


IDL - Cours 2 - idl_03_XML_TAL

Vous avez vu comment :

- structurer des fichiers XML (notion de hiérarchie entre les éléments, attributs, valeurs)
- faire des feuilles de style qui permettent de produire un affichage utile

Rentrons un peu plus dans la partie d'analyse linguistique :


Rappel des étapes, qui constituent un pipeline , l'idée étant qu'on a besoin de l'étape n-1 pour réaliser l'étape n

1.  She's = she is : **segmentation** (ou tokénisation)
2. book = livre, c'est un nom commun (pas le verbe "réserver") : analyse **morphosyntaxique**
3. She's tired = elle est fatiguée, c'est "elle" le sujet de "être fatigué" : analyse **syntaxique**
4. Apple computer : ordinateur Apple : extraction des **entités nommées**.
5. They are best friends = elles sont meilleures amies, car vous savez qu'il s'agit de Jane et Claire // Son ordinateur = his computer. Vous savez que l'ordinateur est à Marc : analyse de **co-références**.
6. mad = énervée, dans ce contexte mad ne se traduit pas par folle ! : analyse **sémantique**

Question 1 : Quelles sont les stratégies mises en place pour développer chacun des modules de ce pipeline ?

Deux grands courants d'approches en NLP :

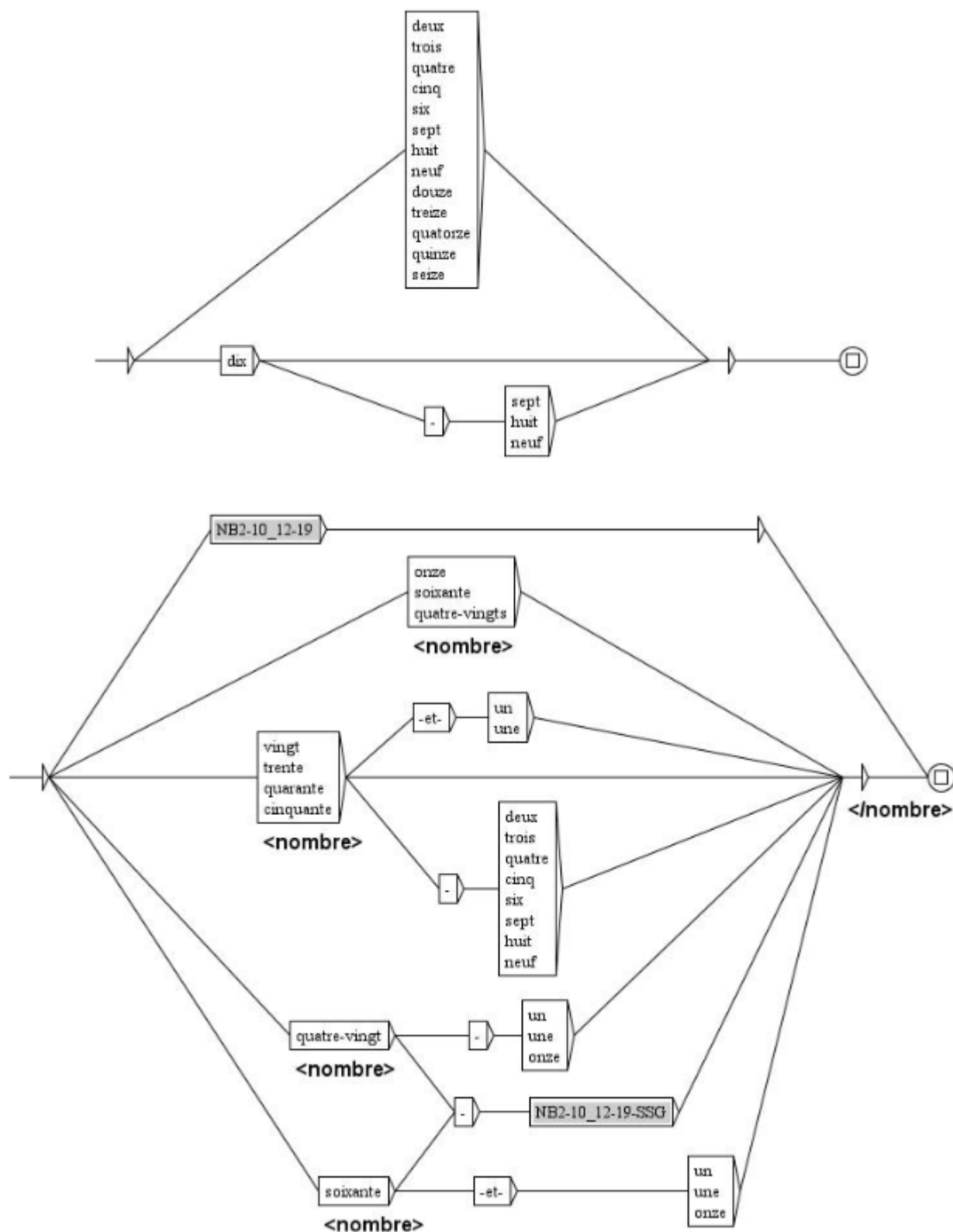
1. Courant n°1 : Approches à base de règles

 idée : si on observe X, alors Y.

Exemple 1 : reconnaissance d'une séquence

Prérequis : Le texte a été séparé en mots au préalable.

Exemple de règles sous forme de graphes pour reconnaître (*et annoter*) les nombres de 2 à 99 :

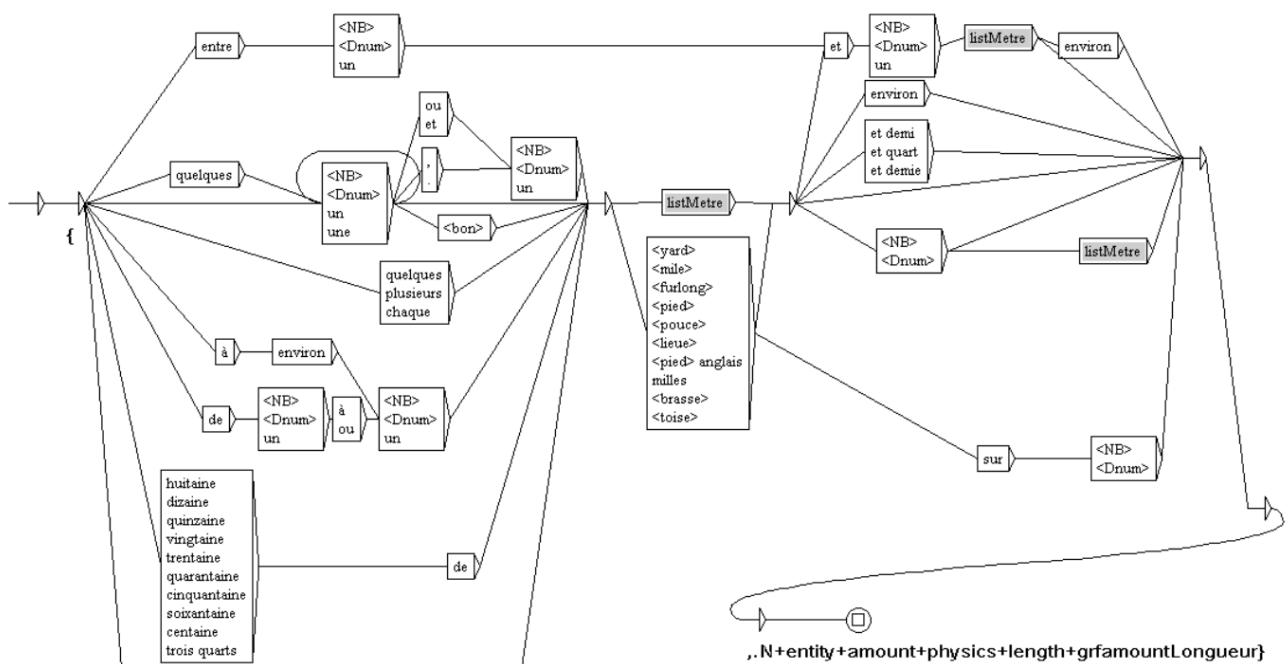


(source : Tutoriel Unitex (https://tln.lifat.univ-tours.fr/medias/fichier/correction-tutoriel-priiseenmain-unitex-annotationcorpus-denis-maurel_1603455974191-zip?ID_FICHE=334600&INLINE=FALSE))

Le programme lit le texte "mot-à-mot" et lorsqu'il reconnaît un motif, les balises `<nombre>` et `</nombre>` sont ajoutées autour des éléments reconnus.

Exemple 2 : reconnaissance de motifs plus complexe : les longueurs

"Vers le sud, une jetée longue de deux mille mètres s'allongeait comme un bras sur la rade de Suez, et à quelques mètres environ de l'extrémité de la jetée se trouvait ma maison"



Remarque : ce graphe reconnaît des structures incorrectes...

- entre 0 et 1 mètre = ok
- entre 1 et 2 mètres = ok
- entre 0 et 0 mètre ???

► Details

Exemple 3 : détecter les phrases incorrectes

"Je mange bleu sur la."

? pourquoi cette phrase est incorrecte ?

et celle-ci : "Je mange souvent une maison sur la conscience." ?

► Details

? À quelle occasion avez-vous déjà fait l'usage d'analyse à base de règles ?

► Details

2. Courant n°2 : Approches à base d'apprentissage

? Quel est l'objectif de l'apprentissage ?

► Details

Apprentissage "supervisé"

L'algorithme s'entraîne sur des données "annotées" (*labeled*), c'est à dire que pour chaque donnée d'entraînement, on renseigne aussi l'output désiré.

- utilisé pour la **classification** : on veut prédire une catégorie (par exemple, la catégorie grammaticale d'un mot)
- la **régression** : on veut prédire une valeur dans un ensemble continu de réels, par exemple le prix d'une maison, le temps passé à regarder une vidéo par un.e utilisateurice de YouTube, etc.

-> données à annoter, bonne interprétabilité

Apprentissage "non supervisé"

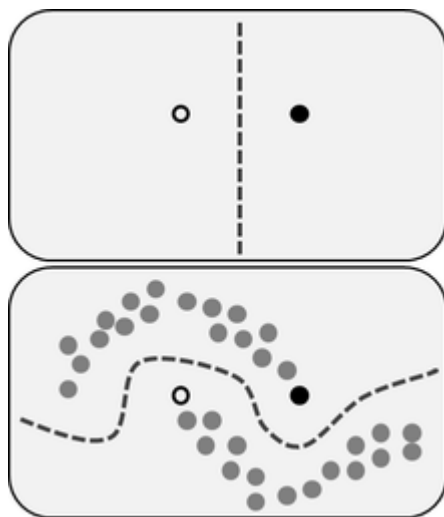
L'algorithme infère des classes de similarité à partir de données "brutes" : "découverte" de structures sous-jacentes permettant de regrouper les données.

-> aucune donnée à annoter, fonctionne sur des tâches plus complexes, plus imprévisibles, interprétabilité plus faible

Exemple : depuis 2018, les (gros) modèles de langue tels que GPT, OpenAI, basés sur des réseaux de neurones, ont supplanté les modèles à base de n-grams qui consistaient à considérer que la probabilité d'apparition d'un mot ne dépendait que des (n-1) mots précédents.

Apprentissage "semi-supervisé"

On utilise à la fois des données annotées (peu) et non annotées (beaucoup)



Source Wikipédia : Apprentissage semi-supervisé (https://fr.wikipedia.org/wiki/Apprentissage_semi-supervis%C3%A9)

-> moins de données à annoter, meilleure précision que le non supervisé complet

Apprentissage auto-supervisé (peu utilisé en NLP)

L'algorithme ajuste son modèle en fonction du résultat des décisions qu'il a prises (exemple : algorithme qui apprend à jouer à un jeu de plateau ou à un jeu vidéo, qui doit optimiser des décisions)

? Quelle est l'approche la plus gourmande en données selon vous ?

? Quelle est l'approche qui requiert le plus d'expertise linguistique selon vous ?

Question 2 : On peut choisir l'approche qui nous semble la plus appropriée pour chaque tâche, mais en fonction de celle-ci on ne va pas s'intéresser aux mêmes "observables" : comment choisir ?

La notion de granularité en NLP

En fonction de la tâche à réaliser et de la méthode employée, on peut s'intéresser à différents niveaux de granularité :

? quels sont-ils ?

► Details

Exemple de tâche : l'identification de langue

The image shows a dark, irregularly shaped object with a rough, textured surface. It appears to be a piece of ancient parchment or a fossilized tablet. The surface is covered with faint, illegible markings that look like ancient script or symbols. There is a small, light-colored, irregular shape near the bottom center, which could be a hole or a piece of damage. The overall appearance is aged and weathered.

Elle a permis à Jean-François Champollion de décrypter les hiéroglyphes. (Voir la pierre

ms=%7B%22x%22%3A0.5474523282432241%2C%22y%22%3A0.6541183090974514%2C%22z%22%3A9.5%2C%22size%22%3A%7B%22width%22%3A0.9253237106921882%2C%22height%22%3A0.6192138849030376%7D%7D))

6 of 9

Salmi salmu 61

1
À u capu di i curisti. Nantu à strumenti à corde. Di Davidiu.
2
O Diu ! ascolta i mio brioni ,
3
Sia attentu à a mio prichera !
4
Da l'estremità di a terra briongu à tè, u core acciaccatu ;
5
Cundùcimi nantu à u scògliu ch'e ùn possu agguantà !
6
Chì sì per mè un aggrondu,
7
Una torra putente, di fronte à u nimicu.
8
Vulariu sughjurnà per sempre in la to tenda,
9
Aggruttammi à l'ascosu di e to ale.
10
- Rifiatu.
11
Chì tù, ò Diu ! stai à sente i mio voti,
12
Mi dai a l'ascita di quelli chì t'èmenu u to nome.
13
Aghjunghji ghjorni à i ghjorni di u rè ;
14
Chì i so anni d'urinu di generazione in generazione !
15
Ch'ellu fermi nantu à u tronu in eternu voltu à Diu !
16
Fà chì a to buntà è a to fideltà vèghjinu nantu ad ellu !
17
Allora cantaraghju u to nome senza cissà,
18
Purtendu ogni ghjornu i mio voti à cumpiimentu.

Psaumes psaume 61

1
Au chef des chantres. Sur instruments à cordes. De David.
2
Ô Dieu ! écoute mes cris,
3
Sois attentif à ma prière !
4
Du bout de la terre je crie à toi, le cœur abattu ;
5
Conduis-moi sur le rocher que je ne puis atteindre !
6
Car tu es pour moi un refuge,
7
Une tour forte, en face de l'ennemi.
8
Je voudrais séjourner éternellement dans ta tente,
9
Me réfugier à l'abri de tes ailes.
10
- Pause.
11
Car toi, ô Dieu ! tu exauces mes vœux,
12
Tu me donnes l'héritage de ceux qui craignent ton nom.
13
Ajoute des jours aux jours du roi ;
14
Que ses années se prolongent à jamais !
15
Qu'il reste sur le trône éternellement devant Dieu!
16
Fais que ta bonté et ta fidélité veillent sur lui !
17
Alors je chanterai sans cesse ton nom,
18
En accomplissant chaque jour mes vœux.

? quelle est la granularité la mieux adaptée à votre avis ?

niveau 'mot'

Idée : utiliser la loi de Zipf !

La loi de Zipf :

"si on classe les mots par ordre de fréquences décroissantes, la fréquence du k -ème mot est approximativement proportionnelle à $1/k$ "

ex. : "the" représente près de **7 % du Brown Corpus** alors que près de la moitié du vocabulaire total du corpus sont des hapax.

Seuls **135 éléments** de vocabulaire sont nécessaires pour couvrir la **moitié du Brown Corpus**.

Rang	Mot	Fréquence
1	<i>the</i>	69 970
2	<i>of</i>	36 410
3	<i>and</i>	28 854
20	<i>I</i>	5 180

SOURCE (https://members.loria.fr/KFort/files/fichiers_cours/IntroTAL.pdf)

► Details

niveau 'caractère' : utilisation des n-grams

► Details

Pour certaines tâches, le "grain caractère" suffit. Mais ce n'est pas le cas de toutes les tâches ! On va parfois avoir besoin de regarder

- les mots ou n-grams de mots (analyse morphosyntaxique),
- les phrases (analyse sémantique),
ex : *"Jamais vous ne m'entendrez dire : "ce produit est génial!"*
- les paragraphes (analyse en co-références)

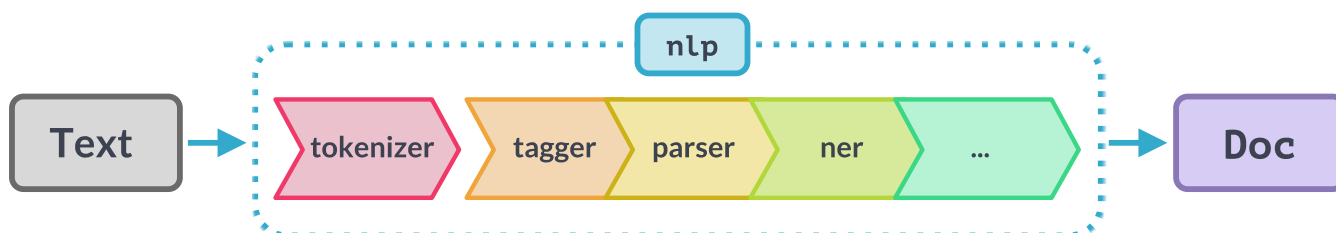
Question 3 : Comment est effectué ce découpage ?

S'atteler à notre tâche numéro 1 : le découpage en phrases - une tâche facile !

? Quelle est l'approche (règles ou apprentissage) la plus adaptée pour notre tâche n°1 : le découpage en phrases ?

? Quid du découpage en "mots" ?

Démo librairie Spacy (python)



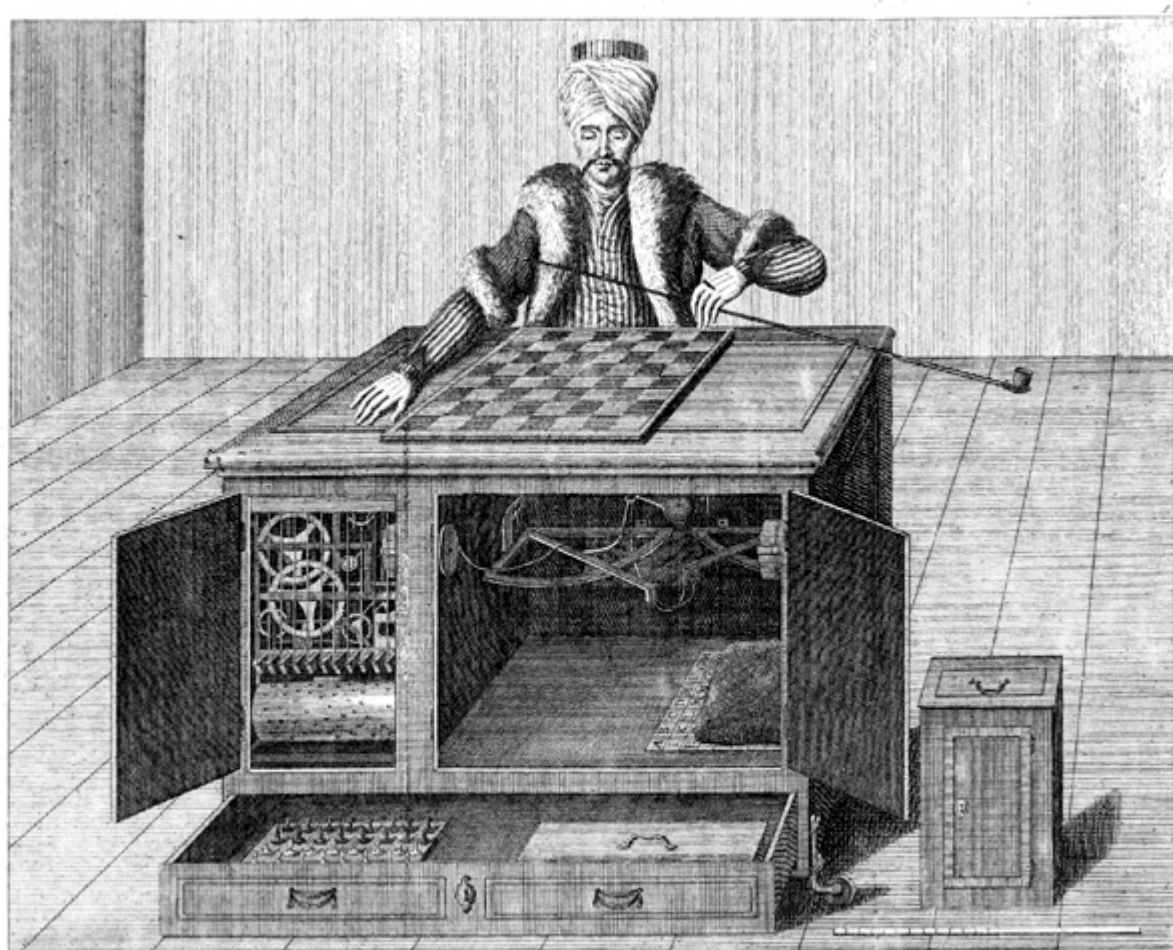
source : <https://spacy.io/api> (<https://spacy.io/api>)

Documentation pipeline nlp (<https://spacy.io/usage/processing-pipelines>)

⚠ Ce n'est pas parce qu'une librairie existe et est joliment présentée qu'elle produit un résultat systématiquement correct.

ex : <https://spacy.io/models/fr> (<https://spacy.io/models/fr>)

Point analyse critique : historique de l'intelligence artificielle



W. de Kempelen del. *Ch. a Mechel, exaud. Basilea.* P.G. Piatz, sc.
Der Schachspieler, wie er vor dem Spiele gezeigt wird, von vorne. Le Joueur d'Echecs, tel qu'on le montre avant le jeu, par devant.

Amazon mechanical turk : plateforme web de production participative (crowdsourcing) qui vise à faire effectuer par des humains, contre (micro-)rémunération, des tâches plus ou moins complexes.

? quand participez vous vous-même à produire des données d'entraînement pour des algorithmes d'"IA" ?

ChatGPT et les travailleurs Kenyans (<https://usbeketrica.com/fr/observations/openai-est-accuse-d-avoir-fait-appel-a-des-travailleurs-kenyans-payes-deux-dollars-de-l-heure-pour-moderer-chatgpt>)

Si il reste du temps : commencez à réfléchir à des expressions régulières pour reconnaître les phrases, puis les mots dans un texte.