Ateliers Informatique TAL

Traitement automatique des langues

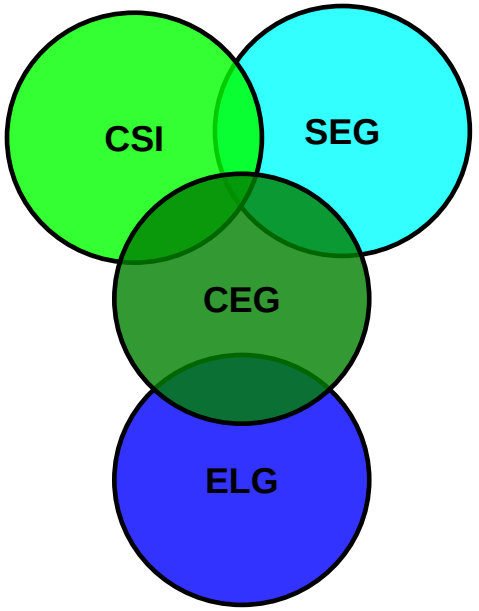
Université d’Ottawa

- Faculté de génie
- Informatique

Disciplines des technologies de l'information

- Nous enseignons quatre programmes à l'Ecole die SIGE

  - **Informatique (CSI)**: Pour les étudiants qui désirent développer des applications et leurs technologies sous-jacentes
  - **Génie logiciel (SEG)**: Pour les étudiants qui veulent se concentrer sur les exigences, designs, et architectures pour systèmes de grande taille, et sur la gestion de projets
  - **Génie informatique (CEG)**: Pour les étudiants plus intéressés par les aspects matériel et par la programmation de bas niveau
  - **Génie électrique (ELG)**: pour les étudiants intéressés par le matériel, les communications (sans-fil, optiques, réseautique), l'électronique de puissance, la robotique.



- Les perspectives de travail sont excellentes pour tous ces programmes. Aussi option **COOP**.

Informatique

- Architecture des ordinateurs et de l'ingénierie
- Infographie et visualisation
- La sécurité informatique et la cryptographie
- Le calcul scientifique
- Réseaux informatiques
- Des systèmes concurrents, parallèles et distribués
- Bases de données et la récupération de l'information
- informatique de la santé
- Sciences de l'information
- Génie logiciel
- L'intelligence artificielle

L’intelligence artificielle

- Les ordinateurs seraient beaucoup plus utiles s’ils pouvaient
  - gérer nos courriels
  - faire nos recherches
  - nous parler ...
- Mais ils sont mêlés par le langage des humains.
- Comment pouvons-nous leur expliquer notre langage?
- (Ou les aider à apprendre, comme font les enfants?)

L’étude de traitement automatique des langues

- Comment les ordinateurs peuvent faire des choses utiles avec les langues naturelles:
  - compréhension du langage naturel
  - génération du langage
- Inclus
  - Linguistique
  - l’informatique théorétique
  - intelligence artificielle
  - Mathématique et de statistique
  - Psychologie
  - sciences cognitives
  - etc

pas une recherche par mot clé

- Google, Bing, Yahoo - 19.4 milliards de recherches en juillet 2013
- basées sur des mots clés
- interface en langage naturel utilise des verbes, des phrases et des clauses pour comprendre la nature de la question

Exemple

- « quel province a une taxe de vente la plus élevée? »
- Recherche mot clés
  - province, taxe, vente
  - résultats: toutes les taxes de vente provinciales, taux
- Recherche langue naturelle
  - résultats: une province

Niveaux de langage

- Phonétique/phonologie/morphologie – les sons et la forme des mots
- Syntaxe: la façon dont les mots se combinent pour former des phrases
- Semantiques: la signification de mots ne peut pas être compris sans connaitre le contexte de leurs emplois

Difficultés

- Ambiguïté
  - Cela peut rajeunir un homme âgé de trente ans.
  - On veut empêcher cette évolution accélérée par l'intervention du gouvernement.
- Inexacte – beaucoup, nombreux, plusieurs
- consiste à raisonner comme le monde
- intégré dans un système social
- toujours en évolution

Traitement automatique des langues

- Fouille des textes
- Système réponses-questions
- Résumé des textes

On va les adapter ensemble pour Français !

Fouille des textes

- l'extraction de connaissances dans les textes
- inclus:
  - Recherche d’information
  - Reconnaissance d'entités nommées
  - Coréférence
  - l'analyse des sentiments
  - Etc.

Index inversé

- Correspondance entre des mots et leurs position dans un ensemble de données
- le même principe qu'un index terminologique
- permettre une recherche plein texte plus rapide

Exemple

- D1 = "L'art de plaire est l'art de tromper.",
- D2 = "Nous avons l’art, afin de ne pas mourir de la vérité.«
  - "art" {D1, D2}
  - "plaire" {D1}
  - "est" {D1}
  - "tromper" {D1}
  - "nous" {D2}
  - "avons" {D2}
  - "afin" {D2}
  - "pas" {D2}
  - "mourir" {D2}
  - "vérité" {D2}

Tokenization

- le processus de casser de texte en mots, phrases, symboles ou d'autres éléments significatifs appelés jetons
  - Phillip, a, sauté, sauts, va, sauter
- pas toujours facile
  - réponses-questions
  - l’art
  - 2ieme

Mots Vides

- sont tellement communs qu'il est inutile de les indexer ou de les utiliser dans une recherche
  - le chien va sauter
  - Les chiens vont sauter

- Est-ce que « le » ou « les » ajoute une valeur?
- distribution est trop élevée
- les prépositions, les articles, les pronoms

Preprocessing

- si nous voulons savoir qui **saute**
- Est-ce que ne nous recherchons uniquement pour « saute » ?
  - Phillip a sauté
  - Phillip sauts
  - Phillip va sauter

Stemming

- le procédé de transformation des mots en leur racine
- la racine d'un mot correspond à la partie du mot restante une fois que l'on a supprimé son préfixe et son suffixe
- la racine ne correspond généralement pas à un mot réel
  - sauté, sauts, sauter = saut
  - cheval, chevaux, chevalier, chevalerie, chevaucher = cheva

- Carry, un algorithme de racinisation pour le français

Applications

- Sécurité - surveillance et analyse
- Marketing - gestion de les relations avec la clientèle
- critiques de films– score global
- biomédicale - Reconnaissance d'entités nommées
  - MONTRÉAL – Le capitaine de l'Impact de Montréal Davy Arnaud ne sera pas de retour avec l’équipe en 2014.
  - L'**asthme** est une maladie du système respiratoire touchant les voies aériennes inférieures et notamment les bronchioles, définie comme étant une gêne respiratoire à l'expiration.

Questions

- Quel pays a été le premier à vendre armements à un pays du tiers monde?
- Combien de morts dans un soulèvement de prison au Venezuela en 1994?
- En 1994, ce qui était plus important que combattre le chômage et le sida?

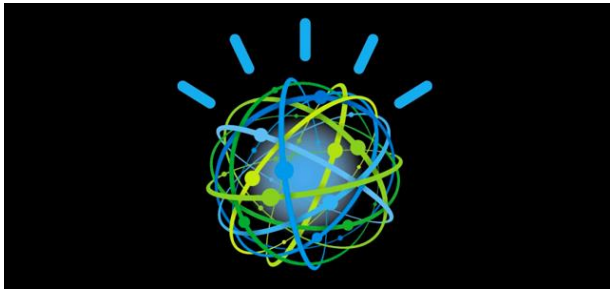
Système réponses-questions

- recherche d'information exploitant des requêtes formulées à l'aide du langage naturel
- 3 buts principaux
  - Comprendre les questions en langage naturel
  - Trouver les informations
  - Répondre à la question
- Habituellement un domaine spécialisé
  - systèmes médicaux
  - réservations aériennes

Exemple de type de requêtes

Questions factuelles	« Où a été brûlée Jeanne d'Arc ? »
Questions booléennes (oui ou non)	« Hosni Moubarak est-il toujours président ? »
Définitions	« Que signifie le sigle IHM ? »
Causes/Conséquences	« Pourquoi la mer est-elle bleue ? »
Procédures	« Comment refaire sa carte d'identité ? »
Listes	« Citer 3 présidents américains »
Requêtes évaluatives/comparatives	« Quelle est la plus grande ville de France ? »
Opinions	« Que pensent les Français de Nicolas Sarkozy ? »

Watson



- Développé par IBM
- gagné, au jeu télévisé Jeopardy! lors de trois épisodes diffusés aux États-Unis les 14, 15 et 16 février 2011
- capable de
  - comprendre l'énoncé des questions
  - buzzer pour prendre la main
  - trouver les réponses en quelques secondes
  - et grâce à une synthèse vocale, énoncer les réponses


THINK
Ken
WATSON
BRAD

DeepQA

- 3 ans
- 20 chercheurs et ingénieurs
- aucun accès à Internet pendant le jeu
- nombreuses ressources collectées - encyclopédies, dictionnaires, thésaurus, des articles Newswire, œuvres littéraires, etc
- parallélisme massif
- estimation de la confiance généralisée
- intégration de la connaissance approfondie et superficielle

Critères d'évaluation

- précision de question-réponse
- vitesse
- estimation de la confiance
- sélection d'indice
- stratégie de paris
- Objectif: récompense d'argent

Nouvelle ère de la technologie

- l' ère de l'informatique cognitive
- transformer les entreprises et la société
- capable de donner un sens à de vastes quantités d'informations non structurées par
  - l'apprentissage
  - le raisonnement
  - l'interaction avec les gens de façon plus naturel
- Un service cloud
  - Watson n’est pas enfermé dans une boîte

Question

- Qu'est-ce que représente DUP?

Résumés automatique de textes

- Une forme de compression textuelle avec perte d'information obtenu au moyen de techniques informatiques.
- 2 approches
  - Extraction
  - Abstraction

Extraction

- extraction des phrases complètes censées être les plus pertinentes du document
- les concaténer de façon à produire un extrait

Abstraction

- L’approche la plus difficile
- vise à rédiger un résumé en générant des phrases pas forcément contenues dans l'original
- La réécriture ou paraphrase est aussi utilisé

Applications

- Résumé d'un texte unique
  - Générique
  - Pertinente aux requêtes
- Résumé de plusieurs textes
  - Collection des documents connexes
  - regroupement de reportages

Multimillionnaire a l'âge de 15 ans

- Nick D’Aloisio, un étudiant de lycée du Royaume-Uni
- En mars 2011, crée l’application Trimit (renommé par la suite Summly) pour iOS
- Résume des textes en 1000, 500 ou 140 caractères
- Vendu en mars 2013 a Yahoo pour 30$ millions US
- Reçu prix de la meilleure application « Intuitive Touch » par Apple en 2012

Evaluation

- intrinsèque vs extrinsèque
  - Intrinsèque
    - évalue la cohérence et le caractère informatif de résumés
  - Extrinsèque
    - testé l'impact de synthèse sur des tâches
    - évalue la pertinence ou la compréhension de la lecture
- Intertextuel et intra-textuelle
  - évaluer un résultat ou comparer plusieurs résultats

Difficultés de l'évaluation

- impossibilité de construire un étalon-or équitable
- une personne peut choisir des phrases différentes à des moments différents
- deux phrases distinctes exprimées en des termes différents peuvent exprimer le même sens
- des résumés écrits par de juges humains

Conclusions et l'avenir

- l'apprentissage est la base de intelligence artificielle
- allons-nous atteindre le singularité technologique
- Questions d’éthique