

IDL Cours 1

Introduction à l'ingénierie des langues

- Ingénierie des langues IDL
- Traitement automatique des langues TAL
- Natural language processing NLP

L'ingénierie des langues au quotidien (Activité 1)

? Quels sont les outils issus de cette discipline que vous utilisez régulièrement ?

▼ Details

- moteurs de recherche (extraction d'information)
- traduction automatique
- agents conversationnels
- correcteurs orthographiques
- suggestion de texte
- outils de recommandation (eg. Netflix)
- détection de spam
- identification de langue (cf. google translate)
- résumé automatique
- reconnaissance / sythèse de la parole

? Quels domaines d'application de ces technologies ?

▼ Details

- la guerre (espionnage, surveillance)
- la santé (diagnostic automatique)
- le droit (détection plagiat, analyse de documents juridiques, recherche jurisprudentielle)
- linguistique (étude de phénomènes linguistiques)
- humanités numériques (traitement de grands volumes de données)
- industries privées : veille de réputation (entreprise, produit), conseil clientèle

Question subsidiaire : qui recrute ? Voir les offres https://www.linkedin.com/jobs/search/?currentJobId=3802132660&keywords=NLP&origin=JOBS_HOME_SEARCH_BUTTON&refresh=true

Couches de traitement sur la langue (Activité 2)

? Quelle est la traduction de la phrase suivante ? Quelles sont les étapes auxquelles vous avez dû procéder pour la traduire et de quelles informations (implicites) avez vous eu besoin ?

🇬🇧 Marc gives his laptop to Jane. She's still mad so she takes her book and his apple computer and goes to listen some rock with Claire. They are best friends.

Traduction humaine (sans ordinateur !)

1. Vers le français

▼ Details

🇫🇷 👍 Marc donne son ordinateur portable à Jane. Elle est toujours énervée donc elle prend son livre et son ordinateur Apple et va écouter du rock avec Claire. Elles sont meilleures amies.

2. Traduisez de nouveau vers l'anglais

▼ Details

🇬🇧 Mark gives his portable computer to Jane. She is still mad so she takes her book and her computer Apple and goes to listen to rock with Claire. They are best friends.

Les étapes (explicites et implicites)

Personne ne donne la solution suivante :

🇫🇷 👎 "Marc donne son ordinateur portable à Jane. Elle's toujours folle donc elle prend son réserver et son Pomme ordinateur et va écouter du pierre avec Claire. Ils sont meilleurs amis."

... car vous procédez naturellement à ces différentes étapes :

1. Segmentation,
2. analyse morphosyntaxique,
3. analyse syntaxique,
4. reconnaissance des entités nommées,

5. analyse sémantique.

Détail des étapes :

► Details

? Est-ce que c'est plus dur que la traduction de python vers de l'assembleur mips ?



▼ Details

- obstacles linguistiques : ambiguïté
- obstacles techniques : quelles sont les méthodes mises en place pour arriver à ce résultat ?

Traduction automatique



- Quelle traduction donne Google Translate ?

▼ Details

  Marc donne son ordinateur portable à Jane. Elle est toujours en colère alors elle prend son livre et son ordinateur Apple et va écouter du rock avec Claire. Ils sont meilleurs amis. => pas mal !

- Quelle traduction donne DeepL ? [https://www.deepl.com/translator#en/fr/Marc gives his laptop to Jane. She's still mad so she takes her book and her apple computer and goes to listen some rock with Claire. They are best friends](https://www.deepl.com/translator#en/fr/Marc%20gives%20his%20laptop%20to%20Jane.%20She's%20still%20mad%20so%20she%20takes%20her%20book%20and%20her%20apple%20computer%20and%20goes%20to%20listen%20some%20rock%20with%20Claire.%20They%20are%20best%20friends)
(<https://www.deepl.com/translator#en/fr/Marc%20gives%20his%20laptop%20to%20Jane.%20She's%20still%20mad%20so%20she%20takes%20her%20book%20and%20her%20apple%20computer%20and%20goes%20to%20listen%20some%20rock%20with%20Claire.%20They%20are%20best%20friends>)

▼ Details

  Marc donne son ordinateur portable à Jane. Elle est toujours en colère et prend son livre et son ordinateur pour aller écouter du rock avec Claire. Elles sont les meilleures amies du monde. => pas mal non plus.

? Mais comment font les outils ?

Revenons à la traduction humaine :

1. Analyse -> construction d'une représentation du sens dans votre tête, sans que vous n'ayez vraiment "pensé" à la traduction
2. Génération -> production de la traduction attendue.
 - traduction de certains mots que vous connaissez "par coeur" (= lexique bilingue) ... mais vous ne connaissez pas la traduction de toutes les

phrases possibles et imaginables par coeur !

- connaissance de correspondances entre certaines structures de phrases:
 - she goes to listen = elle va écouter
 - she goes crazy = elle devient folle.
- connaissance de la langue cible qui vous permet de produire une traduction syntaxiquement correcte.
 - she goes to listen = elle va écouter
 - she goes to Los Angeles = elle va à Los Angeles.

Vous avez emmagasiné assez de connaissances de la langue source, de la langue cible et des liens entre les deux pour traiter des phrases que vous n'aviez jamais vues ni entendues.

? comment avez-vous emmagasiné ces connaissances ? Les machines peuvent-elles faire pareil ?

▼ Details

- manuels
- cours de langue
- traduction ("que veut dire ceci ?")
- livres / musiques / films / conversations

Les machines font (quasiment) pareil : de nos jours, nombre d'outils fonctionnent sur des mécanismes d'apprentissage qui ne sont pas fondamentalement éloignés des nôtres.

Cela n'a pas toujours été le cas ! Les premières approches (qui datent de la seconde guerre mondiale) fonctionnaient par traduction "mot-à-mot", grâce à de petits lexiques bilingues, plus tard améliorées en intégrant la syntaxe c'est-à-dire avec réaligement des phrases. Puis les approches statistiques, fondées sur "l'apprentissage" sont apparues.




? Mais comment fonctionne réellement cet apprentissage ?

- des données : lexiques, corpus, corpus annotés, corpus bilingues alignés ;
- des algorithmes d'apprentissage.

Take away message

Pour faire des outils de traitement des langues de nos jours, on a besoin :

-  de ressources linguistiques ;

-  d'enrichir ces ressources avec des informations d'ordre linguistique ;
-  d'alimenter des algorithmes d'apprentissage ;
-  de savoir évaluer les performances des outils entraînés.

Exemple simple (au tableau) : un outil connaît tout seul la catégorie grammaticale d'un mot.

Qu'allons nous faire dans ce cours ?

- Apprendre à structurer et à enrichir des données pour les rendre exploitables par une machine (TP1, TP2, TP3)
- Apprendre à évaluer des outils d'ingénierie des langues (TP4)
- Coder un petit agent conversationnel à base de règles (TP5)
- Créer un corpus de textes à partir de Wikipédia (TP6)
- Étudier la représentation des mots : comment encoder la sémantique, ie le sens d'un mot ? (TP7) // un ordinateur comprend généralement mieux les chiffres que les mots.
- Tester différentes méthodes de classification de documents (TP8 - TP9)
- Vous aider dans le suivi de votre projet : développer un petit chatbot de recommandation

Avant de se quitter : pourquoi nous nous intéressons à ce domaine ?

La question de la diversité linguistique. Exemple : traduction du français vers le Quechua

image

Modalités d'évaluation :

- TP à rendre avant la séance suivante ;
- chaque séance de TP commence par la correction en classe des TPs de 4 étudiant.es <http://xn--tudiant-9xa.es> tiré.es <http://xn--tir-dma.es> au sort pour être évalué.es <http://xn--valu-9oae.es> Vous devez donc tous et toutes être prêt.es <http://xn--prt-gma.es> à répondre à des questions sur votre rendu qui pourra être montré en cours. Si vous êtes absent.es <http://absent.es> deux fois alors que vous êtes tiré.es <http://xn--tir-dma.es> au sort, vous avez zéro à la note de TP ;
- projet final avec soutenance ;

- + si vous n'apprenez pas le cours d'une fois sur l'autre, des interros surprise avec des questions de cours.