Traitement automatique des langues
Définition(s) & Défis

Intro au TAL — cours n° 1

Guillaume Wisniewski
guillaume.wisniewski@u-paris.fr
septembre 2021

Université de Paris & LLF

Outline

1. NLP & related fields
2. Why do we want to process text with a computer ?
3. Why making computer understand text is difficult ?

À la recherche d'une définition du TAL

Question liminaire

Pourquoi êtes-vous là ?

Pourquoi avoir choisi d'étudier / se spécialiser en TAL ?

Quelque(s) discipline(s) connexe(s)

Informatique = ?

Quelque(s) discipline(s) connexe(s)

Linguistique = ?

Quelque(s) discipline(s) connexe(s)

Intelligence Artificielle = ?

Quelque(s) discipline(s) connexe(s)

Statistiques = ?

Quelque(s) discipline(s) connexe(s)

Apprentissage Statistique = ?

Quelque(s) discipline(s) connexe(s)

Deep Learning = ?

Linguistique

« la langue est une fenêtre sur l'âme »

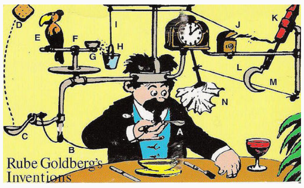
« Language and our thought-grooves are inextricably related, are, in a sense, one and the same. »

Linguistique

« la langue est une fenêtre sur l'âme » (N. Chomsky)

« Language and our thought-grooves are inextricably related, are, in a sense, one and the same. » (E. Sapir)

Informatique



Professor Butts and the Self-Operating Napkin (1931). Soup spoon (A) is raised to mouth, pulling string (B) and thereby jerking ladle (C), which throws cracker (D) past toucan (E). Toucan jumps after cracker and perch (F) tilts, upsetting seeds (G) into pail (H). Extra weight in pail pulls cord (I), which opens and ignites lighter (J), setting off skyrocket (K), which causes sickle (L) to cut string (M), allowing pendulum with attached napkin to swing back and forth, thereby wiping chin.

- = Computer Science ≠ ordinateur
- procedural epistemology = study of the structure of knowledge from an imperative point of view (≠ declarative point of view)
- formalisation et étude de procédés *(process)* → comment réaliser une tâche ?
- maths = "what is", info = "how to"

cf. *Wizard Book* = Structure and Interpretation of Computer Programs,

https://github.com/sarabander/sicp-pdf

Intelligence Artificielle

- un slogan publicitaire
- conférence de Dartmouth (1956)
  - création du terme IA (John McCarthy)
  - objectif : recréer copier ? , dans des machines, les mécanismes du cerveaux humains
- avec beaucoup de hauts et de bas

Les objectifs de la conférence de Dartmouth

« We propose that a 2-month, 10-man study of artificial intelligence be carried out during the summer of 1956 at Dartmouth College in Hanover, New Hampshire. The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer. »

Demande de financement originale de la conférence de Dartmouth

Intelligence Artificielle : est-ce que ça marche ?

Intelligence Artificielle : est-ce que ça marche ?

- oui !
  - Google Translate
  - Calcul d'itinéraires
  - Filtrage de spam
  - ...
- Mais :
  - fondé essentiellement sur les masses de données
    - avis personnel on sait répéter des choses que l'on a déjà vues
  - qu'est-ce que l'intelligence ? intelligence faible / forte
  - intelligence artificielle ou augmentée ?

Exemple : la traduction automatique

Savez-vous traduire les phrases suivantes en anglais ? en swahili ? et un ordinateur ?

1. `Bonjour comment allez-vous ?`
2. `Cette chanson a été diffusée sur les 6 continents`
3. `Et je me retrouvai seul, roulant sous la pluie du jour agonisant, et les essuie-glace étaient en pleine action, mais que pouvaient-ils contre mes larmes ?`

Exemple : la traduction automatique

Savez-vous traduire les phrases suivantes en anglais ? en swahili ? et un ordinateur ?

1. `Bonjour comment allez-vous ?`
2. `Cette chanson a été diffusée sur les 6 continents`
3. `Et je me retrouvai seul, roulant sous la pluie du jour agonisant, et les essuie-glace étaient en pleine action, mais que pouvaient-ils contre mes larmes ?` (Nabokov)

↪ La plupart de nos énoncés sont triviaux (langue « outil ») et faciles à traduire par un ordinateur.

↪ Seuls certains de nos énoncés sont originaux / utilisent toute la beauté de la langue (langue de culture)

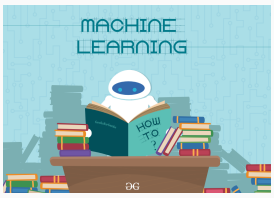
cf. Heinz Wismann, Penser entre les langues

Statistiques

Deux grands objectifs :

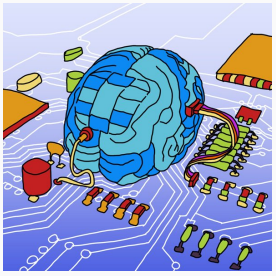
1. décrire/résumer un ensemble de données
2. inférence : est-ce qu'une observation sur un ensemble fini se généralise
   - ↪ tests médicaux
   - ↪ sondages
   - ↪ ...

Apprentissage statistique / automatique



- = *machine learning*
- « méthode » de conception de programme
- plutôt que de décrire comment résoudre un problème (p.ex. traduire une phrase), fournir des exemples et laisser l'ordinateur inférer le programme

Deep Learning

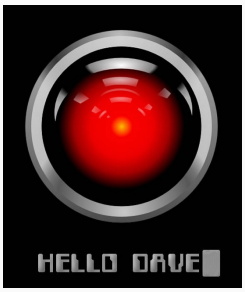


- deep learning *rebranding* de réseaux de neurones
- une des plus anciennes méthodes d'apprentissage statistique

Et le TAL ?

À l'intersection de toutes ces disciplines

Traitement Automatique des Langues



`Dave, stop ... Stop, will you? I'm afraid, Dave ... Dave ... my mind is going ... I can feel it`

Notre objectif :

Faire parler les ordinateurs

↪ interagir avec des données textuelles

↪ interagir avec un ordinateur comme avec un humain (en parlant)

↪ linguistique outillée / linguistique de corpus

Interagir avec des données textuelles

Tarte à la crème

- révolution numérique → essor techniques numériques (informatique + réseau)
- pleins de type de données différentes : sons, images, textes, données de capteurs, ...
- énormément d'échanges sous forme de texte & parole

Quelques ordres de grandeur... (1)

- 40 millions de documents
- chaque année (dépôt légal) :
  - ↪ +70 000 livres
  - ↪ +250 000 périodiques

⇒ « horizon ultime du savoir pour les érudits » (J.-G. Ganascia)

Quelques ordres de grandeur... (2)

- 300 millions de Tweets par jour
- 293 milliards de mails en 2019 (hors spam)

↪ pas toujours la même qualité / intérêt qu'un livre de la BNF...

↪ ...mais il y a des informations intéressantes

De nouveaux défis

Pour nos amis informaticiens/électroniciens

- stocker, garantir l'accessibilité, ...
- 1 image sur un *mur* Facebook = ampoule basse consommation de 20 watts pendant 4h
- 1 année de vie virtuelle = 2 900 litres d'eau = consommation sur 2,5 ans, 365 kWh d'électricité = consommation annuelle de 10 Haïtiens

De nouveaux défis

Pour nous...

- extraire automatiquement des informations et des connaissances de cette masse de données

↪ l'ordinateur doit comprendre des textes

Exemple : le grand débat

- « mouvement des gillets jaunes » → plateformes contributives
  - `https://granddebat.fr/`
  - `https://le-vrai-debat.fr/`

⇒ toutes les données sont librement téléchargeables

	Contributions	Phrases	Mots
GDN	329K	1 161K	29M
VD	98K	256K	6M

⇒ personne ne va « s'amuser » à tout lire !

Exemple de traitements / résultats

Identification des mots les plus « caractéristiques » *(textométrie)*

d'après D. Nouvel `http://damien.nouvels.net/fr/debats2019`

grand débat national pollution (321), aide (317), local (316), polluant (312), taxer (304), agriculture (297), arrêter (296), publique (294), solaire (284), éolien (283), administration (281), plastique (256), écologie (253), production (250), diminuer (246), favoriser (243), développement (243), fiscalité (239), taxe (233), fonctionnaire (232), public (226), énergétique (221), batterie (220), pollueur (211), emballage (206), entreprise (205), diesel (202), recyclage (199), territorial (196), renouvelable (195)

vrai débat vouloir (299), sort (297), candidat (276), voir (275), suivre (264), constitution (254), enfant (249), lire (226), assemblée (225), monnaie (225), loi (222), élection (200), banque (194), aller (193), référendum (186), idée (182), pouvoir (177), article (177), constituant (166), contre (156), libre (156), prof (155), élève (145), souveraineté (141), tirer (141), religion (138), homosexuel (136), droit (136), bonjour (134), parti (133)

Exemple de traitements / résultats

Identification des mots les plus « caractéristiques » *(textométrie)*

d'après D. Nouvel `http://damien.nouvels.net/fr/debats2019`

grand débat national pollution (321), aide (317), local (316), polluant (312), taxer (304), agriculture (297), arrêter (296), publique (294), solaire (284), éolien (283), administration (281), plastique (256), écologie (253), production (250), diminuer (246), favoriser (243), développement (243), fiscalité (239), taxe (233), fonctionnaire (232), public (226), énergétique (221), batterie (220), pollueur (211), emballage (206), entreprise (205), diesel (202), recyclage (199), territorial (196), renouvelable (195)

⇒ écologie

vrai débat vouloir (299), sort (297), candidat (276), voir (275), suivre (264), constitution (254), enfant (249), lire (226), assemblée (225), monnaie (225), loi (222), élection (200), banque (194), aller (193), référendum (186), idée (182), pouvoir (177), article (177), constituant (166), contre (156), libre (156), prof (155), élève (145), souveraineté (141), tirer (141), religion (138), homosexuel (136), droit (136), bonjour (134), parti (133)

⇒ beaucoup plus de variété : éducation, thèmes sociétaux, démocratie, ...

Quelques exemples de tâches / questions

- identification des thèmes
- identification de la polarité
- argument mining

Dans le même ordre d'idées

Hier soir je suis allé voir le dernier film de Tarantino.

...

Un scenario de rêve, des acteurs fantastiques, une intrigue couper le souffle, ...

- identifier que l'on donne son avis sur un film
- identification du film : extraction d'entités nommées, entity linking
- identification de la polarité : opinion mining, sentiment classification, ...

Dans le même ordre d'idées

Hier soir je suis allé voir le dernier film de Tarantino.

...

Un scenario de rêve, des acteurs fantastiques, une intrigue couper le souffle, ... si seulement c'était vrai !

- identifier que l'on donne son avis sur un film
- identification du film : extraction d'entités nommées, entity linking
- identification de la polarité : opinion mining, sentiment classification, ...

Many industrial applications... (1)

Content Filtering

- many website provide (textual content)
  - ↪ newspaper
  - ↪ social networks
  - ↪ ...
- goal : automatically select the content to be displayed
  - ↪ avoid spam
  - ↪ find the most relevant/important 'article'

Many industrial applications... (2)

speech-to-speech translation between several millions of "communication forms"

Machine Translation

- one of the holy grail of Artificial Intelligence
- automatic translation between languages
- help translators (steal their jobs)
- translation for everyone
  - ↪ my grand-mother can go to China without speaking English or Chinese

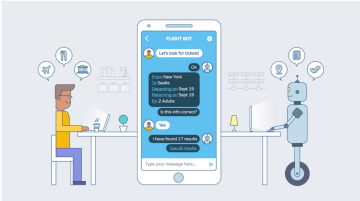
Many industrial applications... (2)

Almost there !

Machine Translation

- one of the holy grail of Artificial Intelligence
- automatic translation between languages
- help translators (steal their jobs)
- translation for everyone
  - ↪ my grand-mother can go to China without speaking English or Chinese

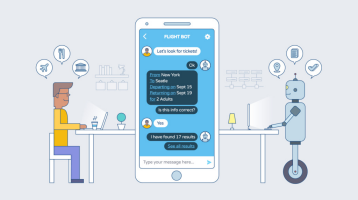
Many industrial applications... (3)



Chatbot

- costumer hot-line
  - ↪ answer to costumer complaints/requests
- route requests, provide information, ask for information, ...

Many industrial applications... (3)



Example

- e.g. automatic booking of flight tickets
  - ↪ find all possible information (destination, seat preferences, dates, ...)
  - ↪ ask the costumer its preferred choice
- automatic or semi-automatic

Linguistique outillée

Documentation linguistique

Définition

- sous-discipline de la linguistique
- description exhaustive des pratiques linguistiques d'une communauté
- deux objectifs
  - ↪ sauvegarder / revitaliser des langues
  - ↪ données empiriques → hypothèses typologiques

Documentation linguistique

Workflow

1. enregistrement
2. transcription (en IPA ou orthographe adaptée)
3. annotation (traductions, gloses, commentaires, ...)
4. archivage & diffusion des données

Est-ce que l'informatique peut-aider ?

↪ acquisition de corpus : LIG-Aikuma

https ://lig-aikuma.imag.fr/

↪ La collection Pangloss : bibliothèque numérique d'enregistrements sonores (≃ 170 langues) ⊕ annotations

https://pangloss.cnrs.fr/

Est-ce que le TAL peut aider ?

- annotation des ressources
  - transcription
  - glose
  - traduction
- fait à la main ⇒ long & coûteux
- objectif : annotation (semi-) automatique
- trouver les similarités / différences entre langues

p.ex. Michaud, Alexis, Oliver Adams, Trevor Cohn, Graham Neubig & Séverine Guillaume. "Integrating automatic transcription into the language documentation workflow : experiments with Na data and the Persephone toolkit". Language

Plus généralement

Est-ce que le TAL peut aider les linguistiques ?

GdR Lift

https://gdr-lift.loria.fr/

1. Extraction de généralisations linguistiques par des méthodes informatiques
2. Linguistique et évaluation des systèmes de traitement automatique des langues
3. Outils de collecte et d'analyse pour les linguistes
4. Données et défis partagés pour une science ouverte
5. Linguistique informatique pour les langues peu dotées ou non documentées

Les grandes défis du TAL

Pourquoi le TAL est-il difficile ?

Principale difficulté

Une jeune femme ou une vielle femme ?

Ambiguïté

1er exemple : dans une base de données...

Le contexte

- entreprise avec plusieurs sites
- chaque employé est rattaché à un site
- objectif : connaitre le nombre d'employés travaillant sur le site de Paris

1er exemple : dans une base de données...

Les bases de données

- système d'information $\simeq$ base de données
- chaque employé est identifié par ensemble de champs ayant chacun une sémantique précise
- contraintes sur les valeurs que peut prendre un champ
- il suffit de i) comparer la valeur du champs à la valeur recherchée ii) compter

n° employé	nom	prénom	site rattachement
2135A	Mickey	Mouse	Paris
1353Y	Donald	Duck	Chartres
1258K	Pat	Hibulaire	Chartres
1469N	Roi	Louie	Paris

1er exemple : dans une base de données...

Est-ce facile ?

Non

- garantir l'exactitude des données (problème sociologique/d'organisation)
- garantir la cohérence des données (problème technique)
- performance (problème technique)

$\Rightarrow$ domaine des bases de données

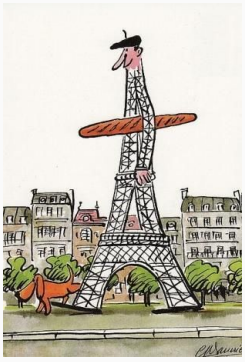
1er exemple : dans une base de données...

Est-ce facile ?

Oui

- chaque site a un identifiant unique
- aucune ambiguité sur le site auquel est rattaché un employé

Dans un corpus de texte...

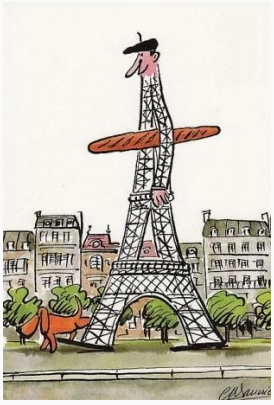


Une nouvelle question :

Trouver tous les « documents » parlant de Paris.

↪ Est-ce toujours aussi simple ?

Différentes manières de nommer Paris ?



Les noms « évidents »

- Paris
- Panam
- la capitale de la France
- la ville lumière
- ...

Les noms « contextuels »

- *Cette ville gigantesque dans laquelle j'ai pass mes plus belles annes*
- la capitale
- le cœur de ce beau pays
-

L'ambiguïté à tous les niveaux (1) : reconnaissance de la parole

Homophones

1re alternative	2e alternative
the tail of a dog	the tale of a dog
the sail of a boat	the sale of a boat

Au niveau des frontières de mots

- It's not easy to wreck a nice beach.
- It's not easy to wreck an ice beach.

L'ambiguïté à tous les niveaux (1) : reconnaissance de la parole

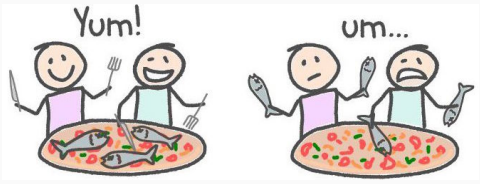
Homophones

- /la/ = la, là, l'a, l'as, las
- belle est jolie *versus* belle et jolie

Au niveau des frontières de mots

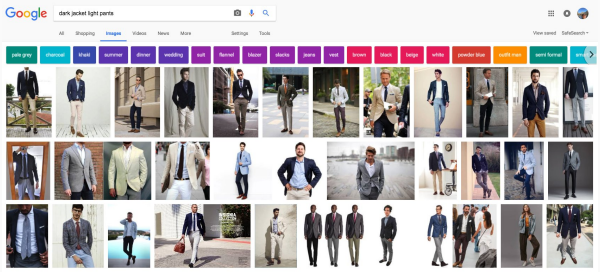
- Les cris du corps enseignant
- l'écris du coran saignant
- les cris du corps en saignant
- l'écris du corps enseignant

L'ambiguité (2) : syntaxe



Je mange une pizza avec des anchoies.

Un vrai problème ?



Trois caractéristiques d'une langue

1. une langue manipule des symboles discrets
2. le sens d'un énoncé se construit en composant le sens des unité
3. une langue est creuse *(sparse)*

Caractère discret des langues

- éléments de base d'une langue écrite = caractère
- combinaison de caractères = mots qui représentent des objets, des concepts, des idées, des événements ou des actions
- aucun lien entre *signifiant* et *signifié*
- la « relation » entre `pizza` et `hamburger` ne peut pas être inférer des symboles

Pourquoi est-ce un problème ?

En vision

Majorelle

(R = 37.65%, V = 31.37% B = 86.27%)

(teinte de 247°, saturation 67% et légèreté 59%)

Cæruleum

(R = 20.78%, V = 47.84% B = 71.76%)

(teinte de 208°, saturation 55% et légèreté 46%)

- possibilité de représenter les couleurs par certaines de leurs propriétés
- des couleurs proches ⇔ propriétés proches
- possibilité de généraliser / passer d'une couleur à l'autre avec de simple opérations mathématiques...
- avec des symboles discrets : besoin d'un énorme dictionnaire / base de connaissances

Sparseness

- version informatique de la « pauvreté du stimulus »
- conséquence des deux caractéristiques de la langue : discret ⊕ combinatoire
- quelque soit le « niveau » auquel on se place : plus que probable que l'on « voit » l'observation pour la première fois

⇒ il ne suffit pas d'acquérir des données
⇒ besoin de « comprendre »