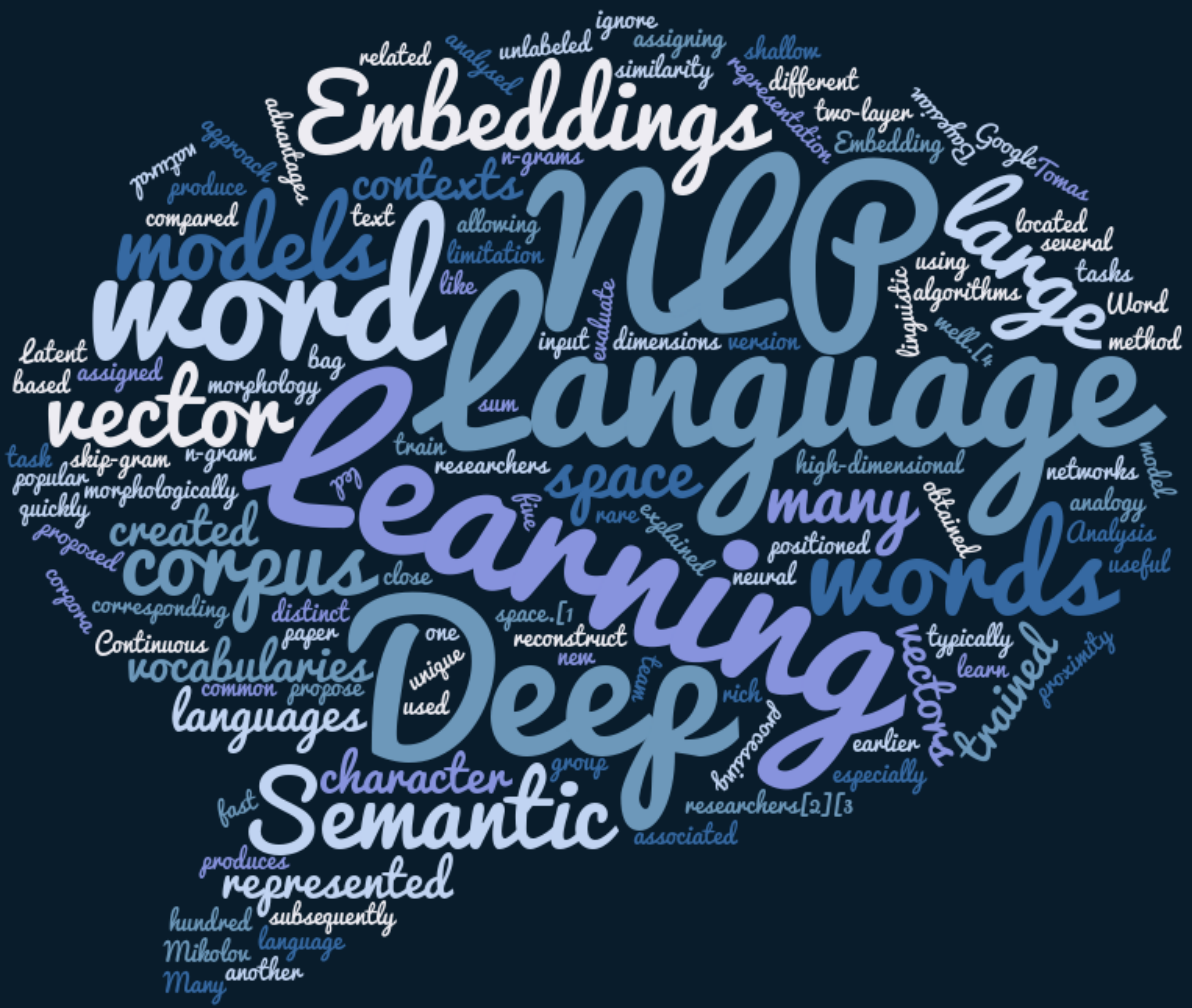

Natural Language Processing

Mourad Akandouch

Notes de cours – INFOY004

Année académique : 2017-2018

ULB - VUB

UNIVERSITÉ LIBRE DE BRUXELLES

VRIJE UNIVERSITEIT BRUSSEL

TABLE DES MATIERES

Chapitre 1

INTRODUCTION

Language shapes the way we think, and determines what we can think about.

- Benjamin Lee Whorf

Qu'est-ce que le traitement du langage naturel (natural language processing[1]) ?

Il faut tout d'abord faire attention à ne pas confondre le « Natural language understanding » avec « Understanding Latural Language », ce cours se concentrant que sur l'un d'eux :

1. *Le Natural Language Understanding (NLU)* est un sous-domaine du NLP, il consiste en la création d'applications mettant en œuvre le langage naturel comme données en entrée dans l'application mais également comme résultat en sortie. On ne calque pas nécessairement le fonctionnement du cerveau mais nous avons plutôt recours à des techniques d'ingénierie.

2. Le Understanding Natural Language qui est plutôt de l'ordre des neurosciences. Pouvons-nous comprendre comment le cerveau humain, en tant que mécanisme biologique, est capable de traiter le langage. Il existe des modèles dans les sciences cognitives.

Dans ce cours, nous nous concentrons principalement sur le NLU et le NLP, pas le second point.

Quelques exemples de NLU

- Il y a un jeu télévisé diffusé depuis 1964 qui s'appelle « Jeopardy! ». Le principe est simple : à partir de réponses communément appelés des indices, trois candidats doivent trouver la question correspondante. Chaque bonne réponse (c'est-à-dire chaque bonne question) rapporte une somme, chaque erreur la fait perdre. Ils peuvent choisir entre six catégories et cinq valeurs d'indices par catégorie.
  En février 2010, IBM Watson bat le meilleur joueur humain au monde dans ce jeu ! Voici la vidéo :
  https://www.youtube.com/watch?v=P18EdAKuC1U.

[1] On va abréger en NLP dans la suite du document

- Les assistants vocaux tels que Siri, Cortana et M, le petit nouveau de Facebook. Il s'agit d'un assistant avec lequel nous pouvons communiquer au travers de l'application Messenger[2].

- Le "speech to speech machine translation". Nous pouvons désormais dicter nos phrases à Google Translate et celui-ci peut instantanément traduire vers une langue cible. La fiabilité des traductions a également été améliorée.

- Il existe encore des tas d'autres exemples.

Pourquoi le NLP est-il si difficile ?

La principale raison rendant le NLP si difficile étant l'ambiguïté pouvant émerger dans nos propos, que ce soit à l'écrit ou à l'oral. Egalement, le fait qu'il y ait tant d'exceptions dans les règles grammaticales sans compter que nous ne savons pas encore très clairement comment l'humain arriver à processer le langage.

Par exemple, la phrase *« I made her duck »*. Cette phrase peut comporter des tas de sens.

- « Je lui ai cuisiné un canard » (*her* ici sert de sujet),
- « J'ai cuisiné son canard » (*her* ici sert de pronom possessif),
- « Je l'ai transformé en canard » (*made* dans le sens *transformer*),
- « Je l'ai faite s'abaisser » (*to duck* peut signifier *s'abaisser*),
- « Je lui ai fabriqué un outil » (*duck* peut signifier un outil),
- « Je lui ai fait boire la tasse » (*duck* peut signifier *boire la tasse*),
- Etc.

Non seulement l'omniprésence de l'ambiguïté est une énorme difficulté en NLP, d'autres choses viennent également se mettre en travers d'un bon processing efficient du langage naturel, par exemple :

1. L'anglais non-standard comme sur Twitter où nous écrivons en abrégé :
   "Great job @justinbieber! were SOO PROUD of what youve done! U taught us 2 #neversaynever & you yourself never should give up either".

2. Les néologismes : *Bromance, Retweet, to google something, unfriend someone.*

3. Les expressions (idioms), par exemple :
   a. *dark horse,* qui se dit lorsque quelqu'un gagne une course de manière innatendue ou lorsque quelqu'un surprend les autres de par son talent ou capacités.
   b. *Get cold feet,* se dit lorsqu'on devient nerveux ou anxieux avant la première tentative de faire telle chose.
   c. *Lose face*, se dit lorsqu'on devient moins respecté.
   d. *Etc.*

4. Noms d'entités piège, par exemple : « *Let it be* was recorded yesterday ».

5. Etc.

En NLP, il y a trois lois fondamentales à prendre en compte et qui ont été dites par Hugo Brandt Corstius :

- *Whatever you do, semantics will screw things up.* Quoi qu'on fasse, la sémantique viendra tout fiche en l'air.

- *Every theory, no matter how explicit it is formulated, will turn out to contain errors when you make a program of it.* Toute théorie, peu importe à quell

[2] https://www.wired.com/2015/08/facebook-launches-m-new-kind-virtual-assistant/

point celle-ci peut être bien formulae, contiendra des erreurs lorsqu'on en fera un programme.

- *The first 80% of accuracy takes litte effort, but further diminishing the gap by half takes double the effort of prevous work. (~law of diminished returns)*
  Les premiers 80% de précision ne demandent que peu d'efforts, mais en diminuant encore l'écart de moitié, on double l'effort du travail précédent.

L'état de l'art actuel du NLP

De nombreux problèmes sont considérés comme quasiment « résolus », d'autres sont en bonne progression et enfin, un certain nombre ont encore énormément de chemin à faire afin de pouvoir être résolus.

À l'heure actuelle, parmi les problèmes quasiment « résolus », nous comptons la détection de spam, la catégorisation de texte *(dire si tel texte appartient à la catégorie Sport, Economie, Technologie, etc.)*, détecter la nature des mots dans une phrase (*sujet, verbe, adverbe, etc.*) et en anglais, la nature des mots se dit « part of speech », souvent abrégé of *POS*. Egalement, parmi les problèmes quasiment résolus, nous trouvons la reconnaissance d'entités nommés[3] qui consiste à extraire d'un corpus linguistique des noms de personnes, monuments, lieux, etc. Enfin, nous avons également presque résolu l'extraction d'information comme le fait de détecter un meeting dans un e-mail.

[3] https://fr.wikipedia.org/wiki/Reconnaissance_d%27entit%C3%A9s_nomm%C3%A9es