Introducion Natural Language Processing

Alison PATOU

Patou.alison@gmail.com

1

Définition

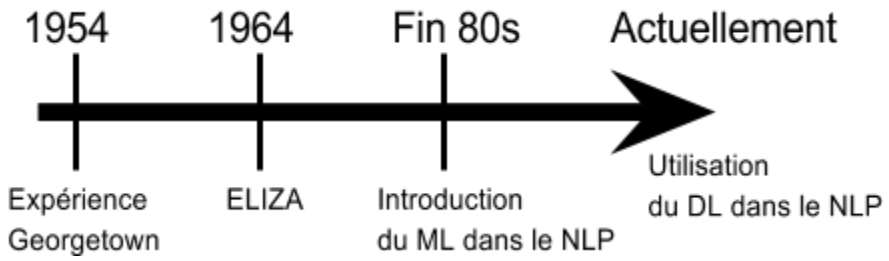
Comment le correcteur orthographique arrive à fonctionner ?

Comment la complétion automatique de phrase fonctionne quand on écrit un mail avec Gmail ?

Comment Siri arrive à retranscrire textuellement ce qu'on lui dit oralement ?

Comment Google arrive à anticiper vos recherches dès la première lettre tapée ?

Pour la petite histoire



l'expérience Georgetown (en) en 1954 qui comportait la traduction complètement automatique de plus de soixante phrases russes en anglais.

Premier chatbot ELIZA

Quelles disciplines ?

- intelligence artificielle
- linguistique
- philosophie

Deux points de vue:

- “Natural language processing” : ingénierie, tâches à résoudre, approche expérimentale par évaluation
- “Computational linguistics” : science, modèles explicatifs, validation par des données

C’est quoi du texte ?

- Une suite de lettres

```
l e  c h a t  e s t ...
```

- Une suite de mots

```
le chat est ...
```

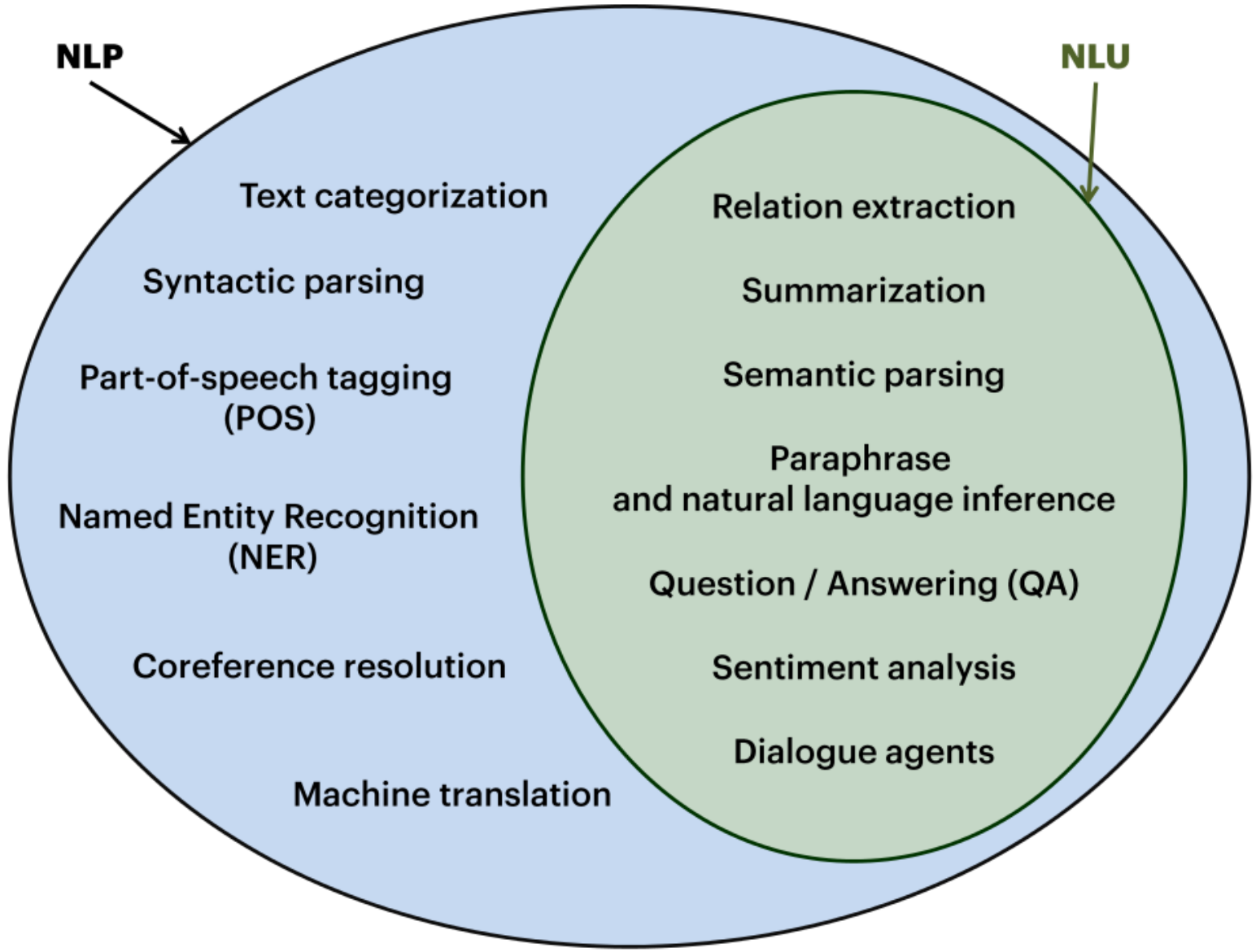
- Un ensemble de mots

Dans l'ordre alphabétique

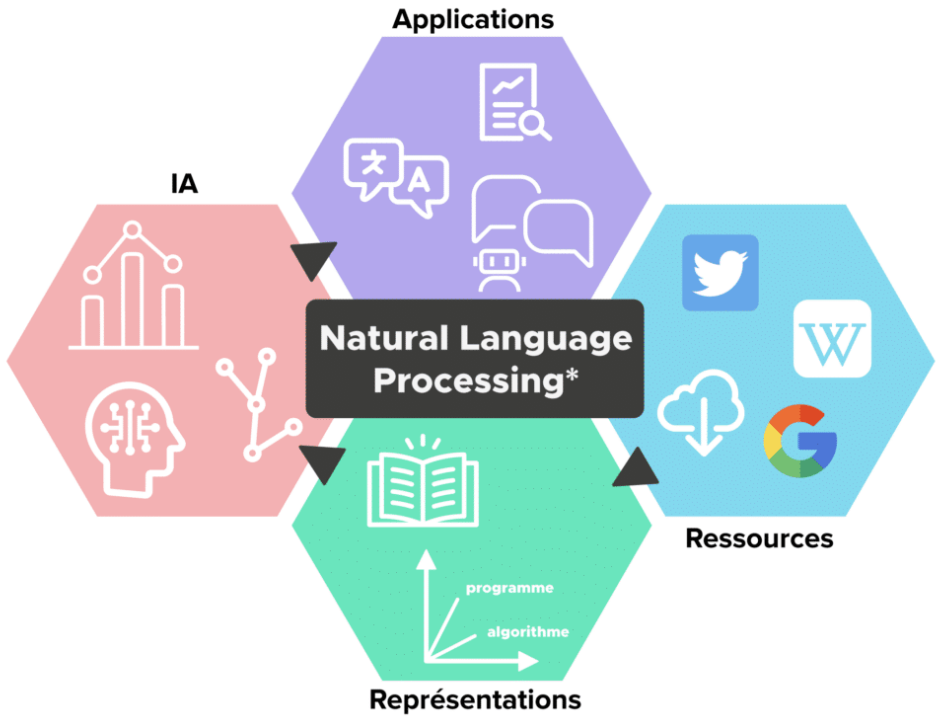
```
chat
est
le
...
```

Définition

Le **NLP** pour **Natural Language Processing** ou **Traitement Numérique du Langage** est une discipline qui porte essentiellement sur la **compréhension**, la **manipulation** et
la **génération** du **langage naturel par les machines**. Ainsi, le *NLP* est réellement à l'interface entre la science informatique et la linguistique. Il porte donc sur la capacité de la machine à interagir directement avec l'humain.


NLP
NLU
Text categorization
Syntactic parsing
Part-of-speech tagging (POS)
Named Entity Recognition (NER)
Coreference resolution
Machine translation
Relation extraction
Summarization
Semantic parsing
Paraphrase and natural language inference
Question / Answering (QA)
Sentiment analysis
Dialogue agents

A quelle problématique répond le NLP ?



*Traitement Automatique du Langage Naturel

DataScientest • com

Classification et catégorisation de texte

- On compte bon nombre d'applications, telles que la recherche Web, le filtrage des informations, l'identification de la langue, l'évaluation de la lisibilité et l'analyse des sentiments.

- L'objectif de la catégorisation de textes est de pouvoir associer automatiquement des documents à des classes (catégories, étiquettes, index) prédéfinies. Nous nous plaçons dans le cadre de l'apprentissage supervisé.

Rappels classification

Observations	Vraie classe	Classe prédite par le modèle
Animal 1	Chien	Chien
Animal 2	Chien	Chat
Animal 3	Chat	Chat
Animal 4	Chien	Chien
Animal 5	Chat	Chien
Animal 6	Chat	Chat
Animal 7	Chat	Chat

Matrice de confusion du classifieur d'animal

Ligne 1 : Deux chiens ont été classifiés comme chiens et un chien a été classifié comme chat.
Ligne 2 : Un chat a été classifié comme chien et trois chats ont été classifié comme chats.

Reconnaissance d’entité nommée

- La tâche principale de la reconnaissance d’entité nommée, est de classer les entités nommées, telles que Pedro Saltimo, Google, Londres, etc., dans des catégories prédéfinies telles que personnes, entreprises, ou villes etc.

Traduction automatique

- Les outils de traduction automatique sont sont très populaires et ce malgré leurs limites. Dans certains secteurs, la qualité de la traduction n’est pas optimal. Pour améliorer les résultats, les spécialistes essayent différentes techniques et modèles, y compris l’approche par réseau de neurones. L’objectif de l’étude de la traduction automatique basée sur les neurones pour le domaine médical des textes est de vérifier les effets de différentes méthodes de formation sur un système de traduction automatique.

Balisage partiel

- Le part of speech tagging (POS), en français l'étiquetage morpho-syntaxique, a de nombreuses applications. Notamment l'analyse syntaxique, la conversion de texte en parole, l'extraction d'informations, etc. Un modèle de CNN a été testé sur les données du Wall Street Journal provenant du jeu de données Penn Treebank III et a atteint une précision de marquage de 97,40%

Reconnaissance vocale

- La reconnaissance vocale a de nombreuses applications, telles que la domotique, la téléphonie mobile, l'assistance virtuelle, l'informatique mains libres, les jeux vidéo, etc. Les réseaux neutres sont largement utilisés dans ce domaine.

Reconnaissance de caractère

- Les systèmes de reconnaissance de caractères ont également de nombreuses applications telles que la reconnaissance de caractère de reçu, la reconnaissance de caractère de facture, la reconnaissance de caractère de contrôle, la reconnaissance de caractère de document de facturation, etc.
- La reconnaissance des caractères par réseau neuronal présente une méthode de reconnaissance des caractères manuscrits avec une précision de 85%.

Quelques applications – Analyse de sentiment

De manière générale, l'**analyse des sentiments** permet de mesurer le niveau de satisfaction des clients vis-à-vis des produits ou services fournis par une entreprise ou un organisme. Elle peut même s'avérer **bien plus efficace que des méthodes classiques** comme les sondages.

Quelques applications – chatbot

- Les **méthodes NLP** sont au cœur du fonctionnement des Chatbots actuels. Bien que ces systèmes ne soient pas totalement parfaits, ils peuvent aujourd'hui **facilement gérer des tâches standards** telles renseigner des clients sur des produits ou services, répondre à leurs questions, etc. Ils sont utilisés par plusieurs canaux, dont l'Internet, les applications et les plateformes de messagerie. L'ouverture de la plateforme **Facebook Messenger** aux chatbots en 2016 a contribué à leur développement.

Vérification orthographique

- La plupart des éditeurs de texte permettent aux utilisateurs de vérifier si leur texte contient des fautes d'orthographe. Les réseaux de neurones sont maintenant intégrés à de nombreux outils de vérification orthographique.

- Dans la vérification orthographique personnalisée à l'aide de réseaux de neurones. Un nouveau système de détection des mots mal orthographiés a été proposé. Ce système est formé sur l'observation des corrections spécifiques apportées par un dactylographe. Il élimine nombre des faiblesses des méthodes de vérification orthographique traditionnelles.

Problèmes

j'croibi1k G1 pb — langue non-standard

tweet; biopic; abracadabrantesque — néologismes
imports d'autres langues

Marie et Jean sont amis. Marie et Jean sont français. — connaissance du monde

Marie et Jean sont de bons amis — contexte

Complexité

- Le langage est complexe à traiter car c'est un système spécialement développé par l'humain, qui regroupe information et sens.

- En effet, les mots seuls, sans contexte, peuvent être mal interprétés d'où l'importance du sens.

- Le langage dépend du contexte social, du monde qui l'entoure, mais aussi, du sens commun et des normes sociales et culturelles.

Sans compter qu'il faut que la machine puisse tenir compte des expressions, des niveaux de langages, etc.

2

Pré-traitement

“ Le traitement automatique du langage naturel vise à créer des outils de traitement de la langue naturelle pour diverses applications. “

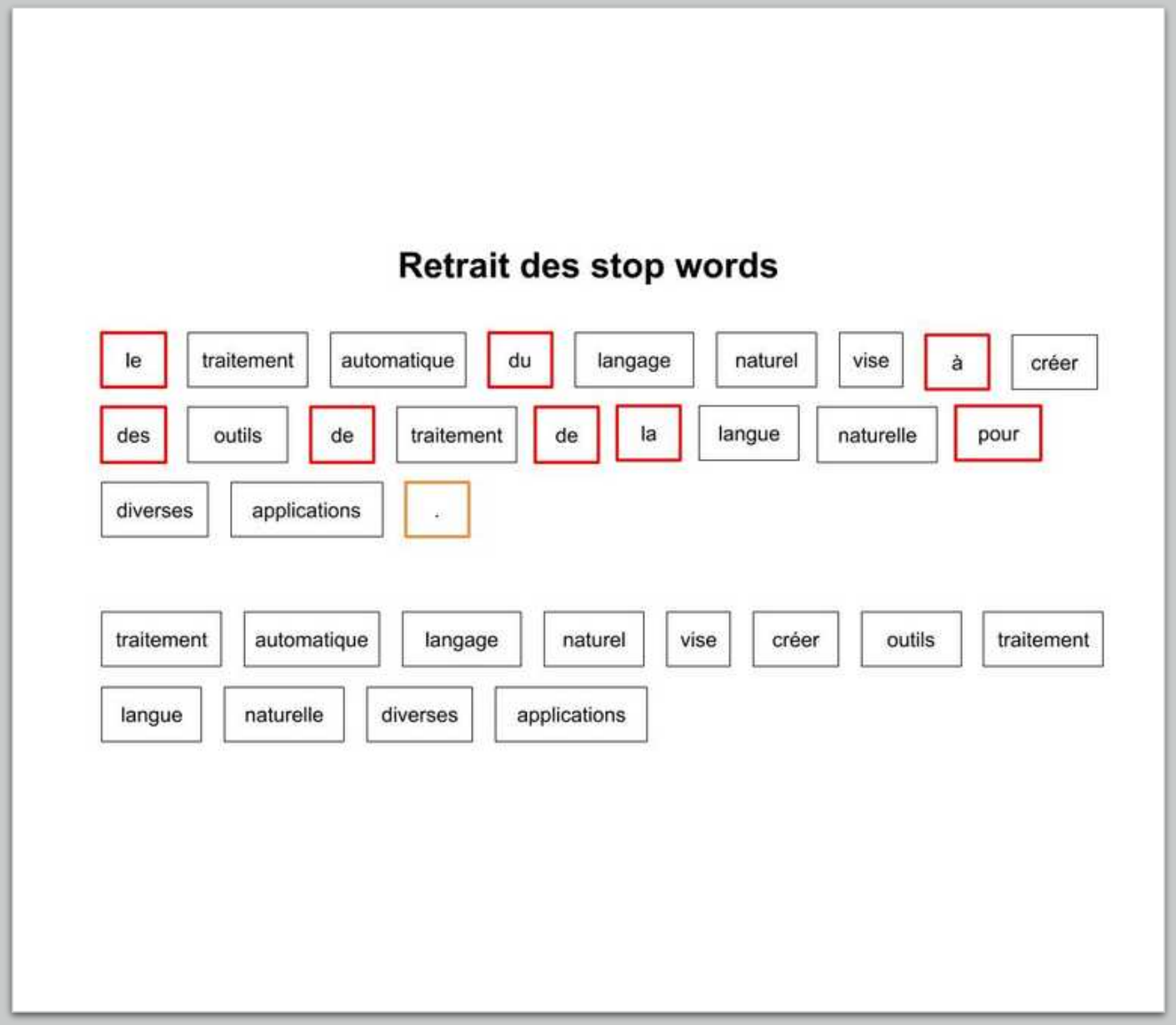
- Nettoyage
- Token (tokenisation)
- Stemmatisation
- Lemmatisation

30 min

Nettoyage

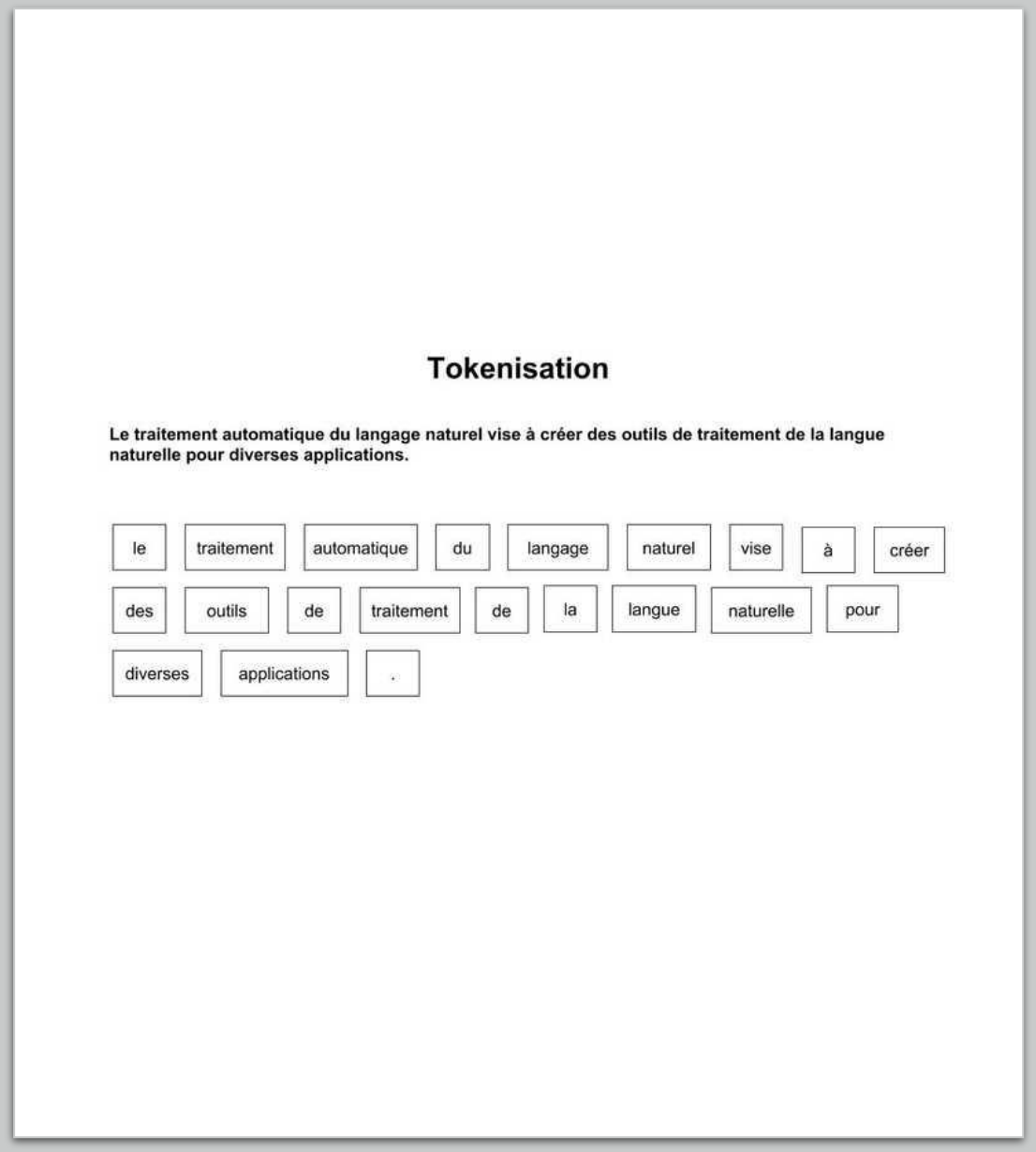
- Comme pour toutes les données que nous avons manipulé précemment, les données textuelles ont parfois (souvent) besoin d'être nettoyées.

- Suppression d'URL
- Suppression d'émoji
- Passer tout en minuscule
- Enlever des caractères spéciaux, des « stopwords », …



Tokenisation

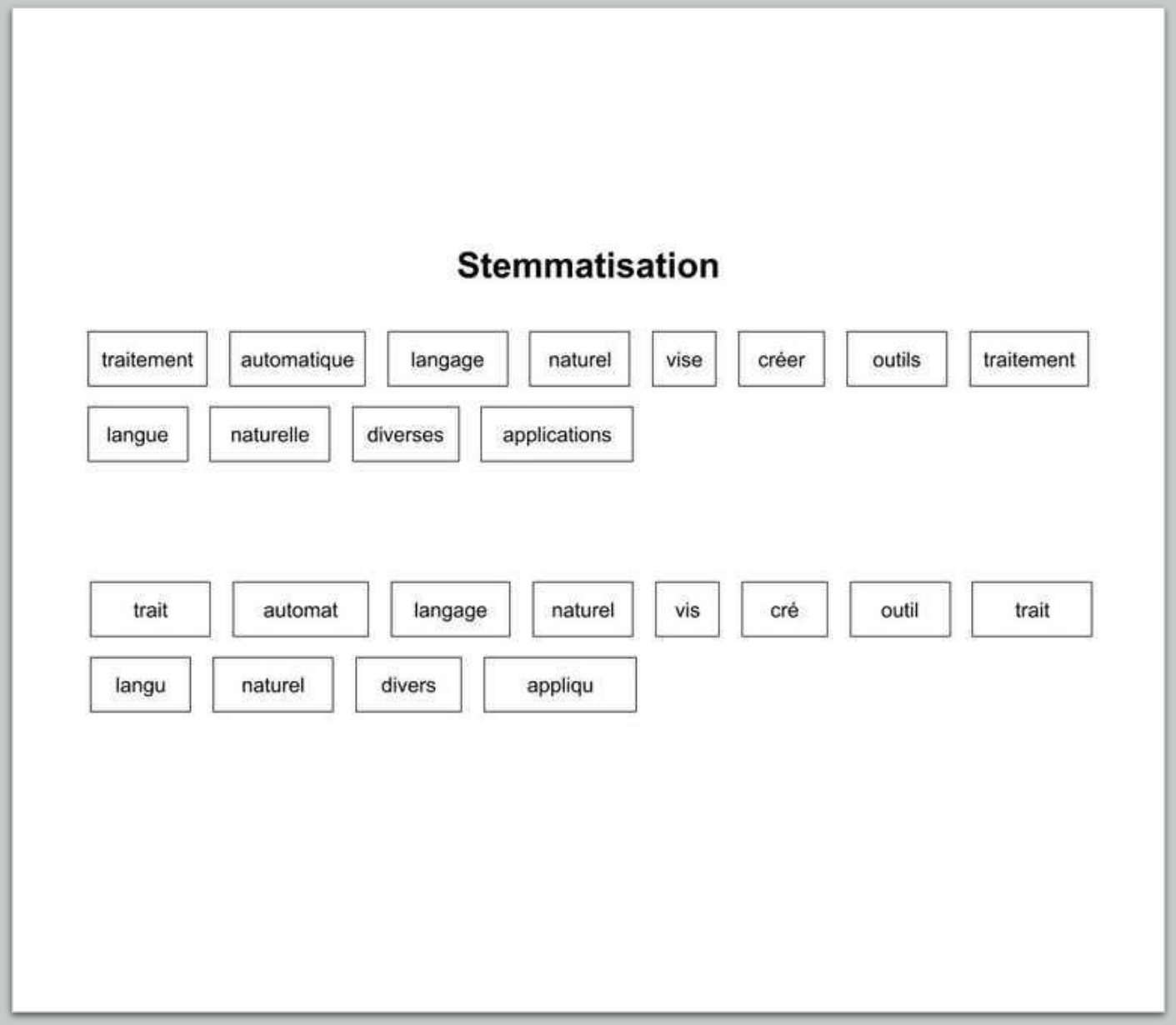
- Il s’agit de décomposer une phrase, et donc un document, en tokens. Un token est un élément correspondant à un mot ou une ponctuation, cependant de nombreux cas ne sont pas triviaux à traiter :
- Les mots avec un trait d’union, exemple : peut être et peut-être qui ont des significations très différentes ;
- Les dates et heures qui peuvent être séparées par des points, des slashs, des deux points ;
- Les apostrophes ;
- Les caractères spéciaux : émoticônes, formules mathématiques.



Stemmatisation

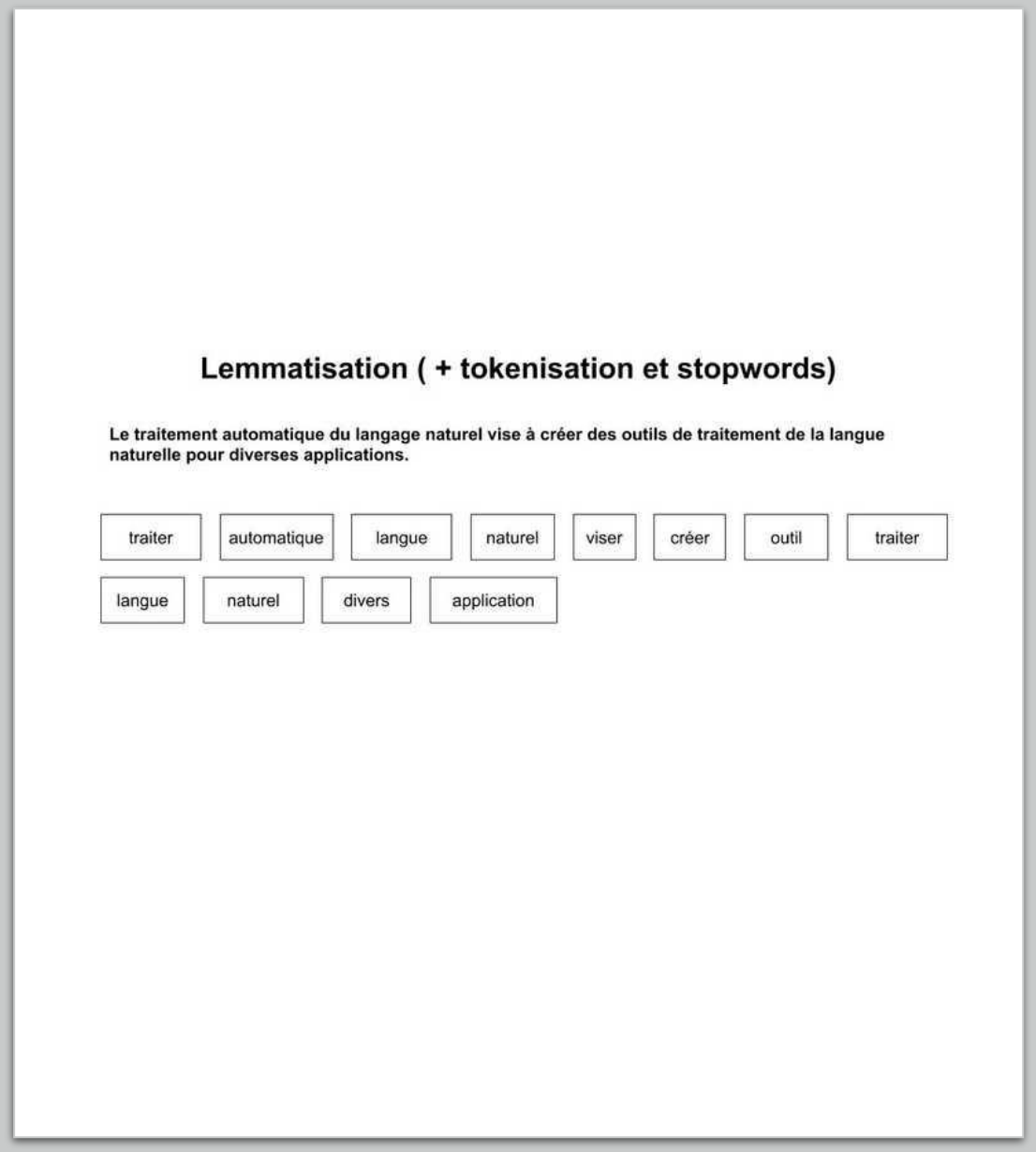
- La stemmatisation (ou racinisation) réduit les mots à leur radical ou racine.

- La méthode est plus « radicale » que la lemmatisation car on n’obtiendra pas un mot connu.



Lemmatisation

- La lemmatisation, qui prend en considération le contexte dans lequel le mot est écrit, a pour but de trouver la forme canonique du mot, le lemme. Par conséquent, elle doit se faire après la transformation des lettres majuscules en minuscules et avant la tokenisation car les mots présents avant et après sont importants pour déterminer la nature du mot.



Bag of Words

- On considère que le monde peut être décrit au moyen d'un dictionnaire (de « mots »). Dans sa version la plus simple, un document particulier est représenté par l'histogramme des occurrences des mots le composant: pour un document donné, chaque mot se voit affecté le nombre de fois qu'il apparaît dans le document

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

3

Analyses

Transformation - TF

Term-Frequency (TF) **:** cette méthode consiste à compter le nombre d'occurrences des *tokens* présents dans le corpus pour chaque texte. Chaque texte est alors représenté par un **vecteur d'occurrences**. On parle généralement de ***Bag-Of-Word***, ou sac de mots en français.

"the black cat eats fish but does not eat other black cats"

Term	Occurrences	Term Frequency
"cat"	2	0.25
"black"	2	0.25
"eat"	2	0.25
"fish"	1	0.125
"other"	1	0.125
Total	**8**	**1**

Machine Learning

- Les approches classiques d'apprentissage automatique peuvent être utilisées pour résoudre des problèmes plus difficiles. Contrairement aux méthodes fondées sur des règles prédéfinies, elles reposent sur des **méthodes qui portent réellement sur la compréhension du langage**. Elles exploitent les données obtenues à partir des textes bruts prétraités via une des méthodes décrites en haut par exemple. Elles peuvent également utiliser des données relatives à la longueur des phrases, à l'occurrence de mots spécifiques, etc. Elles mettent généralement en œuvre un **modèle statistique d'apprentissage automatique** tels que ceux de **Naive Bayes**, de **Régression Logistique**, etc.

Deep Learning

- les **capacités d'apprentissage des algorithmes de *Deep Learning* sont généralement plus puissantes que celles de *Machine Learning* classique**, ce qui permet d'**obtenir de meilleurs scores sur différentes tâches complexes de NLP** dures telles que la traduction

Exemple d’application

- Parmi les tâches de NLP plus complexes, on peut citer la traduction automatique (Machine Translation, MT en anglais). La traduction automatique statistique (*Statistical Machine Translation*, SMT en anglais) repose sur des algorithmes prédictifs qui “apprennent” à partir d’un corpus parallèle, c'est-à-dire un ensemble de textes en plusieurs langues, en relation de traduction mutuelle. La traduction automatique neurale (*Neural Machine Translation*, NMT en anglais) s’appuie sur des algorithmes de Deep Learning.