

Machine Learning for Natural Language Processing

The *Why* and *What* of NLP

Session 1

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 course followed by 1h30 applied *lab* session
- Course Material nlp-ensae.github.io

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 NLP in the "real-world"

Course Evaluation

- Project: Implement NLP algorithm (list of projects given later)
- Outcome : Self-contained **notebook** uploaded to **github** or **google colab**

Today session outline

- Why is language hard ? 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 do with language ?

- 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...¹

- 60 billion web-pages online (1.7 billion websites)
- 48,731,540 Wikipedia pages (open source)

...growing at a fast pace

- 8000 tweets/second
- 2.8 million mail / second (60% spam)
- +500 users / second

¹internet live stats

Why Natural Language Processing ?

Potential Users of Natural Language Processing

- 7.7 billion people use some sort of language (January 2019)
- 4.4 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, Cortona, 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 ?

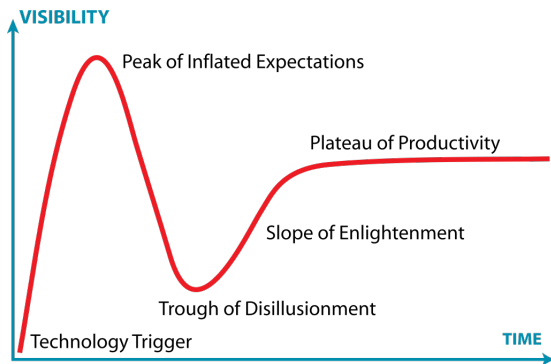


Рис.: Hype Cycle

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 ?


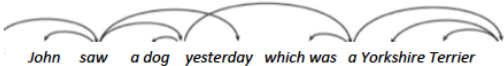
A Definition of Language

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** structure²*

²(Kracht)

Quick introduction to linguistics

Analysis in context	Extra-linguistic context	 <i>Found him in the street inside a bag. I think he is happy with his new life</i> <small>http://img.com/gag/4r1n0wq/Found-him-in-the-street-inside-a-bag-I-think-he-is-happy-with-his-new-life</small>
	Linguistic context	<ul style="list-style-type: none"> — 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.
	Semantic level	The landlord ^{SPEAKER} has not yet REPLIED ^{Communication_response} in writing ^{MEDIUM} to the tenant ^{ADDRESSEE} objecting the proposed alterations ^{MESSAGE} . ^{DNI} ^{TRIGGER}
	Syntactic level	 John saw a dog yesterday which was a Yorkshire Terrier
Sentence- level analysis	Morphological level	<i>brav+itude, bio+terror-isme/-iste, skype+(e)r</i> <i>mang-er-i-ons = MANGER+cond+1pl</i>
	Phonological level	International Phonetic Alphabet [aɪ p ^h i: eɪ]
	Graphemic level	<i>enough, cough, draught, although, brought, through, thorough, hiccough</i>

(Sagot)

Quick introduction to linguistics

6 Levels of analysis

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

→ All NLP problems can be split between one or several of these level of analysis

Why is language hard ?

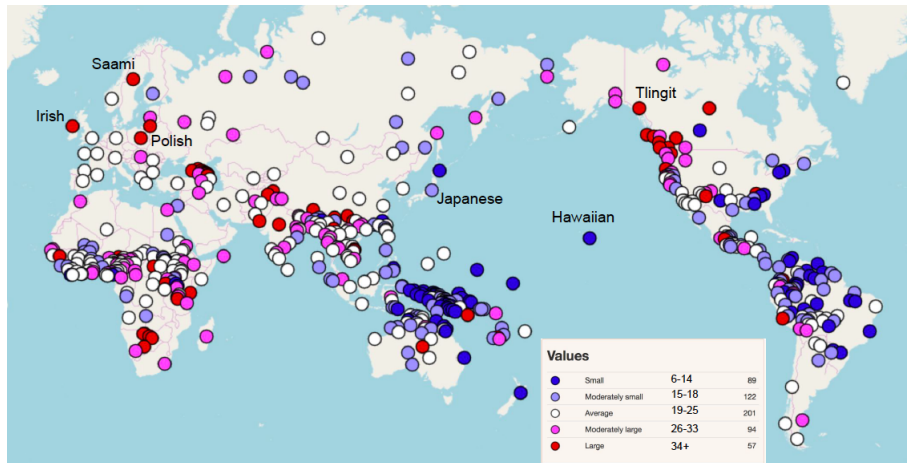
- Language **diversity**
- Language **variation**
- Language **ambiguity**
- Language **sparsity**

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/

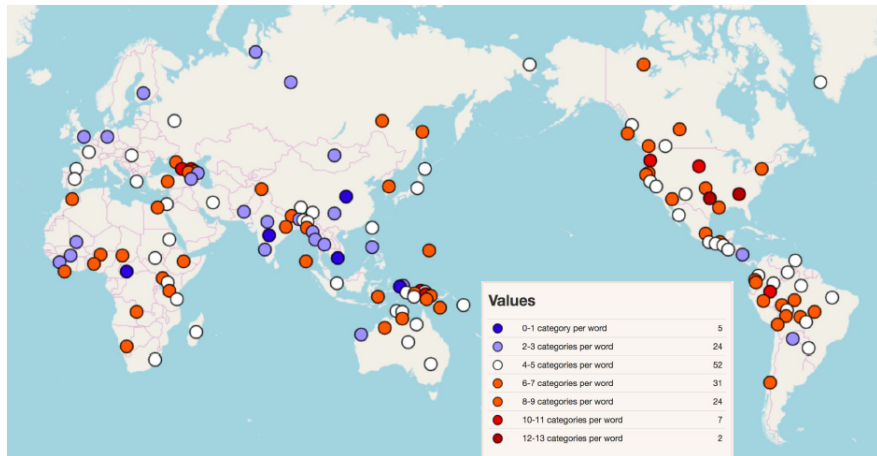
Phonological Diversity



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

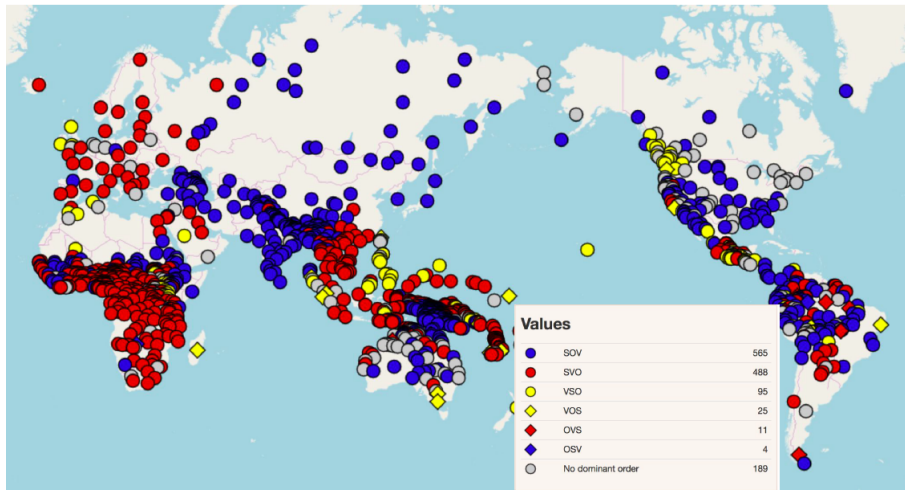
Morphological Diversity



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



Morphology X Syntax

- Word orders freedom and morphology are usually related
- The more freedom in word orders
 - the less information is conveyed by word positions
 - the more information should be included in the "symbols"
 - 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

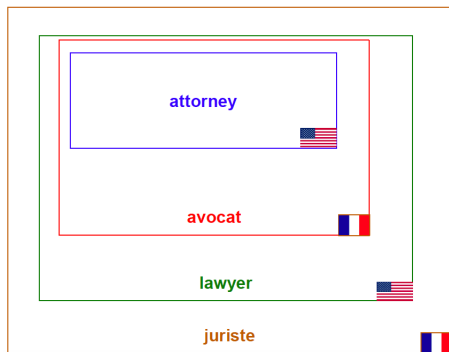


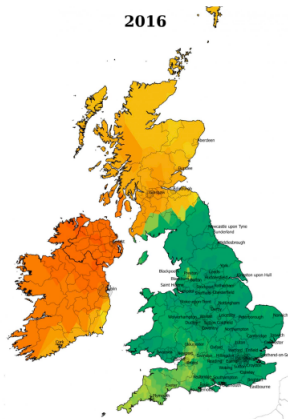
Рис.: Semantic partitioning between English(US) and French:
lawyer vs *avocat*. (Sagot)

Variation

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

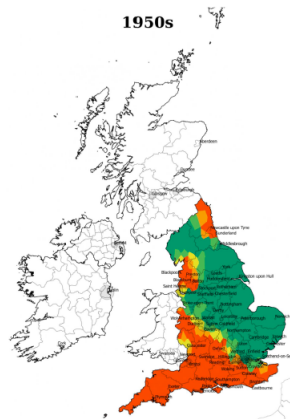
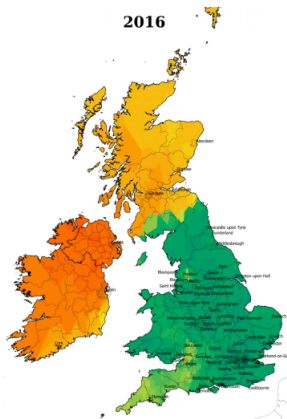
Phonetic Variation

Do you pronounce the
“r” in “arm” ?



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 managementt managemen managemenet
managementt managemet managemetn managemnnt managemnet
managemnt managemrnt managemt managenent managenment managent
managerment managhement managmeent managmement managment managmnt
manament manamgement mananement manangment manasgement
manegement manegment mangaement mangagement mangagment
mangament mangement manggement mangment
mangmt menagement mgmt mgnt
mnagement mngmnt mngmt

Sociolinguistic Variation



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 ?

> loool

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

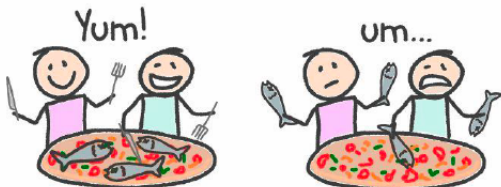
> Ki ca? the dog?

> loool

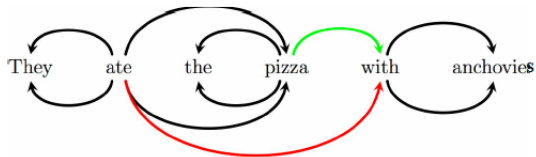
Ambiguity

- Most linguistic observations (speech, text) are open to several interpretation
- 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



Creative Commons Attribution-NonCommercial 2.5
James Constable, 2010



Semantic Ambiguity



The screenshot shows the Wikipedia page for "Michael Jordan (disambiguation)". At the top, there's a navigation bar with "Article" and "Talk" tabs, and a search bar. Below the navigation bar, the title "Michael Jordan (disambiguation)" is displayed. The main content area states: "From Wikipedia, the free encyclopedia" and "Michael Jordan (born 1963) is an American basketball player." It then says "Michael Jordan or Mike Jordan may also refer to:" followed by a list of disambiguation entries under the heading "People". The list includes: Michael Jordan (footballer) (born 1966), English goalkeeper (Arsenal, Chesterfield, Lewes); Mike Jordan (racing driver) (born 1958), English racing driver; Mike Jordan (baseball, born 1863) (1863–1940), baseball player; Michael Jordan (American football) (born 1992), American football cornerback; Michael-Hakim Jordan (born 1977), American professional basketball player; and Michal Jordán (born 1990), Czech ice hockey player. To the right of the list is a "Contents" table of contents with links to "People", "Sports", "Other people", and "Other uses". On the left side of the page, there's a sidebar with various Wikipedia navigation links like "Main page", "Contents", "Featured content", etc.

WIKIPEDIA
The Free Encyclopedia

Not logged in - Talk - Contributions - Create account - Log in

Article Talk

Read Edit View history

Search Wikipedia

Michael Jordan (disambiguation)

From Wikipedia, the free encyclopedia

Michael Jordan (born 1963) is an American basketball player.

Michael Jordan or **Mike Jordan** may also refer to:

People

Sports

- Michael Jordan (footballer) (born 1966), English goalkeeper (Arsenal, Chesterfield, Lewes)
- Mike Jordan (racing driver) (born 1958), English racing driver
- Mike Jordan (baseball, born 1863) (1863–1940), baseball player
- Michael Jordan (American football) (born 1992), American football cornerback
- Michael-Hakim Jordan (born 1977), American professional basketball player
- Michal Jordán (born 1990), Czech ice hockey player

Other people

- Michael B. Jordan (born 1987), American actor
- Michael Jordan (insolvency baron) (born 1931), English businessman
- Michael Jordan (Irish politician), Irish Farmers' Party TD from Wexford, 1927–1932
- Michael I. Jordan (born 1956), American researcher in machine learning and artificial intelligence
- Michael H. Jordan (1936–2010), American executive for CBS, PepsiCo, Westinghouse
- Michael Jordan (mycologist), English mycologist

Contents

- People
 - 1.1 Sports
 - 1.2 Other people
- Other uses
- See also

- 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.*³
- **Zipfian** structure of language

NB : Sparsity is one of the greatest challenge of NLP

³The Origin of Speech, Hockett et. al 1960

Zipf's law

Zipf ' s law can describe many phenomenons of language.

Definition:

f_w frequency of entity w

k frequency rank of entity w

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

Comments

- Zipf law is a Power relation between the rank and frequency
The most frequent entities are much 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 and rank in *Romeo and Juliet* (linear-linear)



Zipfian structure of Language : Lexicon

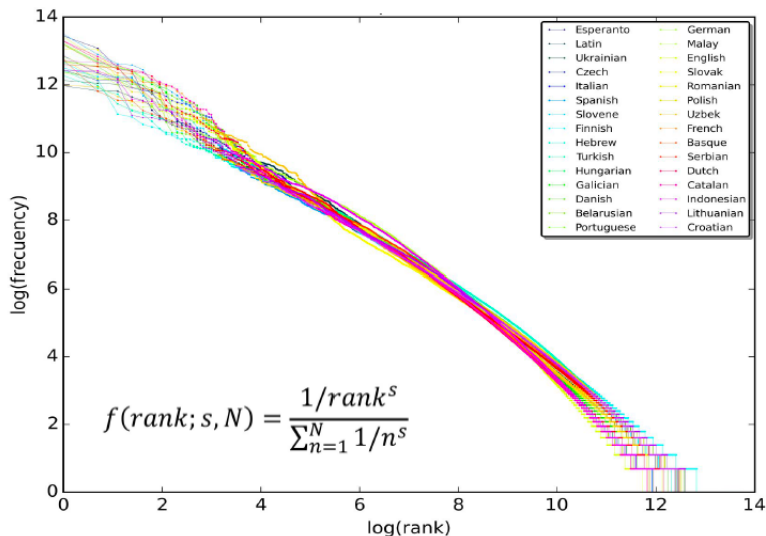


Рис.: rank vs frequency for the top 10M words in wikipedia

Zipfian structure of Language : Syntax

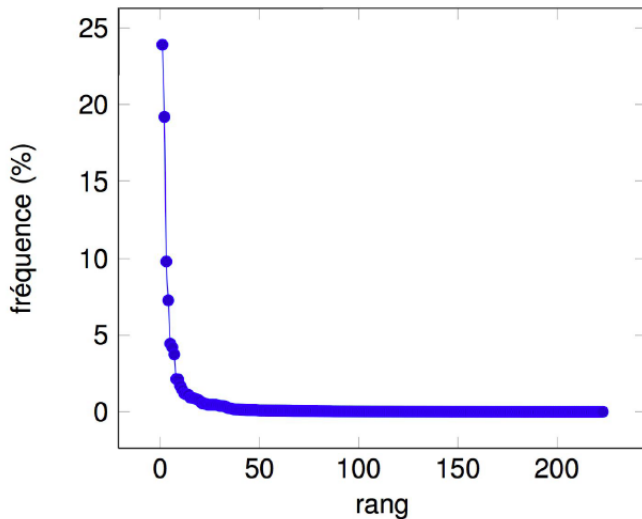


Рис.: rank vs frequency for automatically parsed corpus

Zipf's law and Sparsity

- The Zipf's law is a long tail distribution
- Many entities (words, syntactic structure,...) have a very low frequency
→ 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,
 - artificial intelligence,
 - 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 → Structured Information
- Raw Text → Controlled Text

Brief History of NLP

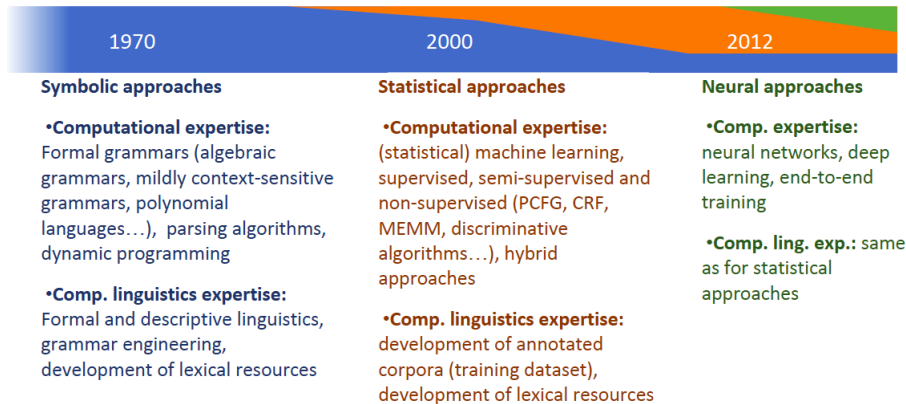
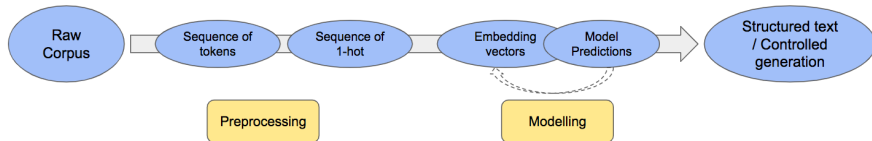
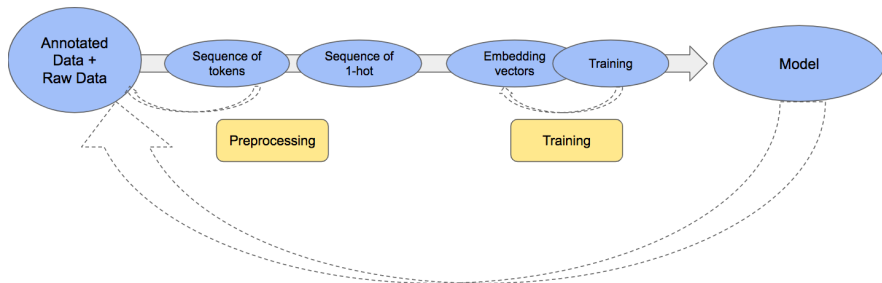


Рис.: (Sagot)

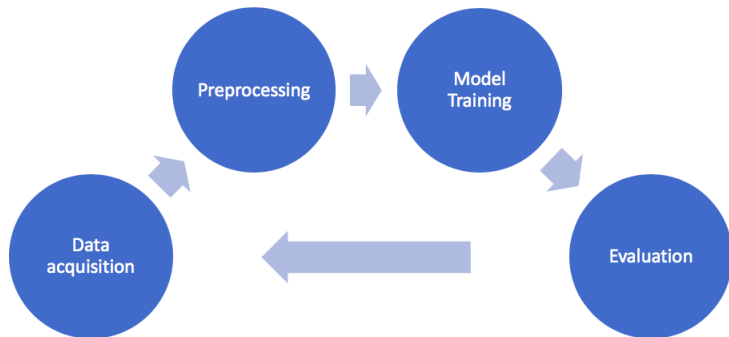
NLP prediction pipeline



NLP training pipeline



NLP in the real-world



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

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 NLP in the "real-world"

References I

[Kracht] Kracht, M. Introduction to linguistics.

[Sagot] Sagot, B. Algorithms for speech and natural language processing,
course ens-saclay.