Introduction au Traitement Automatique des Langues

1 - Introduction Générale

2 - Les applications du TAL

Introduction au traitement automatique des langues

Objectifs de la matière

Le traitement automatique des langues (TAL) vise l'élaboration d'outils et de méthodes capables d'appréhender leur sémantique afin d'en faciliter la prise de connaissance et plus généralement l'exploitation. Selon l'usage que l'on veut en faire, les niveaux d'interprétation peuvent être différents, allant de l'identification de termes pour extraire des mots-clés à des résumés, des traductions ou de la recherche d'informations précises en réponse à des questions. L'objectif de ce module est de présenter les problématiques posées pour le TAL et les principaux modèles pour analyser, synthétiser, exploiter et produire des documents.

Pré requis recommandés **:** programmation, Python, etc.

Unité d'enseignement : UM21
Crédit : 4
Coefficient : 2
Mode d'évaluation : Examen.

Introduction au traitement automatique des langues

Contenu de la matière :

1) Introduction Générale
2) Les applications du TAL
3) Les niveaux de traitement - Traitements de «bas niveau»
4) Les niveaux de traitement - Le niveau lexical
5) Les niveaux de traitement - Le niveau syntaxique
6) Les niveaux de traitement - Le niveau sémantique
7) Les niveaux de traitement - Le niveau pragmatique

Interactions Homme-Machine

Contenu de la matière :

TP : Python

- Python for NLP
- Librairies : nltk, SpaCy, etc.

- Série 1 – Introduction & Bases
- Série 2 – Bas Niveau
- Série 3 – Niveau Lexical
- Série 4 – Niveau Syntaxique

Plan du cours

1. Définitions et généralités
2. Motivations du TAL
3. Brève histoire du TAL
4. Les difficultés du TAL : ambiguïté et implicite
5. Les applications du TAL & exemples
6. Domaines / Topics
7. Outils du TAL – Représentation de connaissances
8. Méthodes et techniques du TAL

Définitions

- Le traitement automatique du langage naturel ou de la langue naturelle (abr. TALN) ou des langues (abr. TAL)
- Ang: Natural Language Processing (abr. NLP)
- Ainsi, le TAL ou TALN est parfois nommé ingénierie linguistique.
- est une discipline à la frontière de la linguistique, de l'informatique et de l'intelligence artificielle, qui concerne l'application de programmes et techniques informatiques à tous les aspects du langage humain.
- Ensemble des recherches et développements visant à modéliser et reproduire, à l'aide de machines, la capacité humaine à produire et à comprendre des énoncés linguistiques dans des buts de communication.
- Traitement du langage naturel et non pas formel (C, Java, etc).
- Traitement de la langue sous forme écrite (texte), le traitement de la parole

Définitions

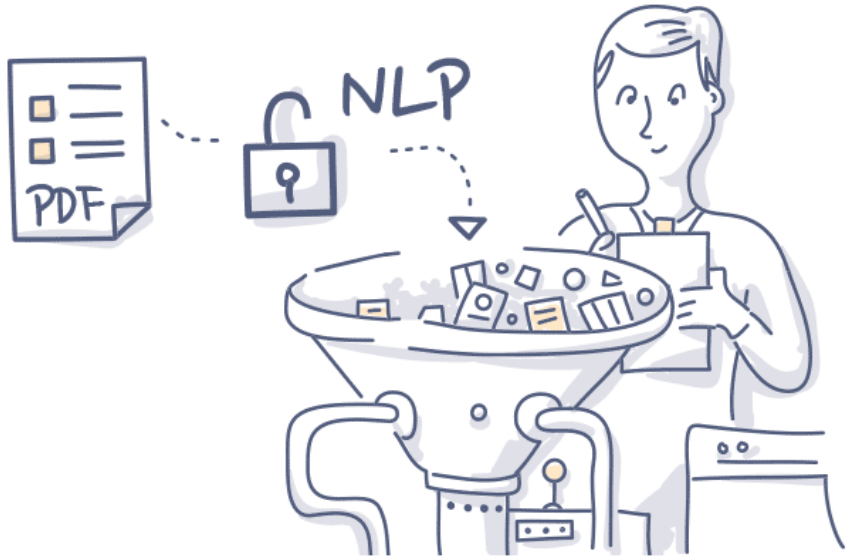
- Le traitement automatique du langage naturel ou de la langue naturelle (abr. TALN) ou des langues (abr. TAL)
- Traitement de la langue sous forme écrite (texte), le traitement de la parole.
- Les données textuelles à traiter se déclinent à l'aune des 3 V (variété, volume, vélocité) – Big Data.
- Elles consistent en des documents écrits, pages Web, emails et autres textes « traditionnels », mais également en contenus de blogs, de réseaux sociaux, en sms, en documents audio transcrits automatiquement, ce qui correspond donc à des types et des qualités de langue très divers.
- Données non structurées ou semi structurées.
- Ces volumes énormes de données textuelles et leurs variétés ont accru le besoin au traitement automatique de la langue.

Motivations du TAL

- Pourquoi s'intéresser à l'automatisation du traitement du langage naturel ?

1. La volonté de modéliser une compétence fascinante (le langage), afin de tester des hypothèses sur les mécanismes de la communication humaine. (Académique)

2. Le besoin de mettre en œuvre des applications capables de traiter efficacement des quantités considérables d'informations « naturelles » aujourd'hui disponibles sous forme électronique. (Pratique)



Brève histoire du TAL

1954 - Traduction automatique, la mise au point du premier traducteur automatique basique. Quelques phrases russes, sélectionnées à l'avance, furent traduites automatiquement en anglais.

1962 - Première conférence sur la traduction automatique est organisée au MIT par Y. Bar-Hillel. « le problème de la traduction automatique est probablement insoluble. »

Entre 1951 et 1954 - Zellig Harris publie ses travaux les plus importants de linguistique (linguistique distributionnaliste).

1957 - N. Chomsky qui publie ses premiers travaux importants sur la syntaxe des langues naturelles, et sur les relations entre grammaires formelles et grammaires naturelles.

1960's - Les élèves de Marvin Minsky développent divers systèmes (BASEBALL (1961), SIR (1964), STUDENT (1964), ELIZA (1966) ...) mettant en œuvre des mécanismes de traitement simples, à base de mots-clés. ELIZA, qui simule un dialogue entre un psychiatre et son patient.

Brève histoire du TAL

1970's - voient le développement d'approches surtout sémantiques (Roger Schank, Yorick Wilks, ...) le rôle de la syntaxe étant pratiquement omis ou, tout du moins considéré comme secondaire.

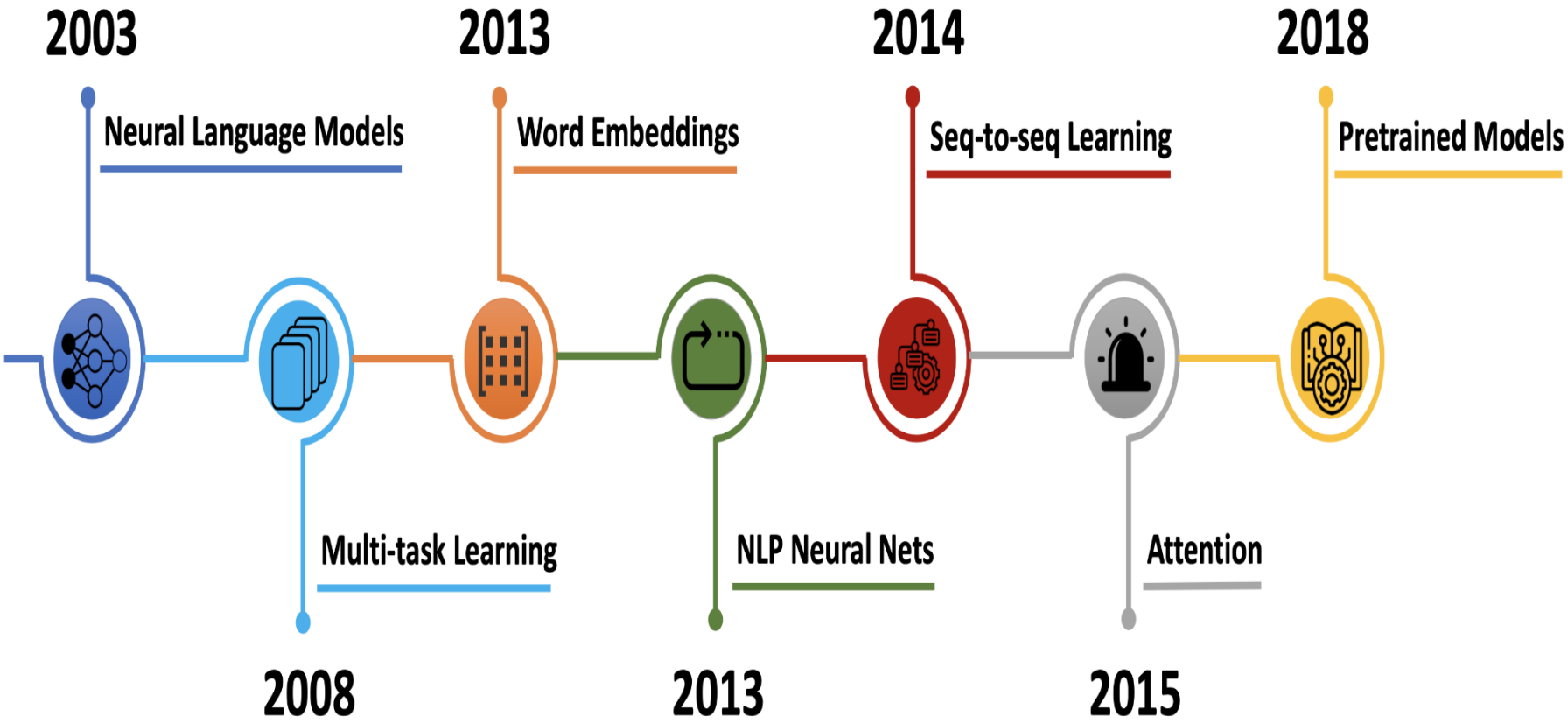
L'importance du contexte, de représentation des connaissances, et le rôle essentiel d'une bonne connaissance du domaine traité pour comprendre un texte est ainsi mis en avant.

Les recherches ont alors cessé de se limiter à l'interprétation de phrases seules pour aborder le traitement d'unités plus importantes comme les récits et les dialogues.

1980's - les propositions issues de l'intelligence artificielle - une partie importante des travaux vise à analyser et à formaliser des mécanismes d'acquisition automatique des connaissances, qui permettent d'extraire directement de lexiques ou de corpus de documents, des règles de grammaire, ou encore des connaissances sémantiques.

Brève histoire du TAL

Histoire récente : Méthodes et techniques IA – Neural NLP

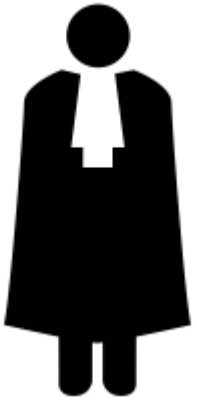


Les difficultés du TAL : ambiguïté et implicite

Les difficultés sont de deux ordres : ambiguïté du langage, et quantité d'implicite contenue dans les communications naturelles.

1 - **Ambiguïté** :

- Le langage naturel est ambigu, à tous ses niveaux (lexical, syntaxique, sémantique, contextuel, etc.)
- Elle se manifeste par la multitude d'interprétations possibles pour chacune des entités linguistiques pertinentes pour un niveau de traitement.
- *Exemple* : lexical sémantique - déterminer si le mot « avocat » se rapporte au domaine juridique ou au domaine alimentaire ?



Les difficultés du TAL : ambiguïté et implicite

Les difficultés sont de deux ordres : ambiguïté du langage, et quantité d'implicite contenue dans les communications naturelles.

1 - **Implicite**:

- L'activité langagière s'inscrit toujours dans un contexte d'interaction entre deux humains, dotés d'une connaissance du monde et de son fonctionnement.

- La machine ne dispose pas de cette connaissance d'arrière-plan, ce qui rend la compréhension complète de la majorité des énoncés difficile, voire impossible.

- Nécessité de disposer de bases de connaissance additionnelles, donnant accès à la fois à un savoir sur le monde (ou le domaine) en général (connaissance statique) et sur le contexte de l'énonciation (connaissance dynamique).

Les difficultés du TAL : ambiguïté et implicite

Les difficultés sont de deux ordres : ambiguïté du langage, et quantité d'implicite contenue dans les communications naturelles.

1 - **Implicite**:

- *Exemple* : désambiguïsation du référent du pronom personnel il :

  - L'étudiant a éteint son smartphone parce qu'**il** ....

    ... **il** était déchargé
    ... **il** voulait révisé

- Figures de style : paradoxes, métaphores, oxymores, antithèses, etc.
  Exemple : Briser la glace.

Les applications du TAL

Les applications du TAL peuvent être regroupé en trois grandes familles :

1 - Le traitement documentaire:

- Les applications qui visent à faciliter le traitement par l'humain des immenses ressources disponibles en langage naturel.

- Exemples:

- La traduction automatique.
- La recherche de documents « intéressants » dans des bases documentaires.
- Le résumé automatique de texte.
- Le routage, classement ou l'indexation automatique de e-documents.
- Questions/Réponses - Correction automatique de réponses écrites.
- La lecture automatisée de documents.
- L'analyse d'un corpus de documents relatifs à un thème donné.
- Des thésauri décrivant des relations entre concepts, WordNet

Les applications du TAL

Les applications du TAL peuvent être regroupé en trois grandes familles :

2 - La production de documents:

- Les applications du domaine de l'aide à la production de texte (la génération de textes), gestion de documents.

- Exemples:

- Correcteurs automatiques,
- La reconnaissance optique de caractères,
- Les correcteurs d'orthographe ou de syntaxe,
- Les correcteurs « stylistiques» ou les aides intelligentes à la rédaction,
- L'apprentissage assisté par ordinateur des langues naturelles,
- La génération automatique de documents à partir de spécifications formelles.

Les applications du TAL

Les applications du TAL peuvent être regroupé en trois grandes familles :

3 - Les interfaces naturelles, IHM:

- Les applications du domaine des interfaces naturelles (i.e. en langage naturel)

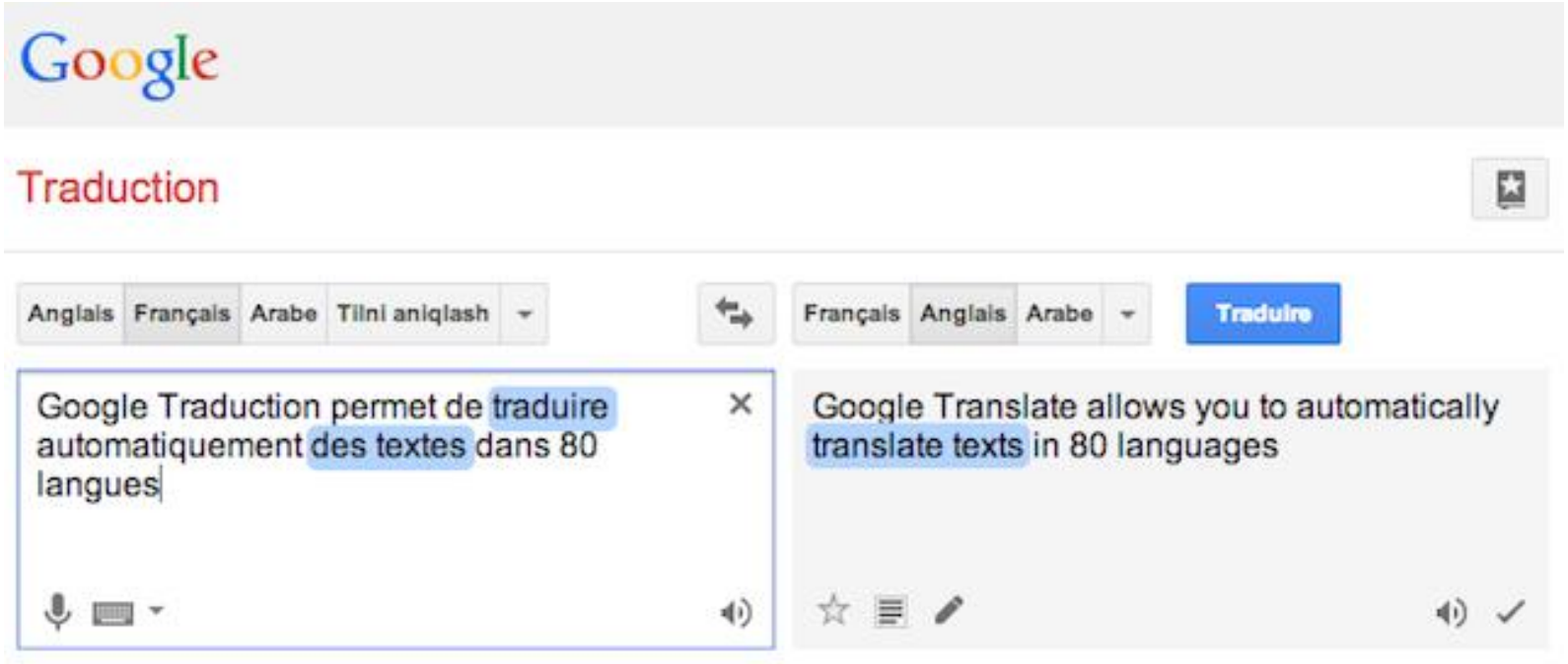
- Exemples:

- Interrogation en langage naturel de bases de données ou de moteurs de recherche.
- Les interfaces vocales, les applications des modules de reconnaissance de parole, synthèse de parole, génération et gestion de dialogue, accès aux bases de connaissance, etc.

Les applications du TAL

1 - Le traitement documentaire:

- La traduction automatique : **Google Traduction**



Les applications du TAL

1 - Le traitement documentaire:

- Le résumé automatique de texte: **Resoomer** - https://resoomer.com/

Les applications du TAL

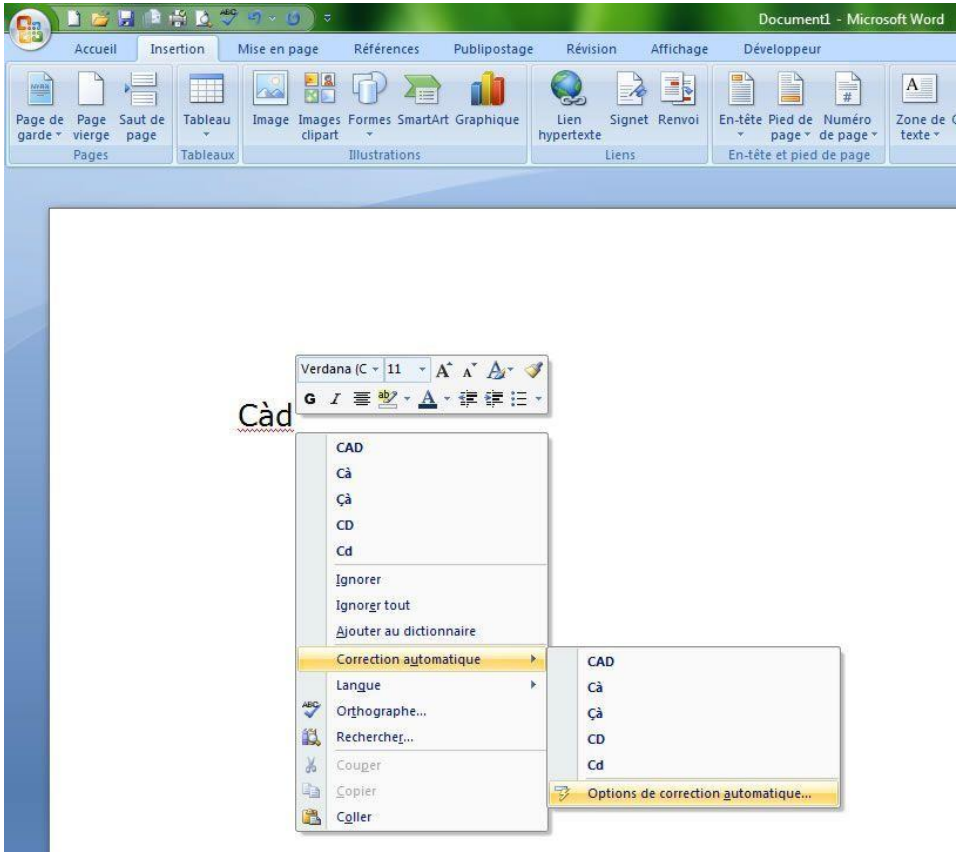
1 - Le traitement documentaire:

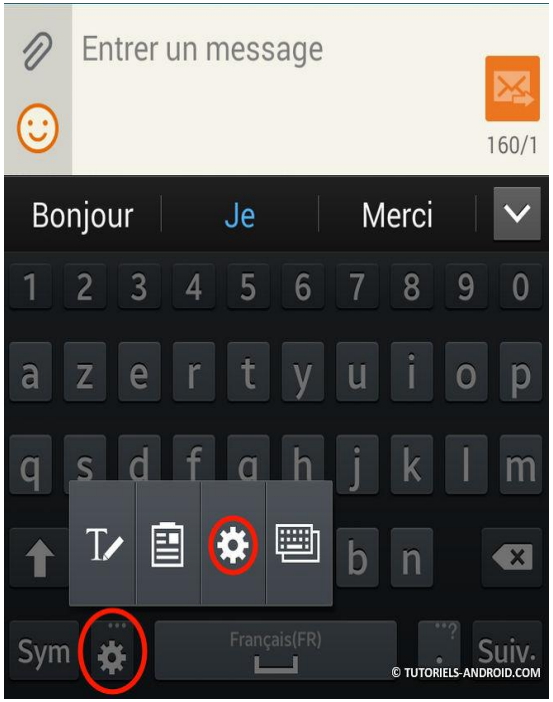
- Questions/Réponses - Correction automatique de réponses écrites.

Les applications du TAL

2 - La production documentaire:

- Correcteurs automatiques.



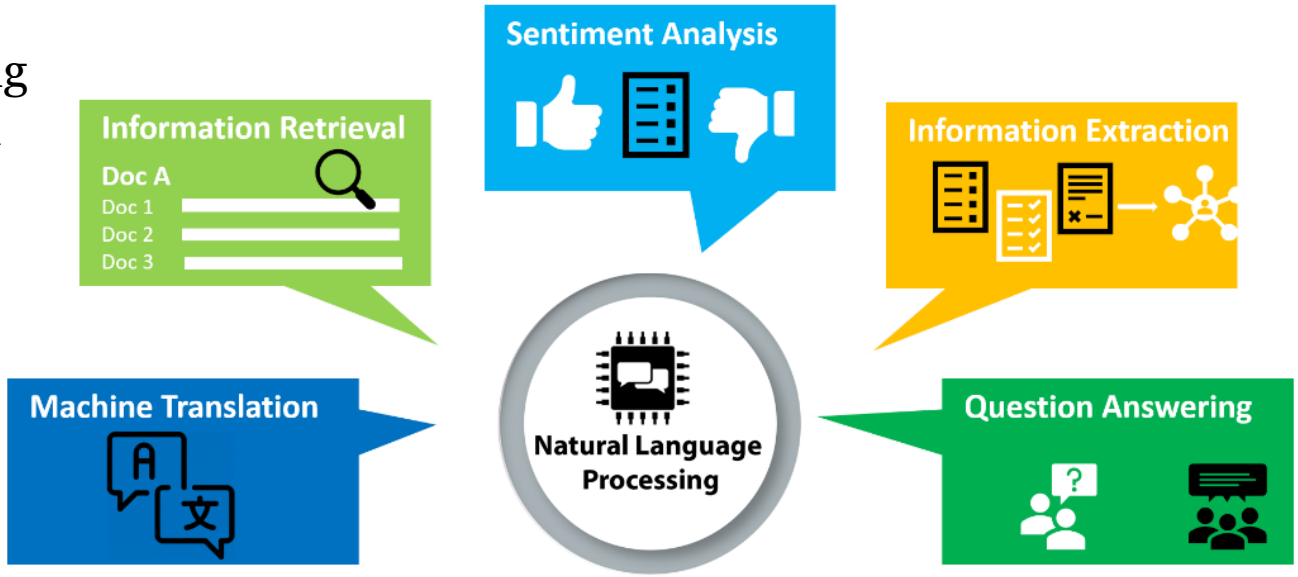


Les applications du TAL

3 - Les interfaces naturelles, IHM : Les interfaces vocales, les applications des modules de reconnaissance de parole, synthèse de parole, etc.

Domaines / Topics du TAL

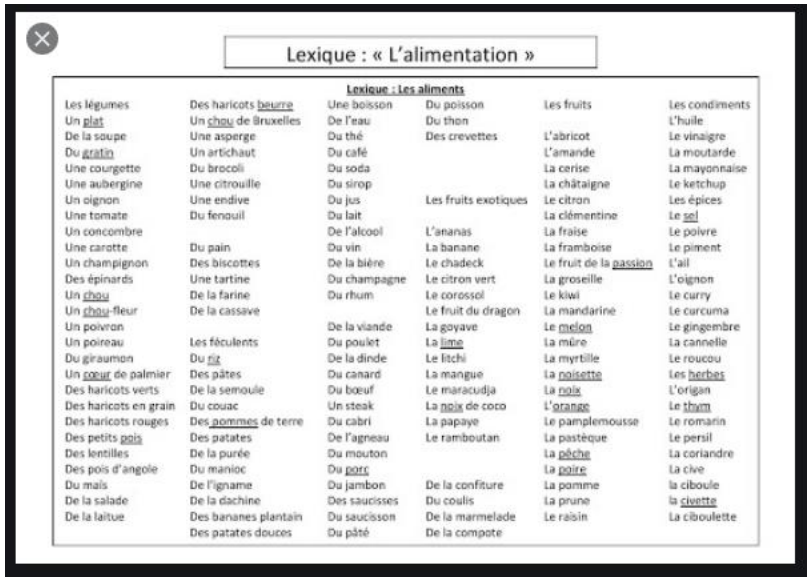
- Machine Translation
- Information Retrieval
- Information Extraction & Named entity recognition
- Text Categorization
- Text Mining
- Summarization
- Sentiment Analysis and Opinion Mining
- Text and Speech Classification
- Topic Modeling
- Question Answering
- Speech recognition
- Dialogue systems
- Etc.



Outils du TAL – Représentation de connaissances

Lexique – Lexicon

- Le lexique d'une langue est l'ensemble de ses mots (somme des vocabulaires utilisés), ou de façon plus précise en linguistique de ses lemmes.
- Dictionnaire - Vocabulaire succinct (d'une langue, d'une science, d'un art, etc.).
- Lexique, avec définitions, restreint à un domaine particulier => **Glossaire - Glossary**



Lexique : « L'alimentation »

Lexique : Les aliments

Les légumes	Des haricots beurre	Une boisson	Du poisson	Les fruits	Les condiments
Un plat	Un chou de Bruxelles	De l'eau	Du thon		L'huile
De la soupe	Une asperge	Du thé	Des crevettes	L'abricot	Le vinaigre
Du gratin	Un artichaut	Du café		L'amande	La moutarde
Une courgette	Du brocoli	Du soda		La cerise	La mayonnaise
Une aubergine	Une citrouille	Du sirop		La châtaigne	Le ketchup
Un oignon	Une endive	Du jus	Les fruits exotiques	Le citron	Les épices
Une tomate	Du fenouil	Du lait		La clémentine	Le sel
Un concombre		De l'alcool	L'ananas	La fraise	Le poivre
Une carotte	Du pain	Du vin	La banane	La framboise	Le piment
Un champignon	Des biscottes	De la bière	Le chadeck	Le fruit de la passion	L'ail
Des épinards	Une tartine	Du champagne	Le citron vert	La groseille	L'oignon
Un chou	De la farine	Du rhum	Le corossol	Le kiwi	Le curry
Un chou-fleur	De la cassave		Le fruit du dragon	La mandarine	Le curcuma
Un poivron		De la viande	La goyave	Le melon	Le gingembre
Un poireau	Les féculents	Du poulet	La lime	La mûre	La cannelle
Du giraumon	Du riz	De la dinde	Le litchi	La myrtille	Le roucou
Un cœur de palmier	Des pâtes	Du canard	La mangue	La noisette	Les herbes
Des haricots verts	De la semoule	Du bœuf	Le maracudja	La noix	L'origan
Des haricots en grain	Du couac	Un steak	La noix de coco	L'orange	Le thym
Des haricots rouges	Des pommes de terre	Du cabri	La papaye	Le pamplemousse	Le romarin
Des petits pois	Des patates	De l'agneau	Le ramboutan	La pastèque	Le persil
Des lentilles	De la purée	Du mouton		La pêche	La coriandre
Des pois d'angole	Du manioc	Du porc		La poire	La cive
Du maïs	De l'igname	Du jambon	De la confiture	La pomme	la ciboule
De la salade	De la dachine	Des saucisses	Du coulis	La prune	la civette
De la laitue	Des bananes plantain	Du saucisson	De la marmelade	Le raisin	La ciboulette
	Des patates douces	Du pâté	De la compote		

Outils du TAL – Représentation de connaissances

Corpus – Corpus

- Recueil / Ensemble fini de documents, artistiques ou non (textes, images, vidéos, etc.) en format électronique, regroupés comme base pour une étude précise.

- En TAL, il permet d'extraire des tendances et notamment de construire des ensembles de n-grammes (une séquence de n mots en TAL).

Sac de mots – Bag of words

- Ensemble fini de mots (=vocabulaire d'un corpus donné).

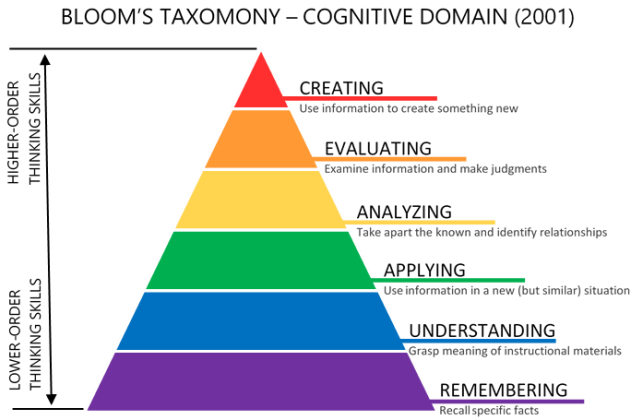
- C'est la manière la plus simple de représenter un document par l'ensemble des mots qu'il contient.

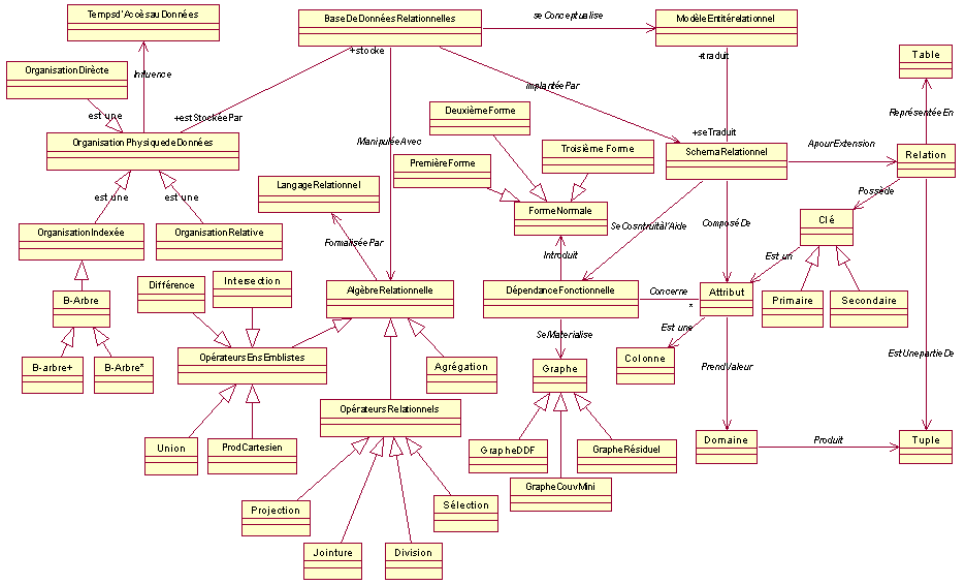
- En pratique, ça peut être par exemple un vecteur de fréquence d'apparition des différents mots utilisés.

Outils du TAL – Représentation de connaissances

Taxonomie – Thesaurus – Ontologie

- La taxinomie est la science des lois de la classification/hiérarchisation.
- Thesaurus est une taxonomie + synonymes et antonymes de mots.
- Ontologie est une sorte de taxonomie + sémantique / raisonnement / inférence. Ensemble structuré des termes et concepts représentant le sens d'un champ d'informations





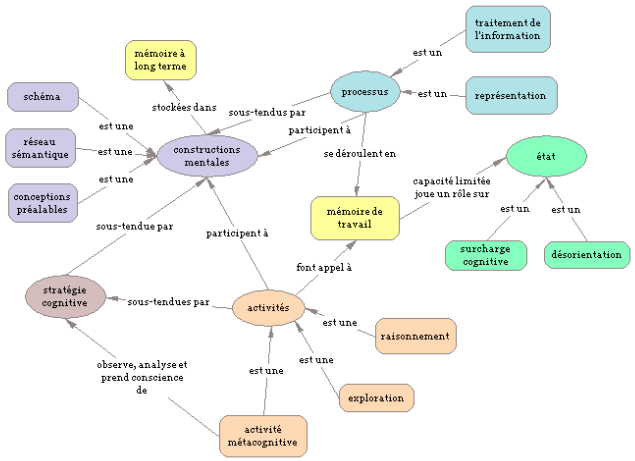
Outils du TAL – Représentation de connaissances

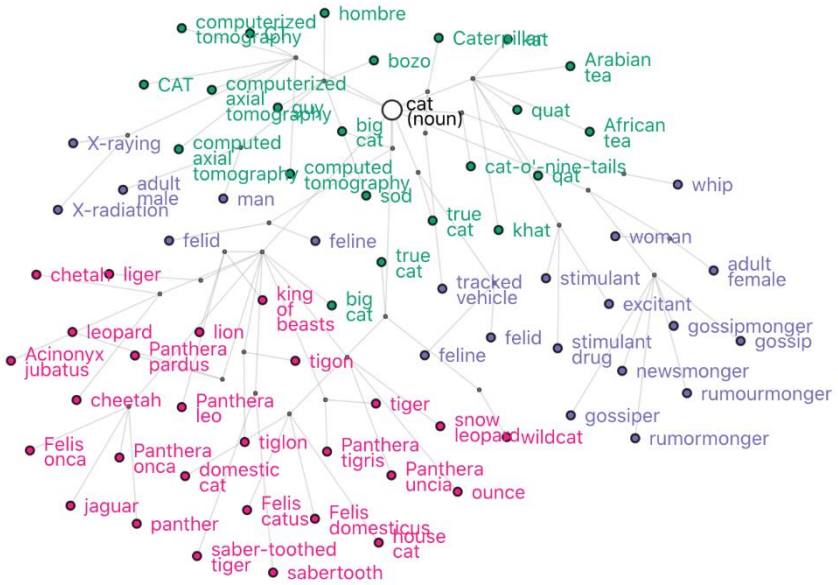
Réseau sémantique – WordNet

- Est une base de connaissances qui représente les relations sémantiques entre les concepts d'un réseau. Graphe orienté et étiqueté.

- WordNet est une base de données lexicale. Son but est de répertorier, classifier et mettre en relation de diverses manières le contenu sémantique et lexical de la langue anglaise (et d'autres).

- http://wordnetweb.princeton.edu/pe





Méthodes et techniques du TAL

Comment implémenter les systèmes TAL ?

- Les méthodes (approches) symboliques : méthodes fondées sur des règles fondées sur l'expertise humaine; modélisation parfois logique, automates, etc.
- Les méthodes stochastiques: méthodes probabilistes fondées sur des calculs statistiques effectués à partir de corpus. Statistical inference.
- Approches empiriques fondées sur les données, où la connaissance est extraite par des techniques d'apprentissage automatique :
- Machine Learning
- Neural networks / Deep Learning

Références

Cours - *François Yvon* – Une petite introduction au Traitement Automatique des Langues Naturelles,
https://perso.limsi.fr/anne/coursM2R/intro.pdf

Article – Marcel Cori - Des méthodes de traitement automatique aux linguistiques fondées sur les corpus
- https://www.cairn.info/revue-langages-2008-3-page-95.htm

Article - Pascale Sébillot - Le traitement automatique des langues face aux données textuelles volumineuses et potentiellement dégradées : qu'est-ce que cela change ?
- https://hal.archives-ouvertes.fr/hal-01056396/document