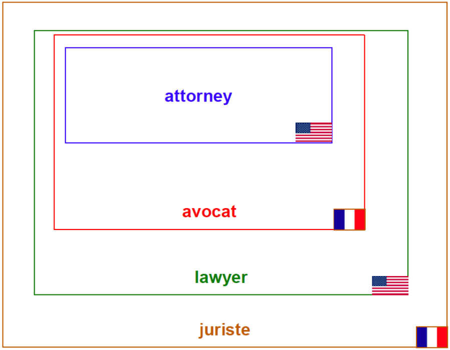
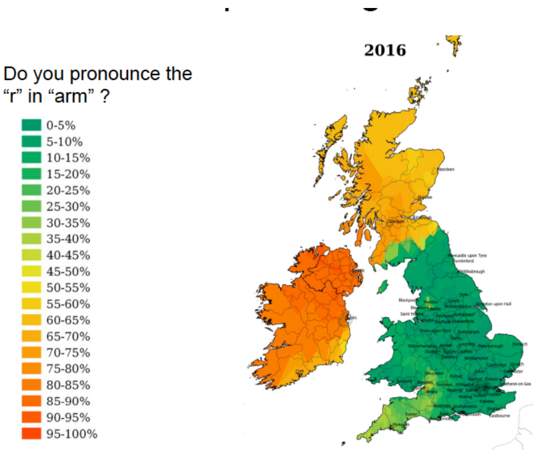


Figure: Semantic partitioning between English(US) and French: *laywer* vs *avocat*. (2)

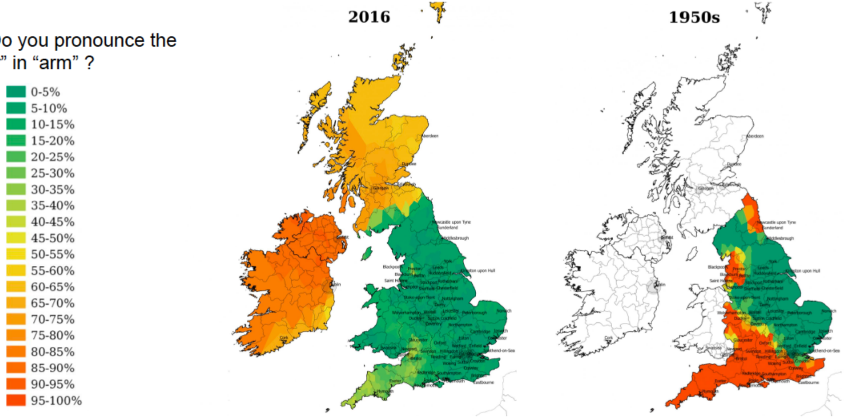
Variation

- Variation at all level of analysis (phonological, morphological, syntactic, semantic)
- Building NLP with such variance is a great challenge


2016
Do you pronounce the
"r" in "arm" ?
0-5%
5-10%
10-15%
15-20%
20-25%
25-30%
30-35%
35-40%
40-45%
45-50%
50-55%
55-60%
60-65%
65-70%
70-75%
75-80%
80-85%
85-90%
90-95%
95-100%

Phonetic Variation

Do you pronounce the
“r” in “arm” ?



Spelling Variation

anagement maagement maanagement
maangement magagement magement
mamagement mamangement manaagement manaement
managaement manageement manageemnt managegment
managemaent managemant managememt managemen managemenet
managementt managemet managemetn managemnent managemnet
managemnt managemrnt managemt managenent managenment managent
managerment managhement managmeent managmement managment managnment
manament manamgement mananement manangment manasgement
manegement manegment mangaement mangagement mangagment
mangament mangement manggement mangment
mangmt menagement mgmt mgnt
mnagement mngmnt mngmt

Figure: Spelling variation of "management" found in Social Media data (2)

T'as vu il l'a bien cherché wsh #AperoChezRicard
> +10000, shah!
> > tabuz, lavé rien fé
> > > ki ca ? le mec ou son chien ?
> > > > Wtf is wrong with him ? #PETA4EVER
> > > > > ki ca ? le chien ?
> > > > > > looool

BING translation:
You saw coming it #AperoChezRicard wsh
> +10000, shah!
> > tabuz, washed anything fe
> > > Ki ca? the guy or his dog?
> > > > WTF is wrong with him? #PETA4EVER
> > > > > Ki ca? the dog?
> > > > > > looool

Figure: Non-Canonical Tweet Translated by Bing (2)

Ambiguity

- Most linguistic observations (speech, text) are open to several interpretations
- We (Humans) disambiguate/find the correct interpretation using all kind of signals (linguistic and extra linguistic)
- Ambiguity can appear at all levels of analysis

Syntactic Ambiguity

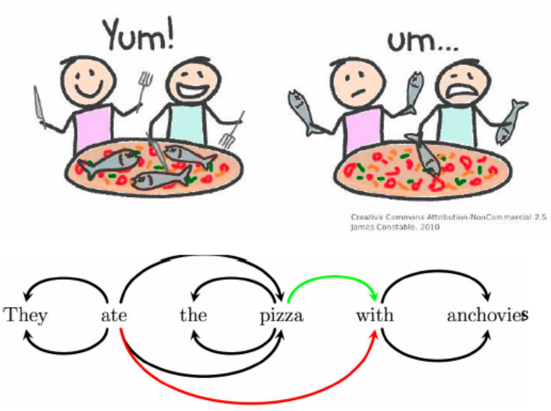


Figure: Syntactic Ambiguity (2)

Semantic Ambiguity

Figure: Semantic Ambiguity (2)

- Name entity
- Polysemy (man)
- Object/Color (cherry)
- Object/Informal (e.g. the book)

Ambiguity examples

- Ambiguity! Some examples of ambiguous headlines:

Iraqi head seeks arms

Enraged cow injures farmer with axe

San Jose cops kill man with knife

Miners refuse to work after death

Two Soviet ships collide, one dies

Dealers will hear car talk at noon

- Ambiguity can be lexical, syntactic, pragmatic

Ambiguity examples

Human: Are there direct flights from Paris to Santiago?

Bot: Yes, there is an Air France flight leaving at 11:40PM.

Human: How long does it takes to go there?

Bot: The flight takes 14h35m.

Human: How much would that cost?

- Needs discourse knowledge, domain knowledge, linguistic knowledge

Sparsity

Data Sparsity is when many entities (words, morphemes, n-grams, ...) in a corpus have very low observed frequency

Sparsity is the consequence of :

- **Combinatorial** structure of language
  Combining meaningless sounds into meaningful morphemes or words and meaningful phrases into sentences.[4]
- **Zipfian** structure of language

NB : Sparsity is one of the greatest challenge of NLP

[4] The Origin of Speech, Hockett et. al 1960

Zipf's law

Zipf's law can describe many language phenomenons.

Definition:
f_w frequency of entity w
k frequency rank of entity w

$$f_w(k) \; \alpha \; \frac{1}{k^{\theta}}$$

Comments

- Zipf law is a Power relation between the rank and frequency
  The most frequent entities are <u>much</u> more frequent than the less frequent ones
- Under Zipf law $log(f_w)$ and $log(k)$ are linearly related

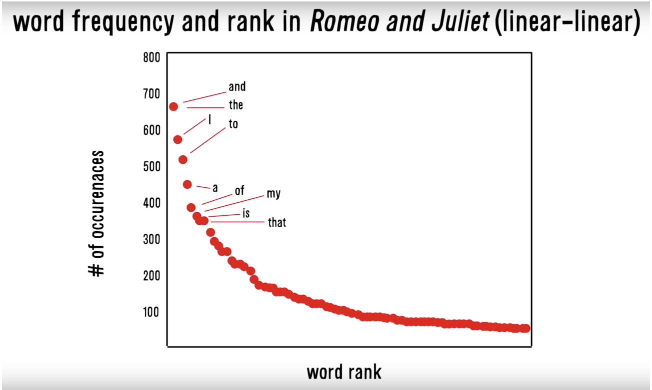
Zipfian structures in Language

Zipf law can be found in many phenomenons in nature.

In Language

- Word frequency
- Syntactic structures frequency

Zipfian structure of Language: Word Frequency



Zipfian structure of Language : Lexicon

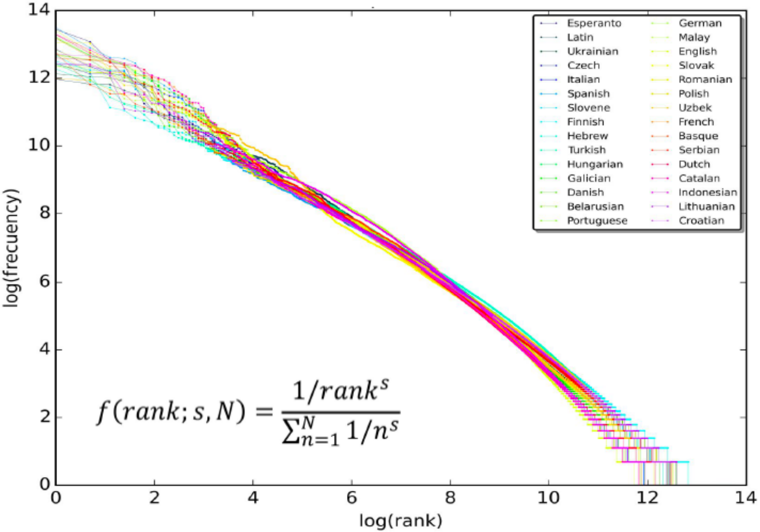


Figure: rank vs frequency for the top 10M words in wikipedia

Zipfian structure of Language : Syntax

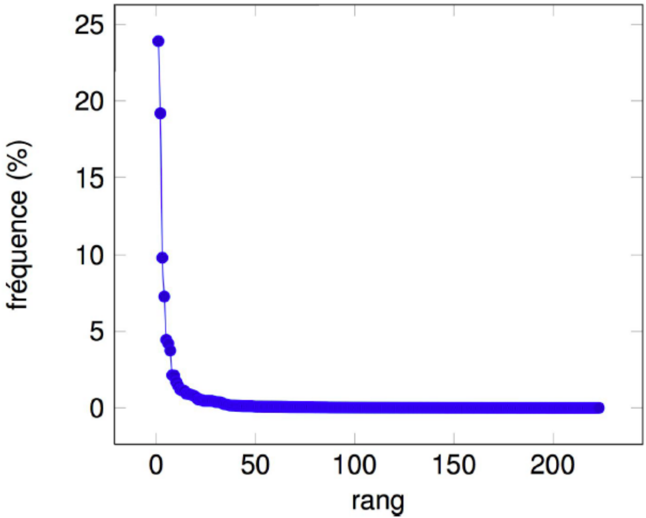


Figure: rank vs frequency for automatically parsed corpus

Zipf's law and Sparsity

- The Zipf's law is a long tail distribution
- Many entities (words, syntactic structure,...) have a very low frequency

  $\rightarrow$ sparsity

What is Natural Language Processing ?

What is Natural Language Processing ?

- Process, analyze and/or produce natural language
- Interact with computers using natural language
- Natural language 'understanding':
  - language as input $\mapsto$ "information" as output
- Natural language generation:
  - "information" as input $\mapsto$ language as output
- Sometimes, both: machine translation, summarization, question answering
- Strongly related fields:
  - machine learning,
  - deep learning
  - (computational) linguistics

What is Natural Language Processing ?

In a nutshell, NLP consists in handling the complexities of language systematically "to do something"

- Raw Text $\rightarrow$ Structured Information
- Raw Text $\rightarrow$ Controlled Text

Brief History of NLP

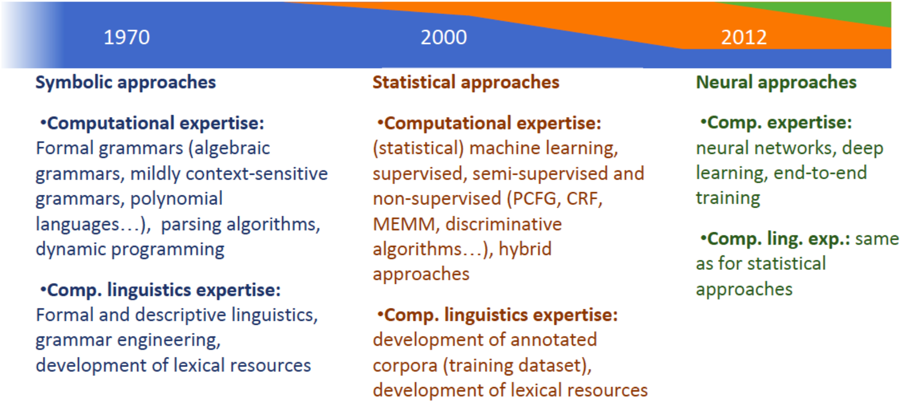
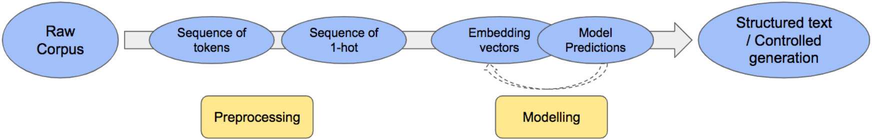
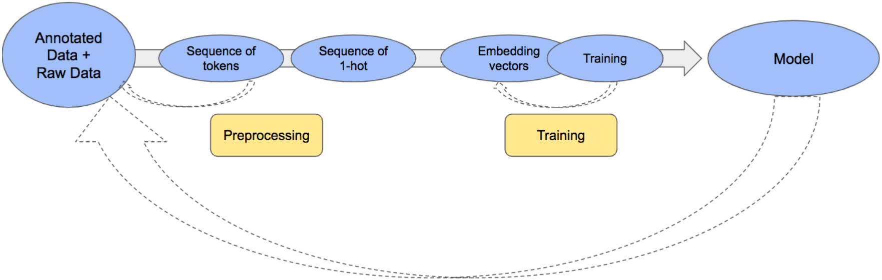


Figure: Brief history of NLP (2)

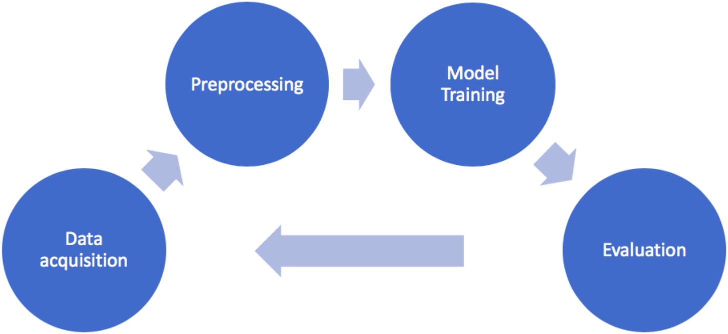
NLP prediction pipeline



NLP training pipeline



NLP in the real-world



- Building NLP systems is an iterative cycle...
- Composed of **Human & Machine Learning**

Outline of the course

1. The Why and What of Natural Language Processing
2. Representing text with vectors
3. Modeling textual data
4. Neural Natural Language Processing
5. Language Modelling
6. Transfer Learning with Neural Modeling for NLP

References I

[Kracht] Kracht, M. Introduction to linguistics.

[2] Sagot, B. (2019). Algorithms for speech and natural language processing, lectures ens-saclay.