Présentation du traitement automatique des langues

Thierry Poibeau
thierry.poibeau@ens.psl.eu
Laboratoire LATTICE (CNRS & ENS/PSL & Sorbonne nouvelle)

Master Humanités numériques, 2025

Définition

- Traitement automatique des langues : Mise au point de programmes informatiques permettant d'analyser les langues dites « naturelles » (par opposition aux langages artificiels, par ex. langages informatiques)

- Analyse de l'écrit et de l'oral (mais ici on ne parlera que de l'analyse de l'écrit)
- Analyse, mais aussi génération

Quelques mots sur l'évolution du domaine

- Part de plus en plus réduite de la linguistique dans le domaine
  - Importance grandissante de l'informatique
  - TAL et Intelligence artificielle
  - TAL et apprentissage automatique (apprentissage profond / deep learning)

- Quoi présenter ?
  - Comment présenter ce domaine ?
  - En quoi ça peut intéresser des linguistes ? Des spécialistes d'humanités numériques ?

Intérêt pour un linguiste ?

- Pourquoi les systèmes actuels fonctionnent relativement correctement, au moins pour certaines tâches ?
- Est-ce que le TAL peut nous enseigner quelque chose sur la langue et sur son fonctionnement ?
  - Quels succès ?
  - Quels échecs ?
- ... et d'ailleurs, comment évalue-t-on un système de TAL ?
  - Qu'est-ce que ça veut dire « fonctionner correctement » ? (... sachant que tous les systèmes automatiques font des erreurs !...)

Intérêt pour les humanités numériques ?

- De plus en plus de données textuelles disponibles, mais comment les traiter ?

  - Qualitative researchers arrive at the médialab bringing rich data and longing to explore them. Their problem is that qualitative data cannot be easily fed into network analysis tools. Quantitative data can have many different forms (from a video recording to the very memory of the researcher), but they are often stored in a textual format (i.e. interviews transcriptions, field notes or archive documents...). The question therefore becomes: how can texts be explored quail-quantitatively? Or, more pragmatically, how can texts be turned into networks?

« Once Upon a Text: an ANT Tale in Text Analysis », 2012
Tommaso Venturini and Daniele Guido (médialab Sciences Po)

Contenu du cours

- 8 cours

- Théoriques et pratiques
- Présentations d’outils/programmes utilisables avec un minimum de connaissances informatiques (python)
- Etat de l’art (trouver les bons interlocuteurs)

Plan du cours

- Cous 1 : Introduction, présentation générale
- Cours 2 : « Niveaux » d'analyse linguistique (morphosyntaxe, syntaxe, sémantique, entités nommées…) / Spacy (Bert)
- Cours 3 : Entity linking ?
- Cours 4, 5 : LLMs, fine tuning… (N. Durandard)
- Cours 6, 7 : Topic modelling (N. Durandard)
- Cours 8 : Aspects éthiques, présentation des travaux de validation du cours, discussion et conclusion

Quelques logiciels potentiellement utiles

- Dictionnaires électroniques, approches à base de règles, automates à nombre fini d'états : Unitex
- Analyse morphosyntaxique : TreeTagger, Stanford parser, UD parser
- Corpus et analyse syntaxique : Spacy, Stanford parser, UD parser
- Outil d'analyse du lexique : SketchEngine
- Modèles de langue : BERT ; GPT, Gemini, Claude, Mistral, DeepSeek…

Quelques problèmes triviaux

(...qui ne sont, finalement, peut-être pas si triviaux tant que cela...)

Problèmes de « bas niveau »

Après Gênes, la famille Benetton dans la tourmente médiatique et financière

Le gouvernement italien accuse la famille, actionnaire à 30% du gestionnaire du viaduc, d'être responsable de la catastrophe, qui a fait au moins 43 morts.

Faire payer les Benetton. Dans tous les sens du terme. C'est la volonté affichée du gouvernement italien qui a désigné la famille d'entrepreneurs vénitiens comme bouc émissaire de l'effondrement, mardi, du viaduc de Gênes.

Face à la tempête médiatique et à la tourmente financière (le titre Atlantia a chuté de plus de 22 % jeudi avant de finir en hausse à + 5,68 % aujourd'hui), la famille Benetton fait bloc mais surtout profil bas. Elle rappelle par communiqué sa volonté de dialoguer avec le gouvernement.

Les Echos, 17/08

Ambiguïté et polysémie

- *La petite brise la glace.*
- *Le boucher sale la tranche.*
- *Flying planes can be dangerous.*

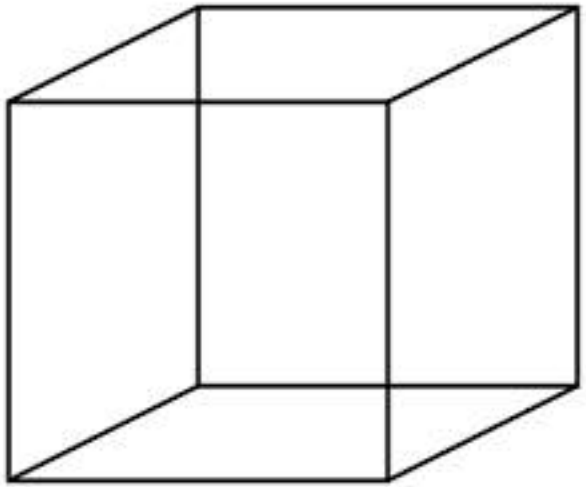
- *Pour pêcher avec son ami, il lui faudrait un autre lieu.*
- *Il a Free, il a tout compris.*
- *Il voit le garçon avec un télescope. (vs avec des cheveux longs)*

« Explosion combinatoire »

- *L'origine du sinistre reste pour l'heure inconnue*. (*Le Figaro*, 19/08)

- l’ : pronom ou nom (voire lettre de l’alphabet)
- origine : commencement / point de départ / milieu de naissance / singularité d’un espace repéré en géométrie …. / verbe « *originer* »
- du : article partitif ou contracté (*de le*)
- sinistre : adjectif / nom / verbe
- reste : nom (nombreux sens) / verbe « *rester* »
- pour : nom ou préposition
- heure : nom (mais « pour l’heure »)

Ambiguïté et perception humaine : le « cube de Necker »

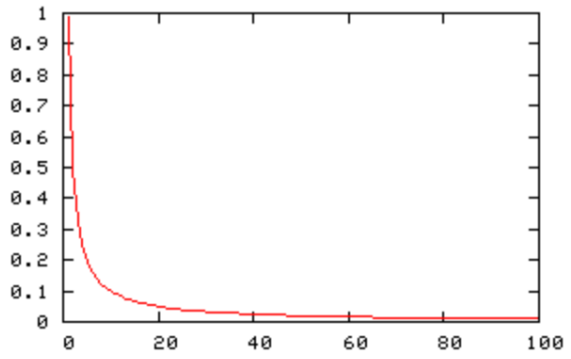


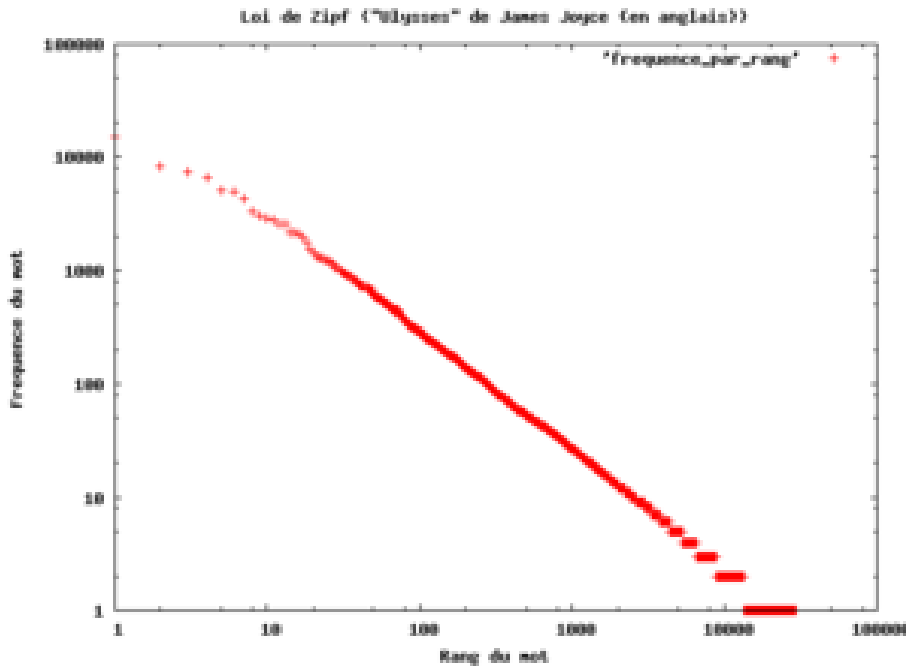
Quelques remarques...

- Les humains ne procèdent (sans doute) pas à une analyse exhaustive des différentes possibilités linguistiques
- Impossibilités de prévoir tous les contextes par une analyse « manuelle »
- Efficacité des ordinateurs (et des techniques informatiques récentes) pour ce type de tâches

Quelques lois statistiques célèbres

- Loi de Zipf (années 1930)
- la fréquence d'utilisation d'un mot dans un texte volumineux est inversement proportionnelle à son rang. (*f(n) = 1/n*)





- Loi de Pareto (80/20)

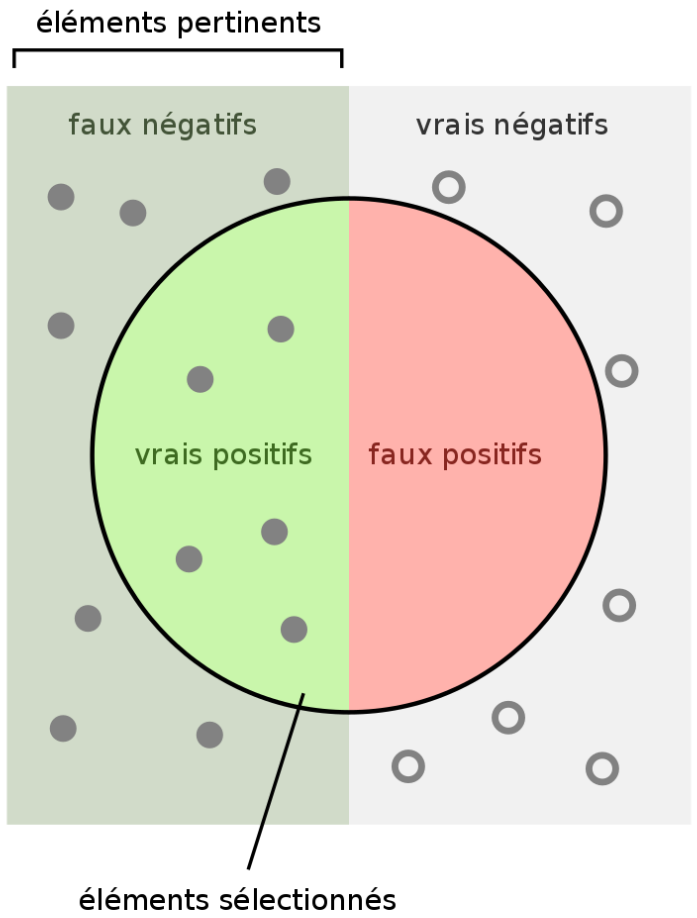
Evaluation : précision et rappel

Evaluation: Précision et Rappel

- Pour un type d'éléments (annotation, document) donné :
  - Est-ce que tous les éléments identifiés sont pertinents ?
  - Est-ce que tous les éléments pertinents ont été identifiés ?
- Mesures de performances :
  - La première question correspond à la précision (bruit)

    P = # éléments pertinents annotés / # éléments pertinents annotés
  - La deuxième au rappel (silence)

    R = # éléments pertinents annotés / # éléments pertinents à annoter
  - F-mesure = 2 * (P * R) / P + R



Combien
de candidats sélectionnés
sont pertinents ?

$$\text{Précision} = \frac{\text{●}}{\text{●●}}$$

Combien
d'éléments pertinents
sont sélectionnés ?

$$\text{Rappel} = \frac{\text{●}}{\text{■●}}$$

Source : Wikipedia, article https://fr.wikipedia.org/wiki/Pr%C3%A9cision_et_rappel