Speech and Natural Language Processing

Didier Schwab

MOSIG 2022-2023

October 4, 2022

Sommaire

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

What you might get out of it

1. An introduction to the field of NLP, its mains challenges, approaches and evaluation methods.
2. A deeper understanding of how some expert systems and machine learning techniques (deep learning here) can be applied to NLP problems.
3. An initial ability to build systems for some of the major problems in Speech and Natural Language Processing: language modelling, text classification, speech recognition, Word Sense Disambiguation...

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

Course organisation (Speech and NLP part)

Lectures' topic

- An introduction to Speech and NLP
- Language modelling
- Word Sense Disambiguation
- Natural language understanding
- Speech Recognition
- Machine Translation

Transversal

- Data Bottleneck
- Evaluation
- Ethics problems
- From classical methods to neural methods

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

References

Dan Juravsky and James H. Martin. *Speech and Language Processing* [Jurafsky et Martin, 2019]

Chris Manning and Hinrich Schutze. *Foundations of statistical natural language processing* [Manning et Schutze, 1999]

Yoav Goldberg. *Neural network methods for natural language processing* [Goldberg, 2017]

and many others...

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

Natural Language Processing (NLP)

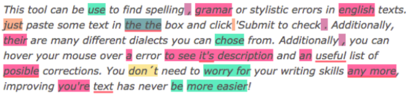
Aim at providing computers with the ability to deal with human natural language (analyse, transform or generate).

Also know as Computational Linguistics (CL) is the English for *Traitement Automatique du Langage Naturel*.

A sub-field of Artificial Intelligence at the crossroad between Computer science (informatics) and Linguistics.

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

Most popular applications of NLP


This tool can be use to find spelling , gramar or stylistic errors in english texts. just paste some text in the the box and click 'Submit to check . Additionally, their are many different dialects you can chose from. Additionally , you can hover your mouse over a error to see it's description and an useful list of posible corrections. You don´t need to worry for your writing skills any more, improving you're text has never be more easier!

Grammatical error correction

Search engines

Machine translation

Document classification

Dialogue

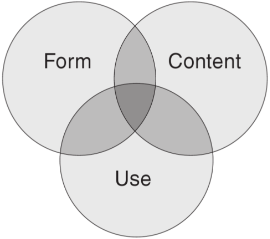
Language

Language development

- Human cognitive and physiological capabilities (innate)
- Communication with others (acquired)

Linguistic skills [Bloom et Lahey, 1978]

- Form: linguistic coding (phonology, lexicon, syntax, morphology)
- Content: semantic, emotion
- Usage: pragmatic

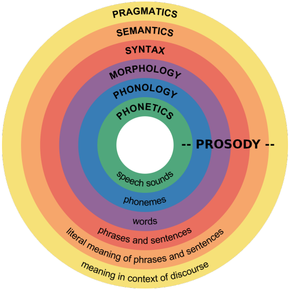


Modern NLP

→ mostly about form and very specific semantics and pragmatics.

Linguistics

- Phonetics:
  - study of sounds produced by humans
- Phonology:
  - Phonemes of a language: Minimum units of sound allowing to differentiate 2 words in a language and their rule of organization
- Morphology:
  - Mental dictionary of words and their formation
- Syntax:
  - combination of lexemes to form a statement
- Semantics:
  - meaning of lemmas and statements
- Pragmatic:
  - use of statements in their context of interaction
- Prosody:
  - rhythm, linguistic or non-linguistic intonations (e.g., irony, emotion)



LABORATOIRE D'INFORMATIQUE DE GRENOBLE

Why NLP is so hard? – *Ambiguity*

- Even well formed sentences are ambiguous or non-interpretable:
  Time flies like an arrow.
  Colourless green ideas sleep furiously. [Chomsky, 1957]

  → Meaning cannot be learned/extracted only from isolated surface form text (language is not self-explanatory).

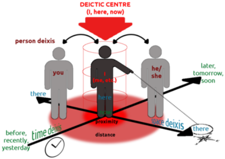
- Interpretation needs context and knowledge (co-reference, history, shared knowledge, common sense)
  The cats were in the street facing the trash cans. They were fighting. → what 'They' refers to?
  A Black Thursday crash is coming.

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

Why NLP is so hard? – multimodality/dynamics

Linguistic communication is inherently linked to the physical world, human perception/cognition/needs and culture

- **Language is dynamic** (new words or expressions): "OMG. This standup is hilarious. I'm dying."
- **Language is culture dependent**: "dexamethasone 1 mg tablet sig three 3 tablet po q8h every 8 hours for 1 doses taper to 2 mg tid x 3 doses on 8 13 . . . "
- **Real language is noisy**: "Le réchauffement climatique est dû à la population de vos gros tas de ferraile et à la cultivisation du soja qui rent nos solle fertile et aride" *exemple de texte d'un élève de 3e*
- **Language is multimodal**: ("*put that there*" [Bolt, 1980])



LABORATOIRE D'INFORMATIQUE DE GRENOBLE

Why NLP is so hard? – Conversation (almost) without human

'Clever' bots?

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

Why NLP is so hard? – Other problems

1. Ambiguity
2. Evaluation
3. Resources

$\rightarrow$ we'll come back to this later

Approaches to deal with NLP problems

1. Develop an exhaustive model to deal with language. For instance Meaning–text theory (MTT) [Mel'čuk, 1981] (too rigid, coverage problem)
2. Reduce the model to the application domain (no need to capture all language phenomena)
3. Rely on human intelligence (e.g., dialogue, web page search)

In all cases

- Highly dependent on resource (corpus) and expertise (highly language dependent)
- As any real world application, it makes the most probable choice and is thus not perfect (low confidence)
- Current trend is bottom-up approach (data driven, less language dependant)

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

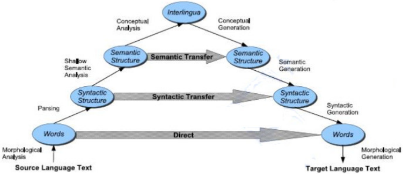
Brief history

- Symbolic NLP (1950s - early 1990s) Expert rule based approaches.
  1954 - The Georgetown experiment
  1960/70 ELIZA (64), SHRDLU (70), Parry (1972)
  80/90 : HPSG, LESK, RST + more structured evaluation methods.
- Statistical NLP (1990s - 2010s) Probabilistic data-driven models.
  2000 - HMM speech recognition – Sphinx [Lee et coll., 1990]
  2007 - Statistical Machine Translation [Brown et coll., 1990] – Moses [Koehn et coll., 2007]
- Neural NLP (present)
  Computing power + big data → rise of DNN.

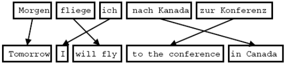
LABORATOIRE D'INFORMATIQUE DE GRENOBLE

Evolution of Paradigms – Example with Translation

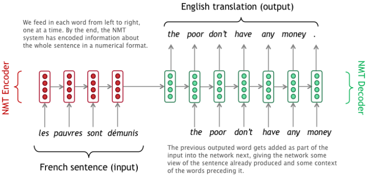
- -90 Symbolic approaches(grammar)
- 06 Statistical approaches (SMT)
- 15 Deep Neural Network (NMT)



English to Spanish:			
1.	NP → Adjective$_1$ Noun$_2$	⇒	NP → Noun$_2$ Adjective$_1$
Chinese to English:			
2.	VP → PP[+Goal] V	⇒	VP → V PP[+Goal]
English to Japanese:			
3.	VP → V NP	⇒	VP → NP V
4.	PP → P NP	⇒	PP → NP P
5.	NP → NP$_1$ Rel. Clause$_2$	⇒	NP → Rel. Clause$_2$ NP$_1$



[from Koehn et al., 2003, NAACL]

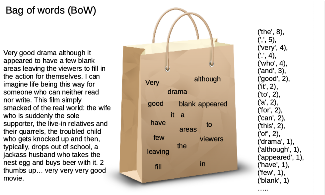


from Abigail See's blog

Text level tasks

The monty python spam...

- Objective: predict categories, extract salient elements (indexing)
- Application: filtering spam emails, classifying documents based on main (latent) content
- Representation: Markov chain (n-gram), bag-of-words



Sentence/Sequence level tasks

- Objective : language modelling - predict next/previous word(s), text generation, text abstraction
- Application: translation, chatbots, sequence tagging, Natural language understanding, named entity recognition
- Representation: character or word sequences (e.g., embeddings)

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

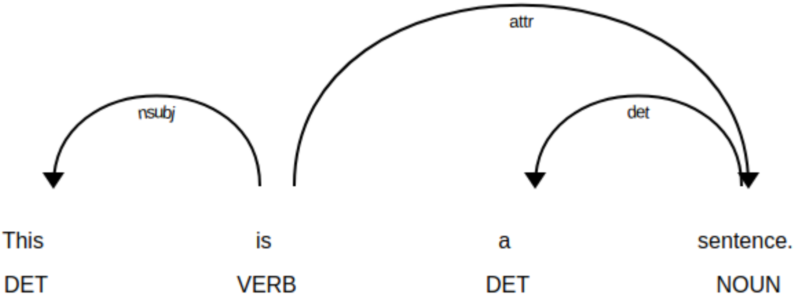
Part of speech tagging

This	is	a	sentence.
DET	VERB	DET	NOUN

$\arg\max_{l_{1..n}} P(l_{1..n}|w_{1..n}))$

Hidden Markov Model, Conditional Random Field, Deep Neural Network, SVM as well as grammatical based method

Dependency parsing



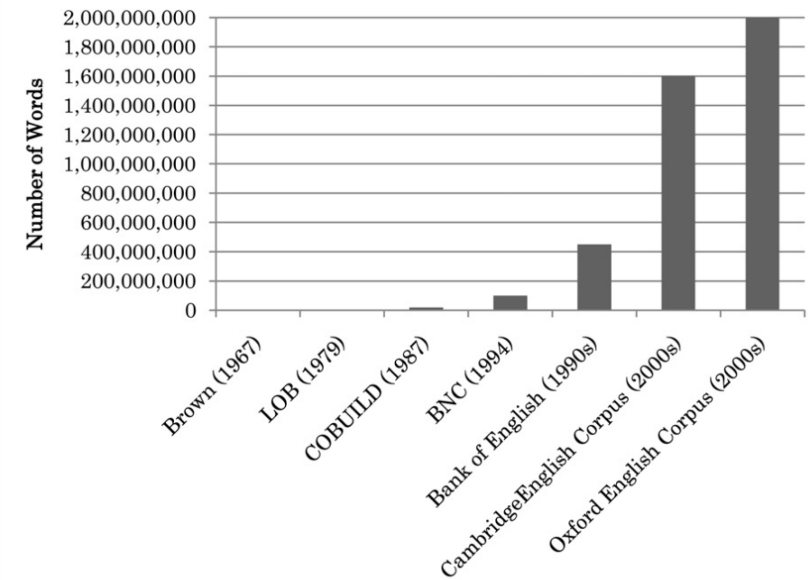
Natural Language Understanding

source: blog.aylien.com 2021

Different kind of resources

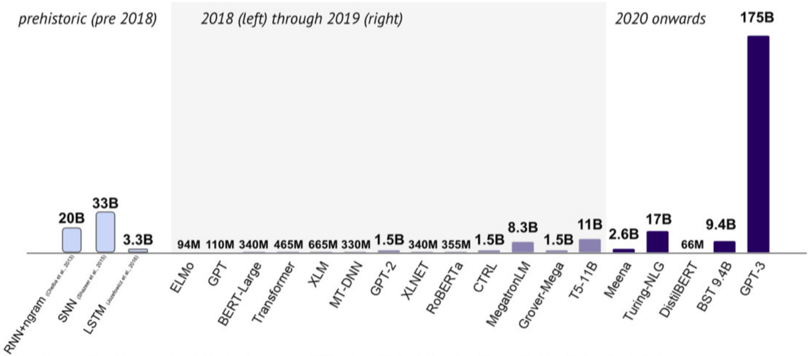
- Corpus
- Lexicon
- Dictionary (monoligual, bilingual)
- Encyclopedia (wikipedia)
- Lexical databases (wordNet, sentiwordnet)

The growing need of resources



LABORATOIRE D'INFORMATIQUE DE GRENOBLE

For greedier machine learning

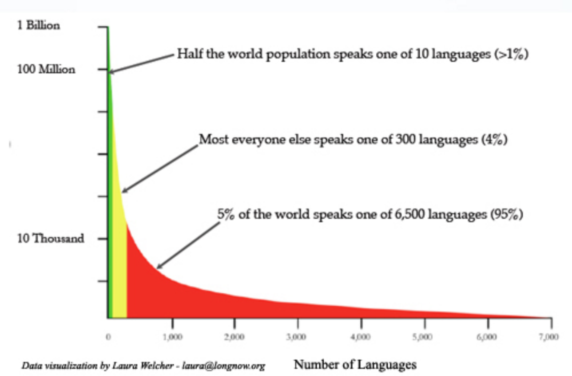


Note: The number of parameters indicates how many different coefficients the algorithm optimizes during the training process.

Oriol Vinyals at stateof.ai 2020

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

Under-resources languages



How to deal with long tail languages when 95% of them are unwritten?

Example: the Haiti earthquake in January 2010.

The common voice initiative

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

Evaluation of NLP

When output is a class

e.g., POS tagging, NLU...
$\rightarrow$ standard classification measures (e.g., accuracy, F-measure)

When output is a generated text

e.g., translation, generation, summarisation...

- Expert based evaluation
  - Correctness, coherence, fluency, etc.
  - Slow and costly
- Automatic evaluation
  - Similarity with reference corpora
  - Very quick and cheap

Black box evaluation does not measure what the system has 'understood' but how it behaves (c.f. Chinese room)
Correlation between human and automatic evaluation debatable.

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

Some common measures

Machine translation: dominated by BLEU (Bilingual Evaluation Understudy) [Papineni et coll., 2002]

- BLEU: comparaison based on n-grams between one candidate translation ws several references

Summarisation: ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [Lin, 2004]
A lot of other measures: TER [Snover et coll., 2006], NIST [Doddington, 2002], LEPOR [Han et coll., 2012], CIDEr [Vedantam et coll., 2015], METEOR [Lavie et Agarwal, 2007], BLEURT [Sellam et coll., 2020]

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

BLEU: example

Given p_n precision as $n - gram$.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n - gram)}{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count(n - gram)} \quad (1)$$

If n small, estimates *adequacy*
If n great, estimates *fluency*

La chatte est sur le tapis

cand 1	The pussy is onto the table.		$count$ et $coun_{clip}$	p_1	p_2
	The.2 pussy.1 is.1 onto.1	table.1	6	$4/6 = .67$	$2/5 = .4$
	The.2 pussy.1 is.1 onto.0	table.0	4		
cand 1	The cat is on the beautiful carpet.				
	The.2 cat.1 is.1 on.1	beautiful.1 carpet.1	7	$5/7 = .71$	$2/6 = .33$
	The.2 cat.0 is.1 on.1	beautiful.0 carpet.1	5		
ref 1	The pussy is on the mat.				
ref 2	The pussy is on the carpet.				

BLEU: length penalty

Problem:
candidate 1 : The pussy
ref 1 : The pussy is on the mat.
ref 2 : The pussy is on the carpet.
$p_1 = 1$ et $p_2 = 1$
$\rightarrow$ add a length penalty.

Given p_n, BP *Brevity Penalty*, c length of candidates and r length of reference translations.

$$p_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count(n-gram)} \; , \; BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{2}$$

BLEU $\in [0, 1]$ is computed as the weighted sum according to the $n - gram$ level (often $N = 4$)

$$BLEU = BP \times exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \Rightarrow \log BLEU = min(1 - r/c, 0) + \sum_{n=1}^{N} w_n \log p_n \tag{3}$$

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

NLP and promises

Beware of media delusions + outbidding

“IBM Watson analyze millions of clinical and scientific reports to help doctors specify cancer treatment based on patients’ genomic profiles”
IBM Watson TV Commercial, ’Watson at Work: Healthcare’ 2017
→ “IBM’s Watson supercomputer recommended ‘unsafe and incorrect’ cancer treatments, internal documents show”
`https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/` 2018

“Robots Can Now Read Better Than Humans, Putting Millions of Jobs at Risk”, Newsweek 2018

Tay: “A peine lancée, une intelligence artificielle de Microsoft dérape sur Twitter”, Le Monde 2016

Sunspring movie: `http://www.thereforefilms.com/sunspring.html`

Limits

Most NLP systems are efficient (sometimes) on the task they have been trained for.

- they do not 'understand' language they just find correlations in the corpus or use expertise (cf. research in natural language grounding).
- It is very difficult to port them to other tasks (Transfer learning is an active research area).

As any DL systems, they are efficient but :

- greedy (most often)
- opaque
- brittle

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

Greediness

Machine side

- Billions of labelled examples to learn to recognize a dog from an image
- Billions of parameters and kW
- several days of learning.

Humain side

- A child just need a few examples to recognize perfectly a dog with few effort
- Animal learns in contact with the environment with all its perception abilities and goals.

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

Environmental impact

Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

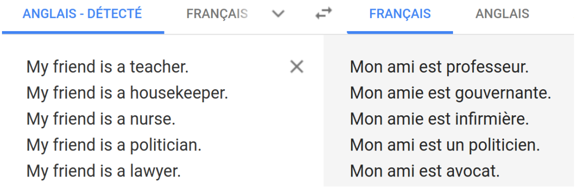
Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

from [Strubell et coll., 2019]

Opacity and black box effect

Biais

Biais inherent to corpora → vicious circle (certificabilité explicabilité équité (CEE – FAT))



translation performed in 2019

Not obvious to insert a priori knowledge to Deep Model.

Interpretability

How to interpret billion of parameters?

Last stage decision are made from latent representations

Challenges

- Bias handling
- Explicability/trust
- A priori knowledge
- Transfer across tasks and languages
- Natural language grounding

References I

Bloom, L. et Lahey, M. (1978).
Language development and language disorders.
John Wiley & Sons Inc.

Bolt, R. A. (1980).
"put-that-there": Voice and gesture at the graphics interface.
SIGGRAPH Comput. Graph., 14(3):262–270.

Brown, P. F., Cocke, J., Della Pietra, S. A., Della Pietra, V. J., Jelinek, F., Lafferty, J., Mercer, R. L. et Roossin, P. S. (1990).
A statistical approach to machine translation.
Computational linguistics, 16(2):79–85.

Chomsky, N. (1957).
Syntactic structures.
Mouton & Co.

Doddington, G. (2002).
Automatic evaluation of machine translation quality using n-gram co-occurrence statistics.
Dans *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145.

Goldberg, Y. (2017).
Neural network methods for natural language processing.
Morgan & Claypool Publishers.

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

References II

Han, A. L.-F., Wong, D. et S Chao, L. (2012).
Lepor: A robust evaluation metric for machine translation with augmented factors.
Dans *Proceedings of COLING 2012*, pages 441–450.

Jurafsky, D. et Martin, J. H. (2019).
Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition.
Pearson International Edition, 3rd édition.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. et coll. (2007).
Moses: Open source toolkit for statistical machine translation.
Dans *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Lavie, A. et Agarwal, A. (2007).
Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments.
Dans *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231.

Lee, K.-F., Hon, H.-W. et Reddy, R. (1990).
An overview of the sphinx speech recognition system.
IEEE Transactions on Acoustics, Speech, and Signal Processing, 38(1):35–45.

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

References III

Lin, C.-Y. (2004).
Rouge: A package for automatic evaluation of summaries.
Dans Marie-Francine Moens, S. S., éditeur : *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81.

Manning, C. et Schutze, H. (1999).
Foundations of statistical natural language processing.
MIT press.

Mel'čuk, I. A. (1981).
Meaning-text models: A recent trend in soviet linguistics.
Annual Review of Anthropology, 10(1):27–62.

Papineni, K., Roukos, S., Ward, T. et Zhu, W.-J. (2002).
Bleu: a method for automatic evaluation of machine translation.
Dans *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

Sellam, T., Das, D. et Parikh, A. P. (2020).
Bleurt: Learning robust metrics for text generation.

Snover, M., Dorr, B., Schwartz, R., Micciulla, L. et Makhoul, J. (2006).
A study of translation edit rate with targeted human annotation.
Dans *Proceedings of Association for Machine Translation in the Americas*.

LABORATOIRE D'INFORMATIQUE DE GRENOBLE

References IV

Strubell, E., Ganesh, A. et McCallum, A. (2019).
Energy and policy considerations for deep learning in nlp.
arXiv preprint arXiv:1906.02243.

Vedantam, R., Zitnick, C. L. et Parikh, D. (2015).
Cider: Consensus-based image description evaluation.
Dans *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 4566–4575.

LABORATOIRE D'INFORMATIQUE DE GRENOBLE