Lecture Series on Predictive Language Models

El Mehdi ISSOUANI

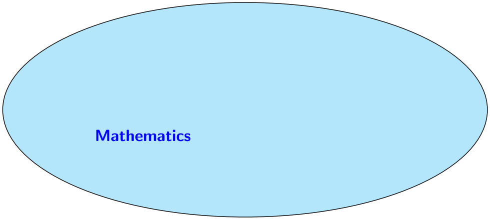
Post-doc: Laboratory of Applied Mathematics of Compiègne (LMAC)
University of Technology of Compiègne (UTC)

Thesis obtained in June 2023 in MODAL'X at University Paris Nanterre

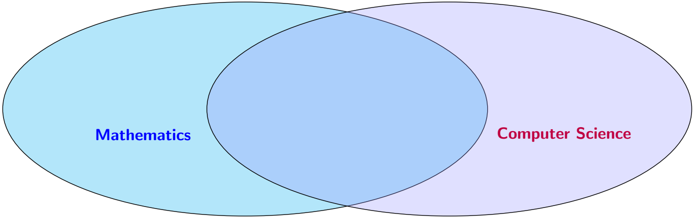
Wednesday 5th February, 2025

Lecture 1 - Introduction to Textual Data and NLP: A Multidisciplinary Perspective

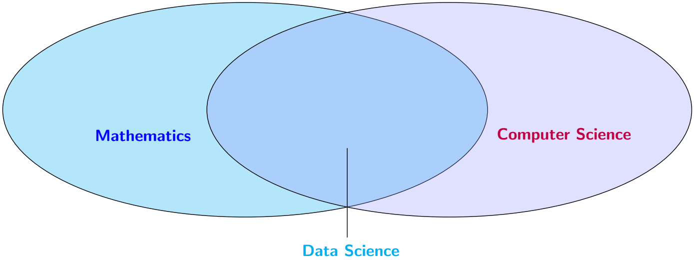
Cross-Disciplinary Approaches



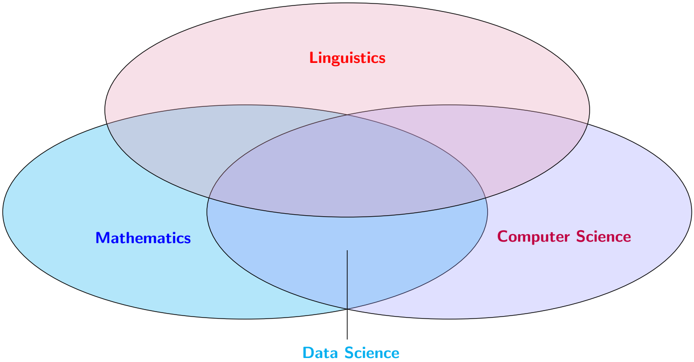
Cross-Disciplinary Approaches



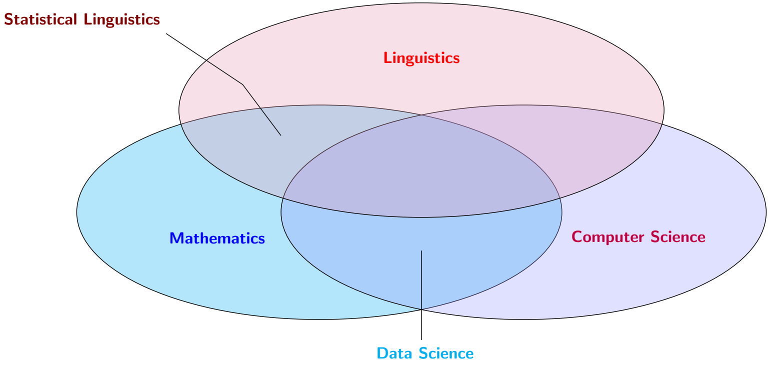
Cross-Disciplinary Approaches



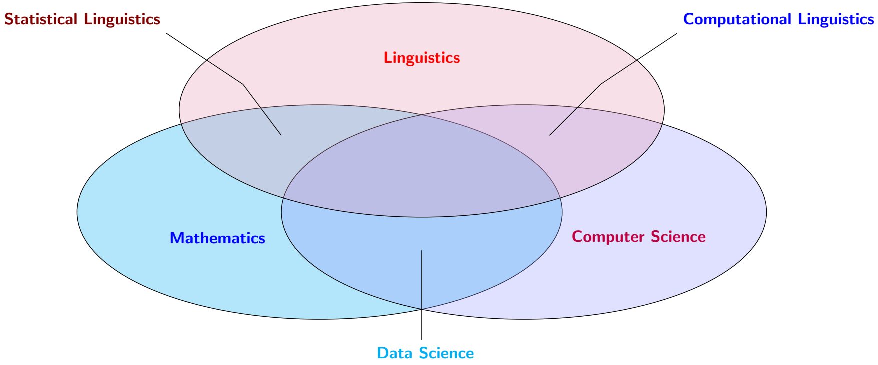
Cross-Disciplinary Approaches



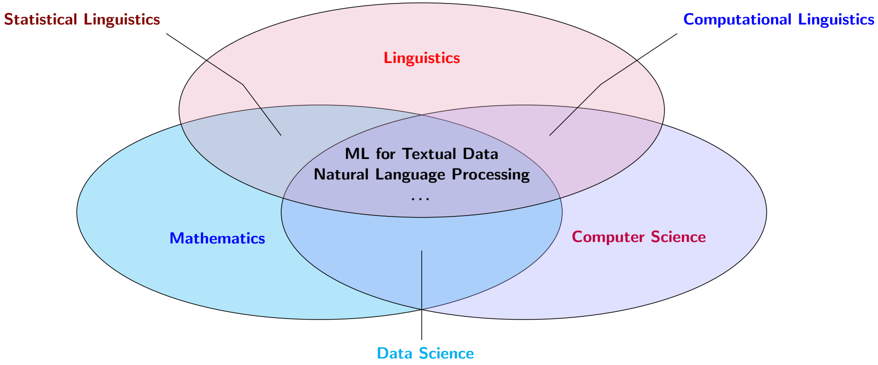
Cross-Disciplinary Approaches



Cross-Disciplinary Approaches



Cross-Disciplinary Approaches



Outline

Linguistic Definitions

Key concepts :

- Corpus
- Linguistic unit[1] (word, sentence, ..)
- Levels of analysis
- Syntactic dependency analysis

[1] Also called lexical units: Defining the Granularity of the Observations

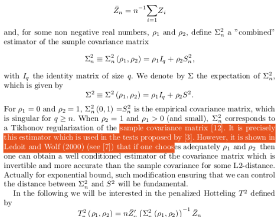
Linguistic Definitions

Key concepts :

- Corpus
- Linguistic unit[1] (word, sentence, ..)
- Levels of analysis
- Syntactic dependency analysis
- Spatial representation method (STR)
- Identification of *Topics*
- Machine Learning
- ...

[1] Also called lexical units: Defining the Granularity of the Observations

Corpus



$$\bar{Z}_n = n^{-1}\sum_{i=1} Z_i$$

and, for some non negative real numbers, ρ_1 and ρ_2, define Σ_n^2 a "combined" estimator of the sample covariance matrix

$$\Sigma_n^2 \equiv \Sigma_n^2\left(\rho_1, \rho_2\right) = \rho_1 I_q + \rho_2 S_n^2,$$

with I_q the identity matrix of size q. We denote by Σ the expectation of Σ_n^2, which is given by

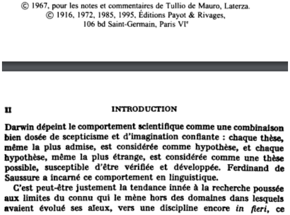
$$\Sigma^2 \equiv \Sigma^2\left(\rho_1, \rho_2\right) = \rho_1 I_q + \rho_2 S^2.$$

For $\rho_1 = 0$ and $\rho_2 = 1$, $\Sigma_n^2\left(0,1\right) = S_n^2$ is the empirical covariance matrix, which is singular for $q \geq n$. When $\rho_2 = 1$ and $\rho_1 > 0$ (and small), Σ_n^2 corresponds to a Tikhonov regularization of the sample covariance matrix [12]. It is precisely this estimator which is used in the tests proposed by [3]. However, it is shown in Ledoit and Wolf (2000) (see [7]) that if one chooses adequately ρ_1 and ρ_2 then one can obtain a well conditioned estimator of the covariance matrix which is invertible and more accurate than the sample covariance for some L2-distance. Actually for exponential bound, such modification ensuring that we can control the distance between Σ_n^2 and S^2 will be fundamental.

In the following we will be interested in the penalized Hotteling T^2 defined by

$$T_n^2\left(\rho_1, \rho_2\right) = n\bar{Z}_n'\left(\Sigma_n^2\left(\rho_1, \rho_2\right)\right)^{-1}\bar{Z}_n$$

Figure: a pdf document written in latex



© 1967, pour les notes et commentaires de Tullio de Mauro, Laterza.
© 1916, 1972, 1985, 1995, Éditions Payot & Rivages,
106 bd Saint-Germain, Paris VIᵉ

II INTRODUCTION

Darwin dépeint le comportement scientifique comme une combinaison bien dosée de scepticisme et d'imagination confiante : chaque thèse, même la plus admise, est considérée comme hypothèse, et chaque hypothèse, même la plus étrange, est considérée comme une thèse possible, susceptible d'être vérifiée et développée. Ferdinand de Saussure a incarné ce comportement en linguistique.

C'est peut-être justement la tendance innée à la recherche poussée aux limites du connu qui le mène hors des domaines dans lesquels avaient évolué ses aïeux, vers une discipline encore *in fieri*, ce

Figure: Scanned document

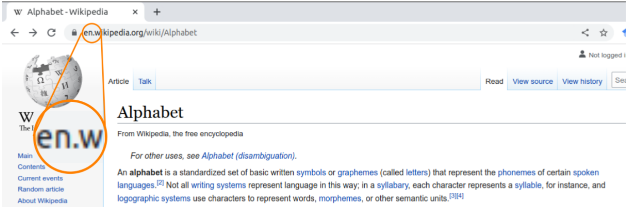
- Here are some English tagged corpora

  Brown corpus and Penn Treebank corpus [MMS93]
- Here are examples of parallel corpora

  Europarl, ParaSol, EUR-Lex

Corpus

- In our case, we'll use internet
- Digital textual resources and encyclopedias
  - linguee
  - Wikipedia (example: the simple version and the standard version in English)
- Local databases such as nltk


Alphabet - Wikipedia
en.wikipedia.org/wiki/Alphabet
en.w
Not logged i
Article
Talk
Read
View source
View history
Alphabet
From Wikipedia, the free encyclopedia
For other uses, see Alphabet (disambiguation).
An alphabet is a standardized set of basic written symbols or graphemes (called letters) that represent the phonemes of certain spoken languages.[2] Not all writing systems represent language in this way; in a syllabary, each character represents a syllable, for instance, and logographic systems use characters to represent words, morphemes, or other semantic units.[3][4]
Main
Contents
Current events
Random article
About Wikipedia


fr.wikipedia.org/wiki/Alphabet
Wikipédia
L'encyclopédie libre
Rechercher sur Wikipédia
fr.w
Alphabet
Article
Discussion
Lire
Pour les articles homonymes, voir Alphabet (homonymie).
Cet article ne cite pas suffisamment ses sources (novembre 2016).
En pratique : Quelles sources sont attendues ? Comment ajouter mes sources ?
Un alphabet (de alpha et bêta, les deux premières lettres de l'alphabet grec) est un système d'écriture constitué d'un ensemble de symboles dont chacun représente, par exemple, un des phonèmes d'une langue.
Introduction [modifier | modifier le code]
Chacun des symboles d'un alphabet ou graphèmes, est appelé « lettre ». Dans les cas les plus simples, chaque lettre correspond à un phonème de la langue et inversement. Certaines lettres peuvent recevoir un ou plusieurs diacritiques afin d'étendre le stock de graphèmes si celui-ci est insuffisant pour noter les sons de la langue ou permettre d'éviter les ambiguïtés. De la même manière, un alphabet peut être étendu par l'utilisation de digrammes ou encore de lettres supplémentaires.
Histoire
Alphabets
Alphabets récents et de transcription
Alphabets imaginaires
Alphabets littéraires
Système de codage
Phrases alphabétiques
Notes et références
Voir aussi
Bibliographie
Articles connexes

Levels of linguistic analysis

In linguistics, there are different analysis levels[2]:

- Morphological
- Lexical
- Syntactic
- Semantic
- Pragmatic

[2] In this course, we will not consider the phonetic aspect or the phonological level, neither the pragmatic

Text Generation Example - Translation

Here are some well-known **text translation** methods:

- Syntactic Translation
- Lexical Translation
- Hybrid Translation
- Statistical Machine Translation

Text Generation Example - Translation

Here are some well-known **text translation** methods:

- Syntactic Translation (direct speech, subject-verb structure, etc.)
- Lexical Translation (synonyms, definitions, etc.)
- Hybrid Translation (combination of syntactic and lexical approaches)
- Statistical Machine Translation (seq-to-seq)

Representations

Two Approaches:

- High-dimensional sparse vectors (discrete representation)

MaxEnt and Feature-Based Models

$$\begin{cases} x = (0, 1, 0, ..., 1, 0) \in \mathbb{R}^q \\ q \text{ is very large, } O(q) \approx \text{hundreds of thousands or millions.} \end{cases}$$

- Low-dimensional dense vectors (continuous representation)

Word2Vec

$$\begin{cases} x = (0.54,\ -0.312,\ 3.1,\ ....,\ -2.344,\ 0.543) \in \mathbb{R}^q \\ q \text{ is small, } O(q) \approx \text{hundreds or thousands.} \end{cases}$$

Representations

- The advantage of the first method: easy to interpret. However, it is computationally expensive and slows down programs.

- The second method makes it difficult to interpret the role of each component. However, the vectors can capture semantic meaning.

1. Tokenization
2. Part Of Speech Tagging (**POS tagging**)
3. Chunking
4. Parsing (Syntactic Analysis)

1. Tokenization
2. Part Of Speech Tagging (**POS tagging**)
3. Chunking
4. Parsing (Syntactic Analysis)

Tokenization example

"He called Mr. Green at 2 p.m. in St. Louis, Mr. White did not answer. He then left him a voice mail message."

Sentence tokenization[1]

"He called Mr. Green at 2 p.m. in St. Louis, Mr. White did not answer."
"He then left him a voice mail message."

Word tokenization

{*"He", "then", "left", "him", "a", "voice", "mail", "message", "."*}

[3] See sentence boudaries detection using maximum entropy approach in [Rat98]

POS tagging example (pos tagging is an important task [FH10])

Time	flies	like	an	arrow	.
↓	↓	↓	↓	↓	↓
NN	VB	PRP	DT	NN	.

Fruit	flies	like	a	banana	.
↓	↓	↓	↓	↓	↓
JJ	NN	VB	DT	NN	.

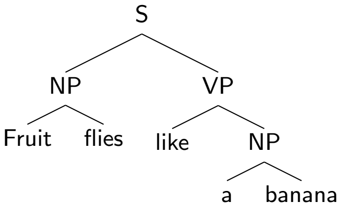
I	saw	a	girl	with	a	telescope	.
↓	↓	↓	↓	↓	↓	↓	↓
PRP	VBD	DT	NN	IN	DT	NN	.

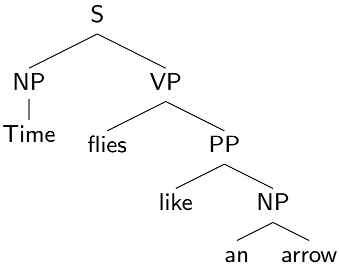
PRP: personal pronoun - **VB**: Verb - **VBD**: Verb in Past tense - **DT**: Determiner - **NN**: Noun singular or mass - **IN**: Preposition - **NNP**: Proper Noun singular - **JJ**: Adjective

Chunking example

"(Fruit flies)$_{\textbf{NP}}$ like (a banana)$_{\textbf{NP}}$." and "(Time)$_{\textbf{NP}}$ flies like (an arrow)$_{\textbf{NP}}$."

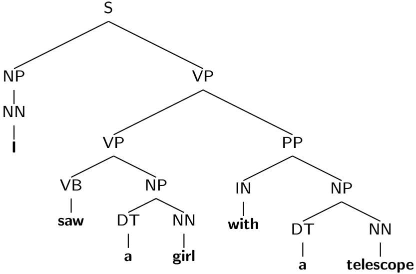
Parsing example n$^\circ$1

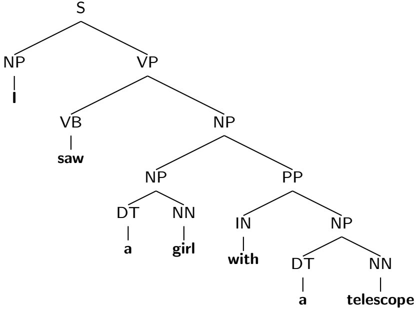




Parsing examples n°2

I	saw	a	girl	with	a	telescope	.
↓	↓	↓	↓	↓	↓	↓	↓
PRP	VBD	DT	NN	IN	DT	NN	.





Mathematical models for POS-Tagging

Some recent methods that mainly have good performances and properties:

- Hidden Markov Models (Brants 2000 [Bra])
- Maximum entropy approaches (Ratnaparkhi 1996 and [R$^+$96])
- Transformation-based learning (Brill 1994 [Bri94])
- An overview of these and other approaches in (Manning and Schütze 1999 [MS99])

- POS Tagging task can be considered as a **classification** problem.

$$\begin{cases} cl: & \mathrm{X} \longrightarrow \boldsymbol{Y} \\ & x \longmapsto y \end{cases}$$

So for a given sentence $w_1, ..., w_N$ (containing a random number of words N), the goal of POS tagging is to find the best tag sequence $t_1^*, ..., t_N^*$.

Mathematical Modeling

Notations:

- n : Corpus length (Dataset size)
- N : A phrase length (which is random variable)
- T : Tagset
- $t_1, ..., t_N$ is a tag-sequence
- $w_1, ..., w_N$ is a word-sequence (a phrase or sentence)
- t^* : The most likely tag for the word w
- p : Theoretical distribution
- $\hat{p}$: Estimation of the distribution p

Mathematical Modeling

Probabilistic Models: For a given N words sequence or a sentence of N words $w_1, ..., w_N$

- Either the conditional probability is maximized:

$$t_1^*, ..., t_N^* = \underset{t_1, \ldots, t_n \in \mathrm{T}}{\arg\max} \left[p\left(t_1, ..., t_N | w_1, ..., w_N\right) \right]$$

  Or alternatively the joint probability is maximized:

$$t_1^*, ..., t_N^* = \underset{t_1, \ldots, t_N \in \mathrm{T}}{\arg\max} \left[p\left(t_1, ..., t_N, w_1, ..., w_N\right) \right]$$

Thank you for your attention!

Bibliographie

- Thorsten Brants, *Tnt: A statistical part-of-speech tagger*.
- Eric Brill, *Some advances in transformation-based part of speech tagging*, Proceedings of the Twelfth National Conference on Artificial Intelligence (Vol. 1) (Menlo Park, CA, USA), AAAI '94, American Association for Artificial Intelligence, 1994, pp. 722–727.
- Anna Feldman and Jirka Hana, *A resource-light approach to morpho-syntactic tagging*, Brill, 2010.
- Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini, *Building a large annotated corpus of english: The penn treebank*, Comput. Linguist. **19** (1993), no. 2, 313–330.
- Christopher D. Manning and Hinrich Schütze, *Foundations of statistical natural language processing*, MIT Press, Cambridge, MA, USA, 1999.
- Adwait Ratnaparkhi et al., *A maximum entropy model for part-of-speech tagging*, Proceedings of the conference on empirical methods in natural language processing, vol. 1, Philadelphia, PA, 1996, pp. 133–142.
- Adwait Ratnaparkhi, *Maximum entropy models for natural language ambiguity resolution*, Ph.D. thesis, Philadelphia, PA, USA, 1998, AAI9840230.

Now let's install Python! ☺