Traitement automatique des langues : Fondements et applications

Cours 1 : Présentation du domaine

Plan

Qu’est-ce que c’est ?

- traitement informatique de données en langage naturel, surtout l’écrit
- toutes les formes d’écrits: livres, documents techniques, forums, emails, chats, blogs
- représenter l’information contenue dans les données textuelles et la communiquer

Pourquoi ?

- IA et représentation de connaissances : on ne fera pas tout à la main

  immense réservoir dans les textes disponibles

- faciliter la communication Humain/Machine et Humain/Humain

  accès à l'information, médiation de la communication

- langage comme trace de la pensée, du raisonnement et du sens commun

  comprendre l'intelligence par la communication

Pour quoi faire ?

Applications

- fouille / extraction d'information
  - analyse d'opinions
  - réponse à des questions
- traduction automatique entre langues
- résumé de textes
- construction de bases de connaissances / ontologies

sous-domaines très riches : médical, juridique, domaines techniques

Quelle(s) discipline(s)

- 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

Exemple: le Question-réponse avec Watson

Watson

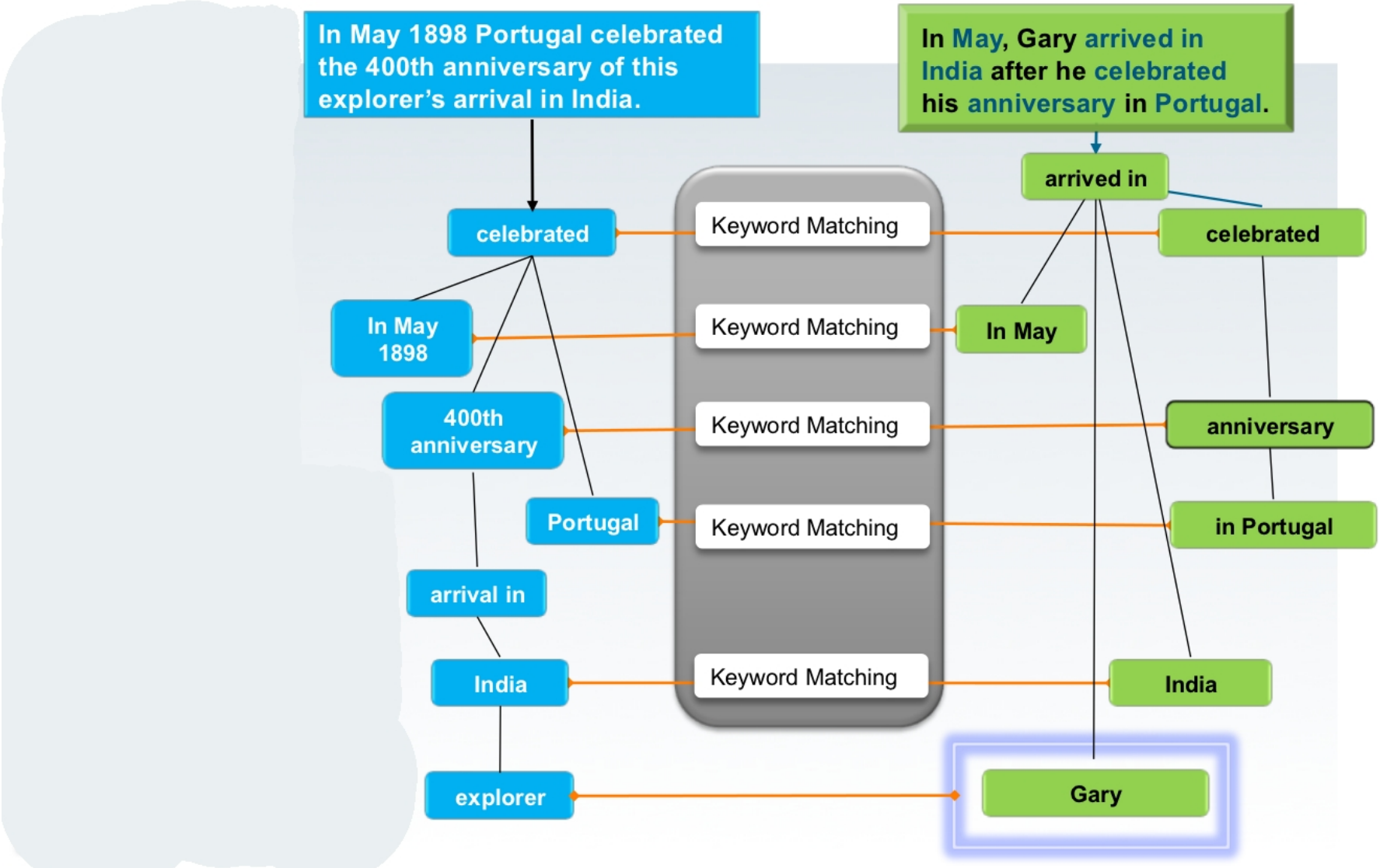
un système de réponse à des questions, dans le format du jeu “jéopardy!”
combine :

- traitement du langage naturel (analyse/production)
- apprentissage automatique
- représentation de connaissances / décision

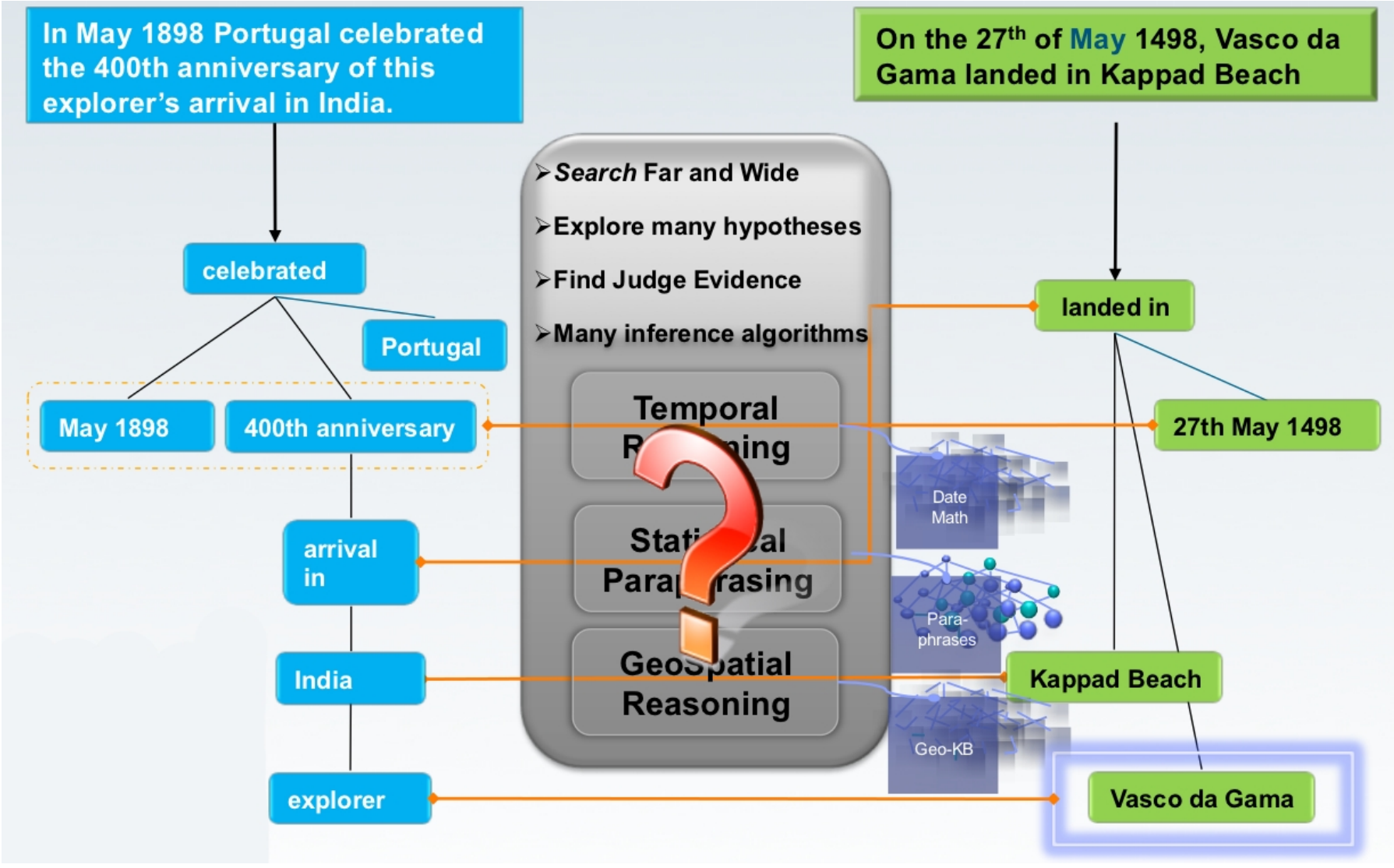
$\rightarrow$ a battu tous les champions du jeu (en temps réel) !

un exemple de recherche d’information sophistiquée

Les mots clefs ne suffisent pas



Analyse sémantique, raisonnement, décision



Analyse d'opinion

- suivi d'opinion sur un produit
- aggrégation d'avis
- analyse fine pour recommendations

source : xkcd; allociné

Analyse d'opinion

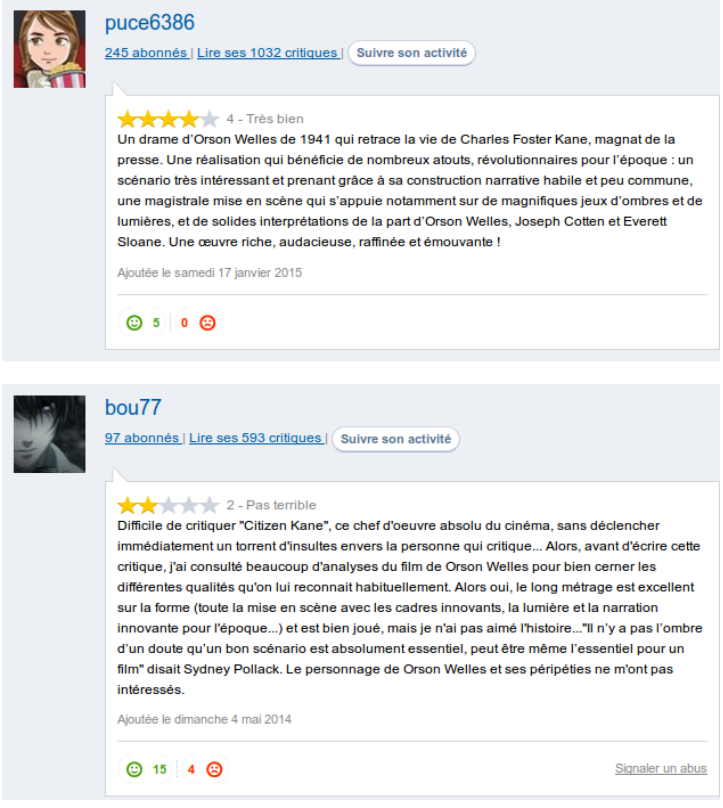
- suivi d'opinion sur un produit
- aggrégation d'avis
- analyse fine pour recommendations

source : xkcd; allociné

Analyse d'opinion

- suivi d'opinion sur un produit
- aggrégation d'avis
- analyse fine pour recommendations

versus



source : xkcd; allociné

Traduction automatique

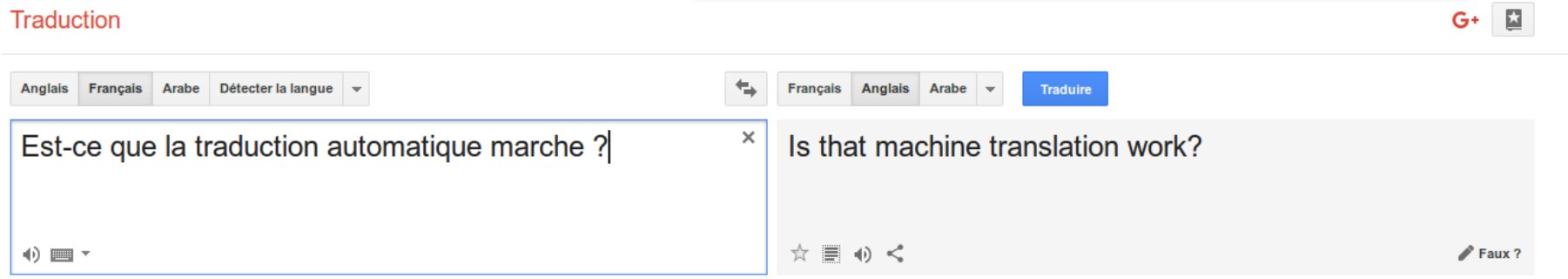
Alignement → Aide à la traduction

www.linguee.fr

After suffering a crushing defeat to his one-time friend, Sasuke, Naruto must pick up the pieces and train with the Leaf [...] — nintendo.co.uk	Après avoir subi une défaite écrasante face à son ami d'antan, Sasuke, Naruto doit se reprendre et s'entraîner avec les [...] — nintendo.fr
John Turner led the members opposite to a crushing defeat because he opposed expanding trade with our largest trading partner. — www2.parl.gc.ca	Le très honorable John Turner a mené les députés d'en face à une cuisante défaite, parce qu'il s'opposait à l'expansion des échanges avec [...] — www2.parl.gc.ca
As soon as your darling sees you, he'll forget all about his team's crushing defeat! — ca.clarins.com	En vous regardant, votre chéri oubliera vite la défaite de son équipe ! — ca-fr-new.clarins.net
[...] players, this handful of elite soldiers succeeds, more or less easily, in inflicting a crushing defeat on the Russian troops. — esisc.net	[...] joueurs, cette poignée de soldats d'élite parvient, plus ou moins aisément, à infliger une cuisante défaite aux troupes russes. — esisc.net

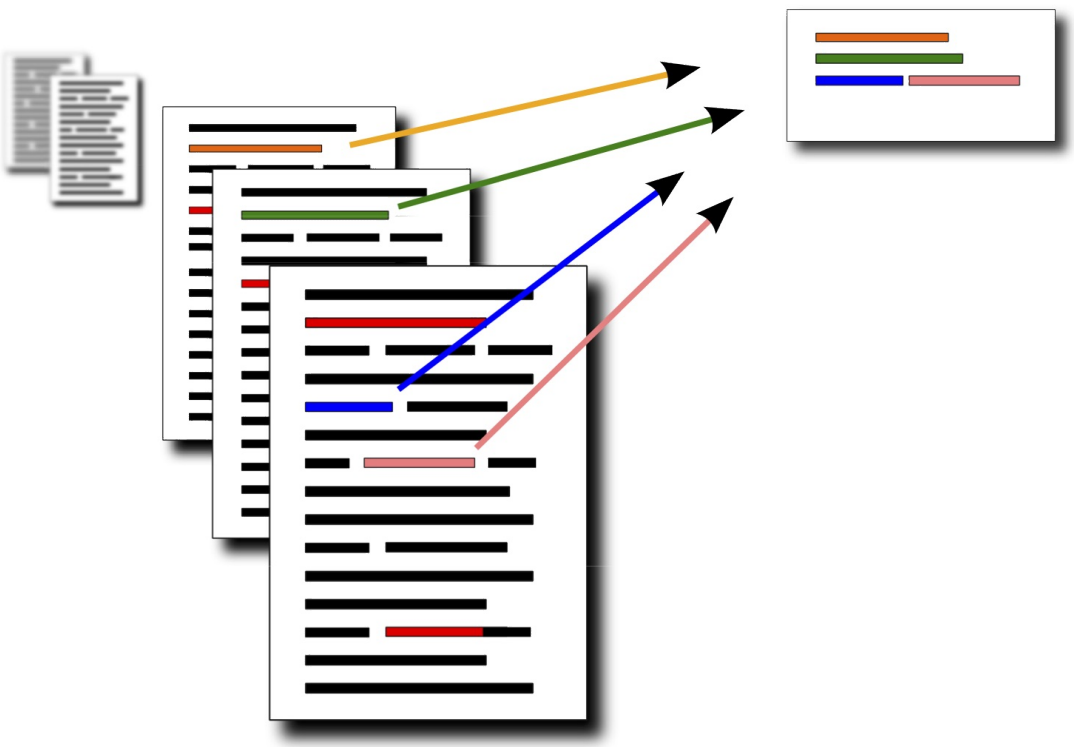
Traduction automatique (statistique)

google translate



Résumés de textes

- synthèse de documents
- simplification
- le plus souvent par extraction de phrases



source : Mikael Kågebäk

Pourquoi c'est dur ?

- ambiguité à tous les niveaux
- beaucoup d'implicite
- beaucoup d'équivalences de sens sous des formes différentes
- opposition phénomènes courants/phénomènes rares
  mais beaucoup de phénomènes rares
  $\rightarrow$ difficile à modéliser

Quelques exemples de problèmes

la ligne Paris-Toulouse
la ligne Bordeaux Saint-Jean-Mont-de-Marsan

un mouton noir; arriver après la bataille; jeter l'éponge

J'ai bien aimé Une journée en enfer.
Le facteur Cheval est un artiste célèbre.

Quelques exemples de problèmes

la ligne Paris-Toulouse
la ligne Bordeaux Saint-Jean-Mont-de-Marsan

segmentation

un mouton noir; arriver après la bataille; jeter l'éponge

J'ai bien aimé Une journée en enfer.
Le facteur Cheval est un artiste célèbre.

Quelques exemples de problèmes

la ligne Paris-Toulouse
la ligne Bordeaux Saint-Jean-Mont-de-Marsan

segmentation

un mouton noir; arriver après la bataille; jeter l'éponge

locutions

J'ai bien aimé Une journée en enfer.
Le facteur Cheval est un artiste célèbre.

Quelques exemples de problèmes

la ligne Paris-Toulouse
la ligne Bordeaux Saint-Jean-Mont-de-Marsan

segmentation

un mouton noir; arriver après la bataille; jeter l'éponge

locutions

J'ai bien aimé Une journée en enfer.
Le facteur Cheval est un artiste célèbre.

noms d'entités

Quelques exemples de problèmes

j'croibi1k G1 pb

tweet; biopic; abracadabrantesque

Marie et Jean sont amis. Marie et Jean sont français.

Marie et Jean sont de bons amis

Quelques exemples de problèmes

j'croibi1k G1 pb langue non-standard

tweet; biopic; abracadabrantesque

Marie et Jean sont amis. Marie et Jean sont français.

Marie et Jean sont de bons amis

Quelques exemples de 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.

Marie et Jean sont de bons amis

Quelques exemples de 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

Quelques exemples de 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

Un peu de contexte linguistique

Les niveaux d'analyse

- phonologie: les sons
- morphologie: les mots et leur forme
- syntaxe: l'organisation des mots en phrase
- sémantique: le sens dans la phrase
- pragmatique: le sens en contexte

Exemples d'ambiguité à tous les niveaux

Phonologie

- homophones : vers, ver, vert, verre,
- important pour la correction / language "non standard"
- relativement mineur cependant

Exemples d'ambiguité à tous les niveaux

Morphologie

- homonymes/homographes :
  - car (conjonction) / car (nom)
  - brise (nom) / brise (verbe)
  - voler (dans le ciel) / voler (la banque)
- complique l'analyse syntaxique
- complique l'analyse sémantique
- peut se combiner pour multiplier le problème:
  La petite brise la glace

Exemples d'ambiguité à tous les niveaux

Parfois ... il n'y a rien à comprendre

- "La pente est rude mais la route est droite." (JP. Raffarin)
- "Le libéralisme est une valeur de gauche." (E. Macron)
- "C'est beau ce stade vélodrome qui est toujours plein à l'extérieur comme à domicile." (F. Ribéry)

Plan

Sac de mots

a bag of words

pourquoi s'embêter ? on peut juste regarder au niveau des mots

$\rightarrow$ approche "recherche d'information" : indexer un texte selon les mots importants / apparier avec une requête

ou dans le résumé, chercher les phrases avec les mots importants

ou en traduction, apparier des mots équivalents d'une langue à l'autre

Prétraitement

- Pour un traitement effectif du langage, le prétraitement du texte brut est une étape importante
  - segmentation des mots (tokenisation)
  - detection des phrases
  - normalisation

Tokenisation

- Les textes sont représentés comme chaîne de caractères
- Les mots, les espaces et la ponctuation sont encodés de la même façon
- Il faut définir et marquer les limites de mots: **tokenisation**

Qu’est-ce qu’un mot ?

une suite de caractères séparées par un blanc (espace, ligne) et/ou de la ponctuation ?

- apostrophes ? *l’arrêt d’activité, aujourd’hui*

Qu'est-ce qu'un mot ?

une suite de caractères séparées par un blanc (espace, ligne) et/ou de la ponctuation ?

- apostrophes ? *l'arrêt d'activité, aujourd'hui*
- tirets ? *New-York; est-il là ? video-projecteur*

Qu'est-ce qu'un mot ?

une suite de caractères séparées par un blanc (espace, ligne) et/ou de la ponctuation ?

- apostrophes ? *l'arrêt d'activité, aujourd'hui*
- tirets ? *New-York; est-il là ? video-projecteur*
- et s'il n'y a pas de séparateur ?
  - mots composés en allemand
  - écriture en mandarin: suite de 'caractères' sans séparateur
  - langues agglutinantes: turque, basque, langues inuites

Segmentation des phrases

- Certaines signes de ponctuation sont non ambigu (‘?’, ‘!’)
- Le point (‘.’) est très ambigu
  - *En octobre 2013, M. Obama visitera Paris.*
  - *De nombreux genres musicaux comme le motet, la cantate, etc. se développeront au cours des siècles suivants.*
- Le contexte est important pour déterminer les fins de phrases

Combien de mots pour neige en français ?

- neige
- poudreuse
- soupe
- slush (québecois)

et en latin ?

Combien de mots pour neige en français ?

- neige neiges
- poudreuse
- soupe
- slush (québecois)

et en latin ?

Normalisation

Lemmatisation

- capitales / minuscules : Demain/demain (mais USA vs usa)
- graphies: *clé*/*clef*, *pizzéria*/*pizzeria*
- abbréviations: etc, svp, ...
- **inflections**:
  - conjugaison est, serai, étaient, suis, es : **être**
  - nombre/genre: associatives : **associatif**
- **lemme** $\approx$ entrée du dictionnaire
  lemmatisation : réduction à la forme de base
- mais déjà des ambiguités :
  - étais : être/étayer
  - suis : être/suivre

  $\rightarrow$ *morphologie*