Traitement Automatique des Langues Naturelles

Cours 1: Introduction

Chloé Braud, Philippe Muller

Master IAFA 2024-2025

Présentation

Semestre 9

- Cours/TD, 22 heures
- TD en partie avec machine
- TP, 6h ; 2 groupes
- une Master class 2h

Modalités de Contrôle des Connaissances

- CT : 70%, examen écrit de deux heures
- CCTP : 30%, note : devoir et/ou compte-rendus de TP, règle "16" : ABI $\rightarrow$ note 0, ABJ $\rightarrow$ coef 0

Pourquoi un cours de TAL ?

Bonne question, demandons à un modèle de langue:

`https://gemini.google.com/app/883b552d8343624d`

Objectifs généraux

- connaitre les enjeux, les questions, et les grandes applications du TAL
- se familiariser avec les aspects spécifiques du langage naturel
- savoir choisir le modèle adapté à un problème donné
- savoir mettre en place une méthodologie expérimentale
- aborder le domaine et ses utilisations avec un regard critique

Plan du cours

1. introduction
2. à la recherche du mot; initiation à l'extraction de connaissances
3. l'importance de la syntaxe du langage
4. la sémantique : lexicale, formelle, distributionnelle
5. grandes architectures de modèles et applications
6. l'ère des modèles préentrainés et du transfert
7. le TAL contextuel: modélisation de documents et du dialogue ("discours")

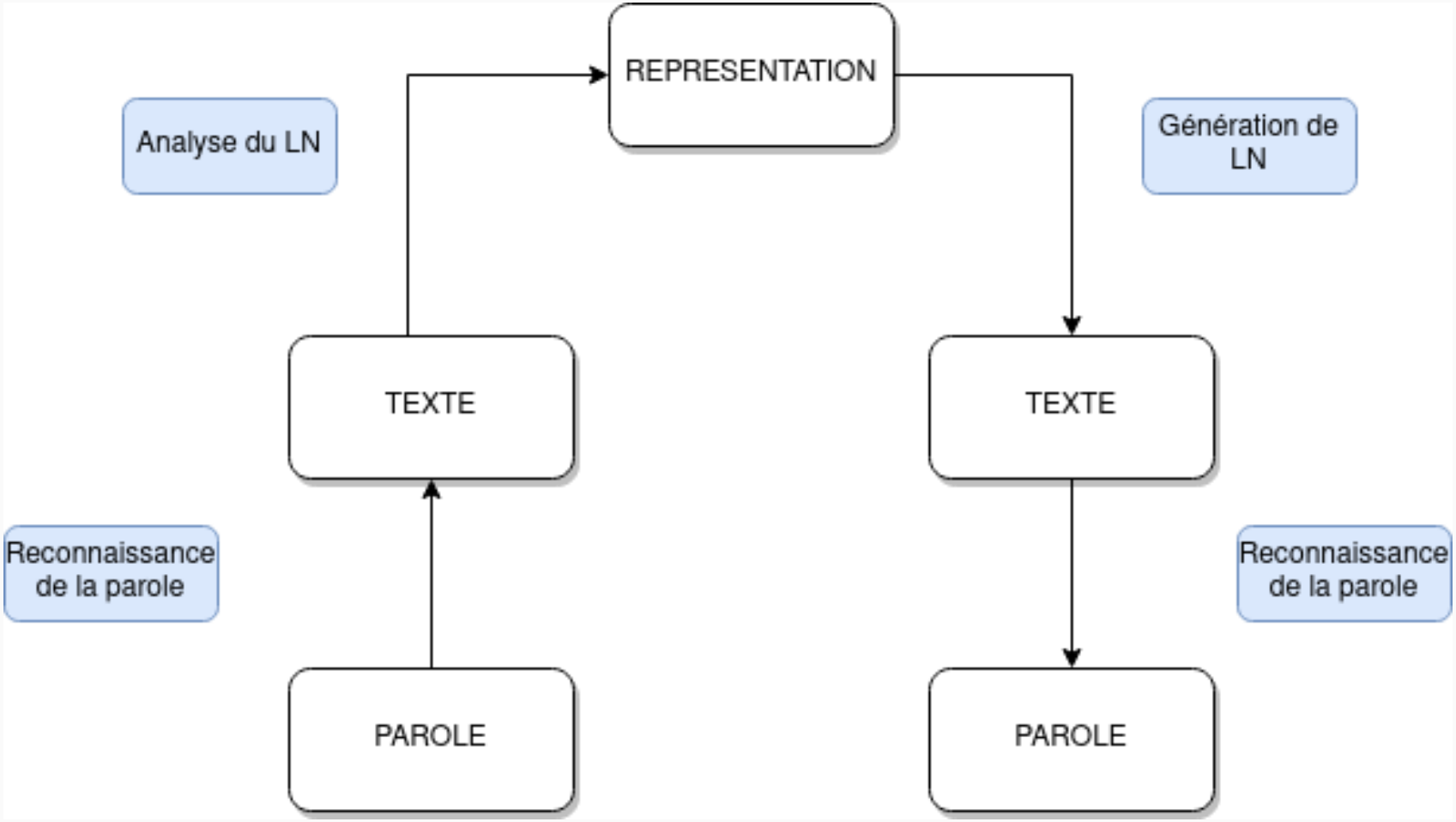
Plan du jour

- introduction au domaine: grandes questions, applications, difficultés
- une question centrale : quelles sont les bonnes unités d'analyse d'un texte

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

Les technologies de la langue



Pourquoi ?

- IA et représentation de connaissances :

  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
- génération de textes
- système de dialogue
- construction de bases de connaissances / ontologies

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

Quelle(s) discipline(s)

- intelligence artificielle / apprentissage automatique
- 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

$200

If you're standing, it's the direction you should look to check out the wainscoting.

$1000

The first person mentioned by name in 'The Man in the Iron Mask' is this hero of a previous book by the same author.

$600

In cell division, mitosis splits the nucleus & cytokinesis splits this liquid *cushioning* the nucleus

$2000

Of the 4 countries in the world that the U.S. does not have diplomatic relations with, the one that's farthest north

Watson

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

combine :

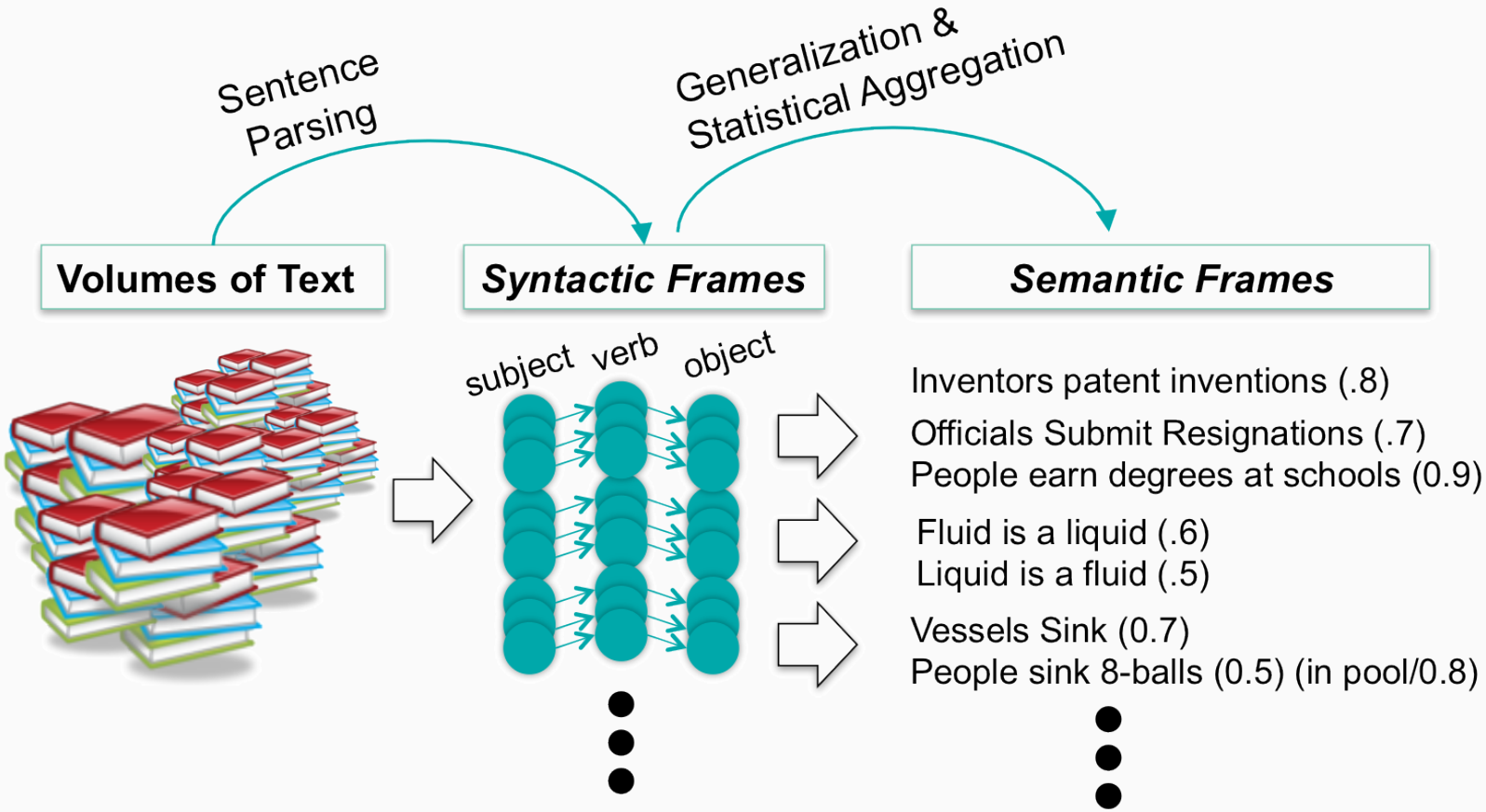
- traitement du langage naturel (analyse/production)
- apprentissage automatique
- représentation de connaissances / décision

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

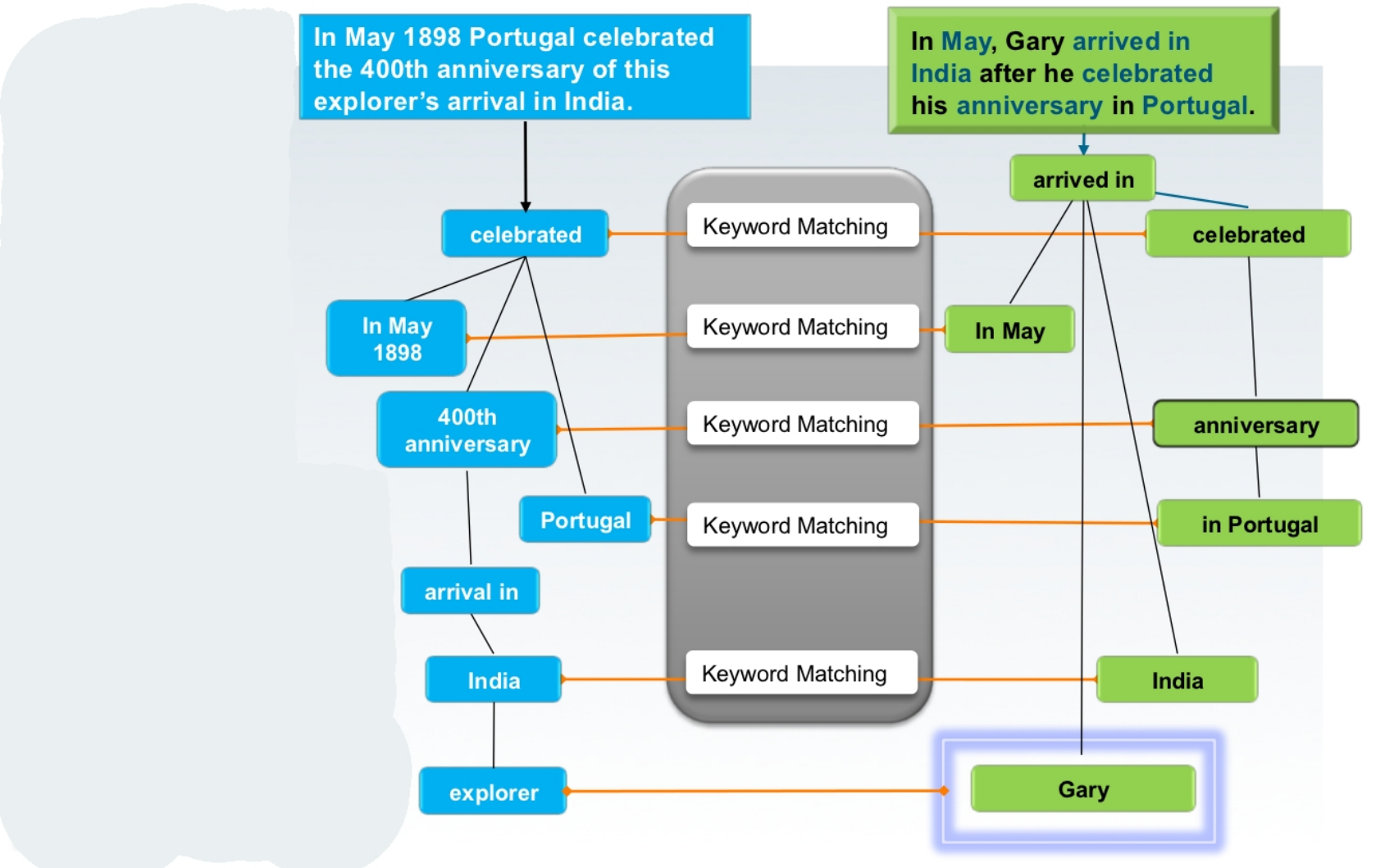
un exemple de recherche d'information sophistiquée

date de 2010 mais une partie de son fonctionnement est encore typique de certaines applications

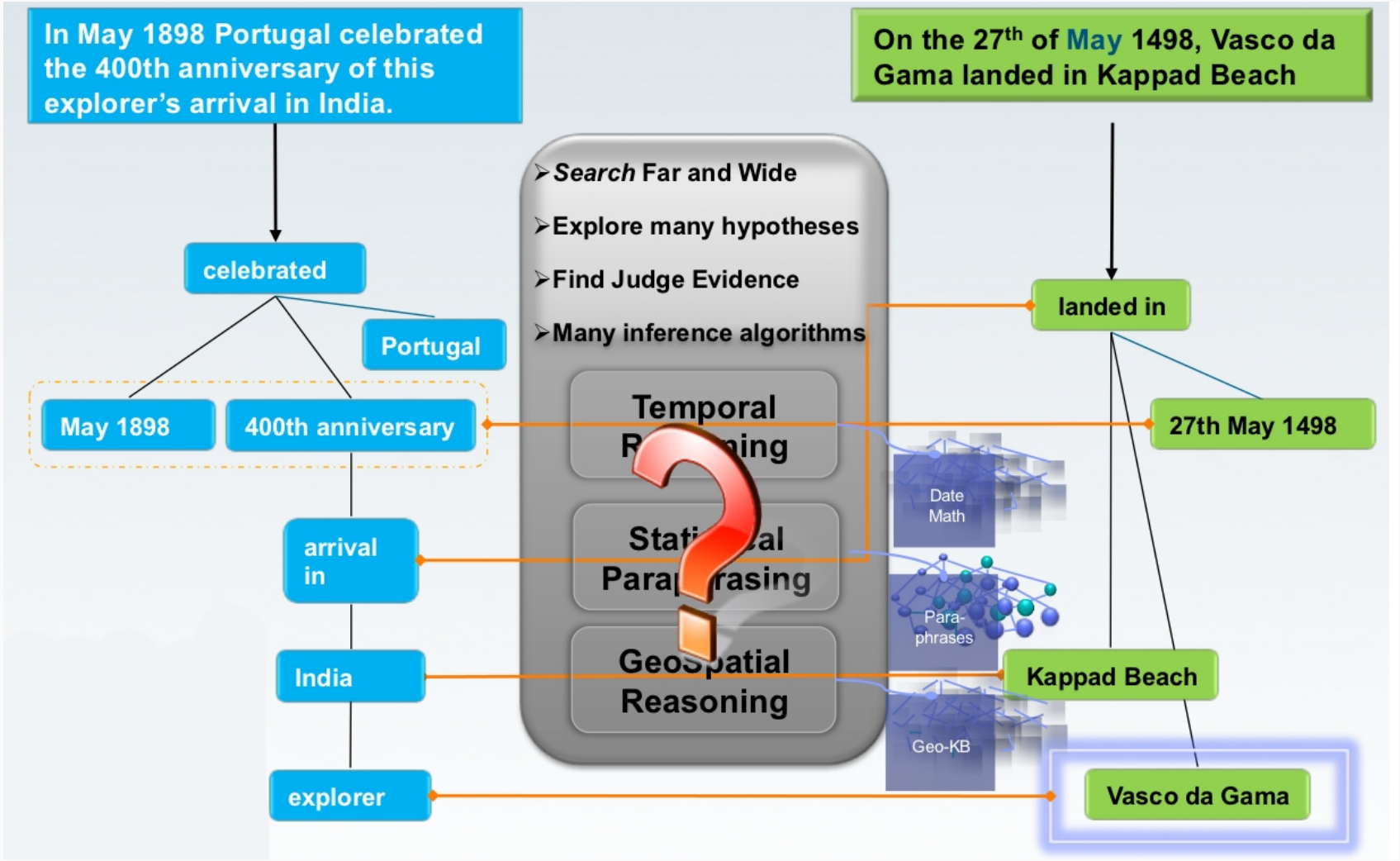
Watson construit ses connaissances à partir de texte



Les mots clefs ne suffisent pas



Analyse sémantique, raisonnement, décision



Exemple: Analyse d'opinion

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

source : xkcd; allociné

Exemple: Analyse d'opinion

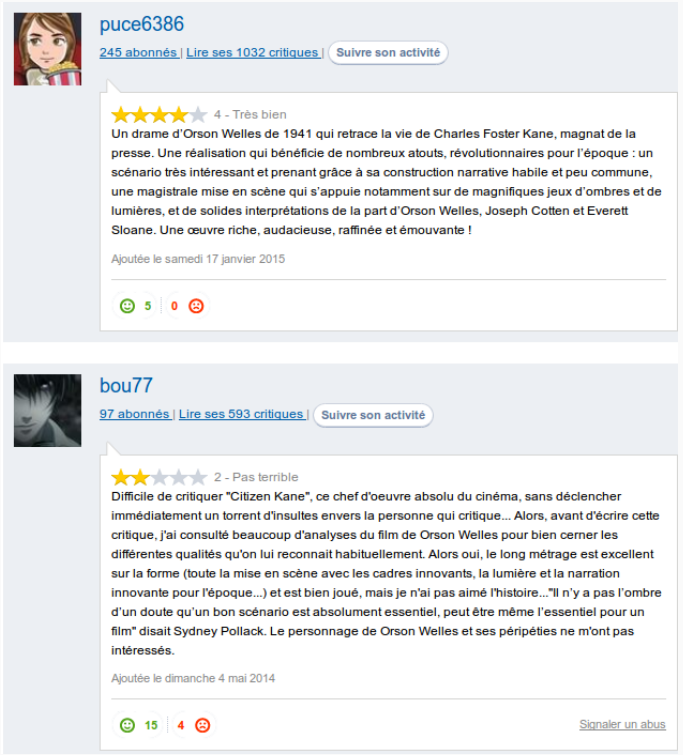
- suivi d'opinion sur un produit
- aggrégation d'avis
- analyse fine pour recommendations

source : xkcd; allociné

Exemple: Analyse d'opinion

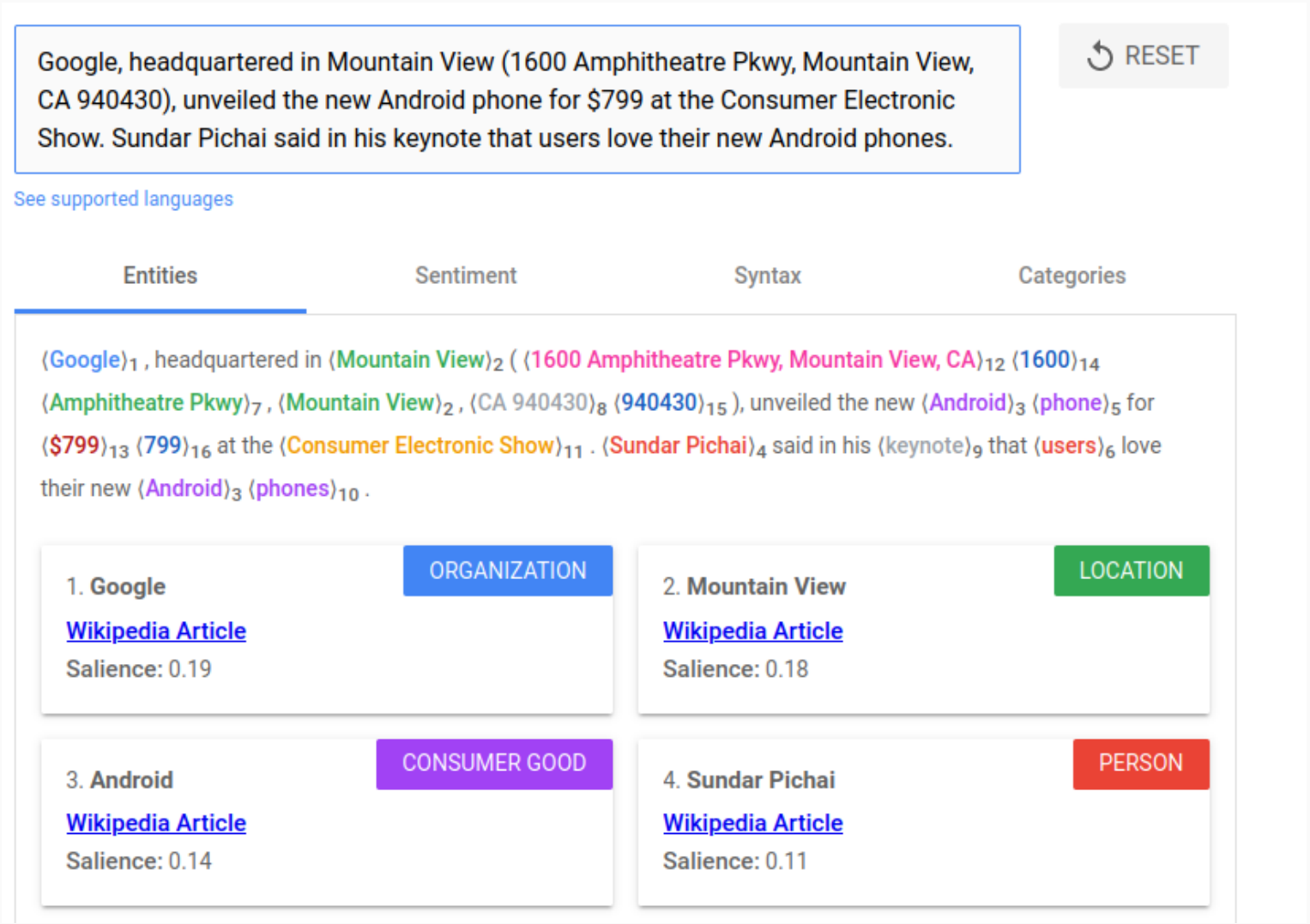
- suivi d'opinion sur un produit
- aggrégation d'avis
- analyse fine pour recommendations

versus



source : xkcd; allociné

Exemple : Extraction d'information

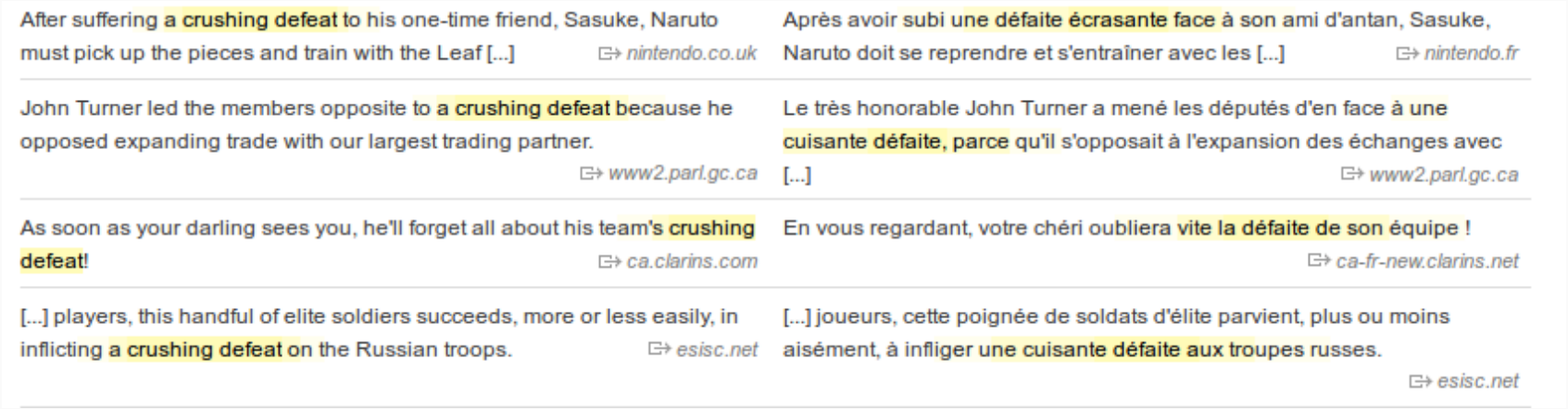


→API google

→API IBM

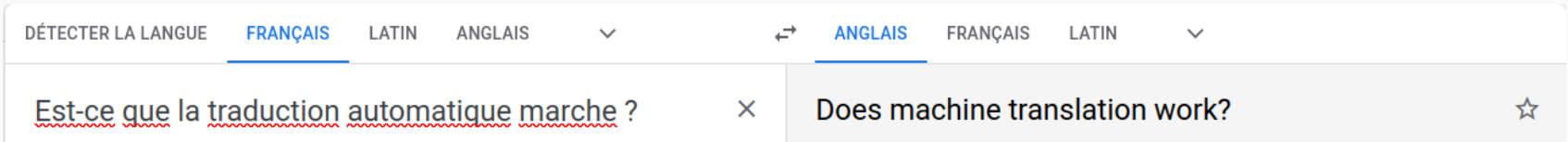
Exemple : Traduction automatique

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

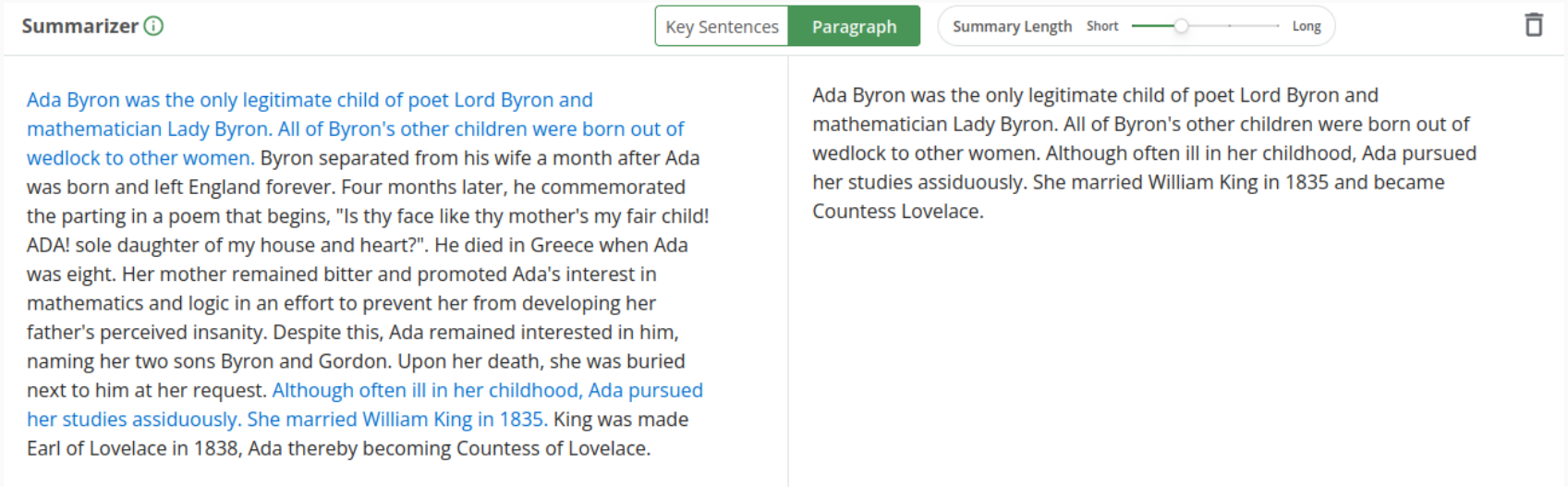
Traduction automatique google translate/DeepL



Exemple : Résumés de textes

- synthèse de documents
- simplification
- résumé extractif ou abstractif

Exemple: Quillbot



One model to rule them all ?

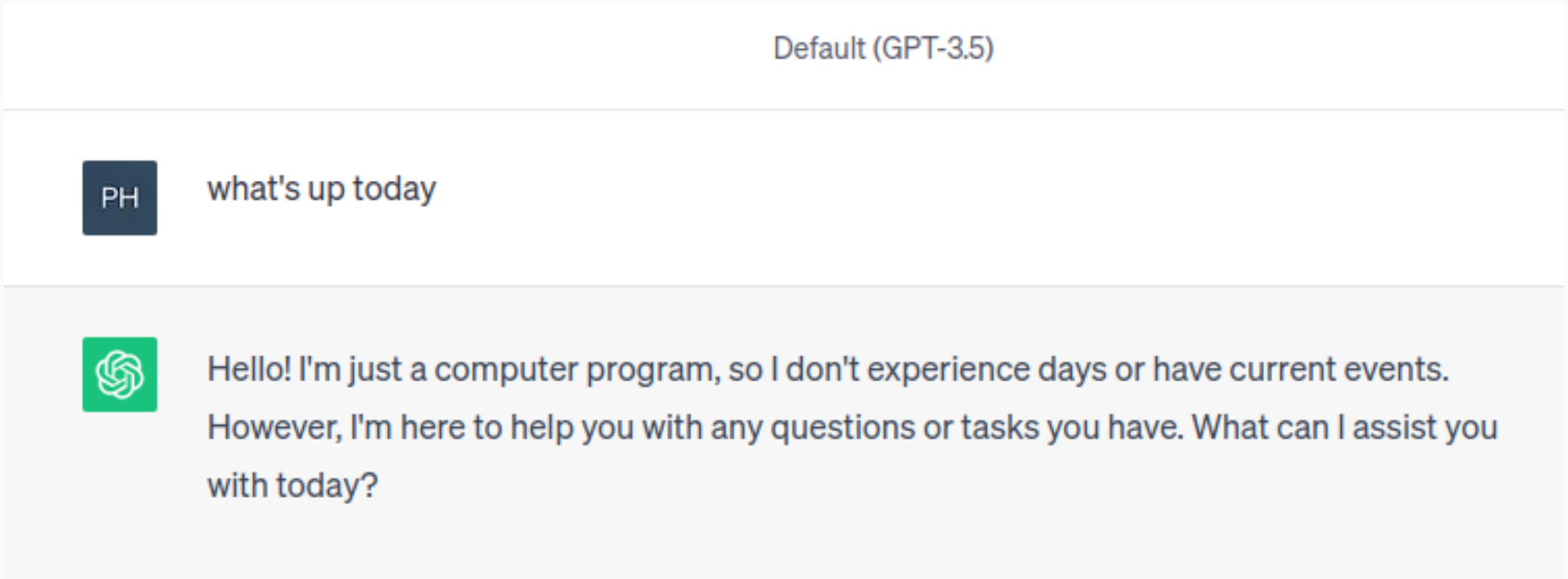
Default (GPT-3.5)

PH

what's up today

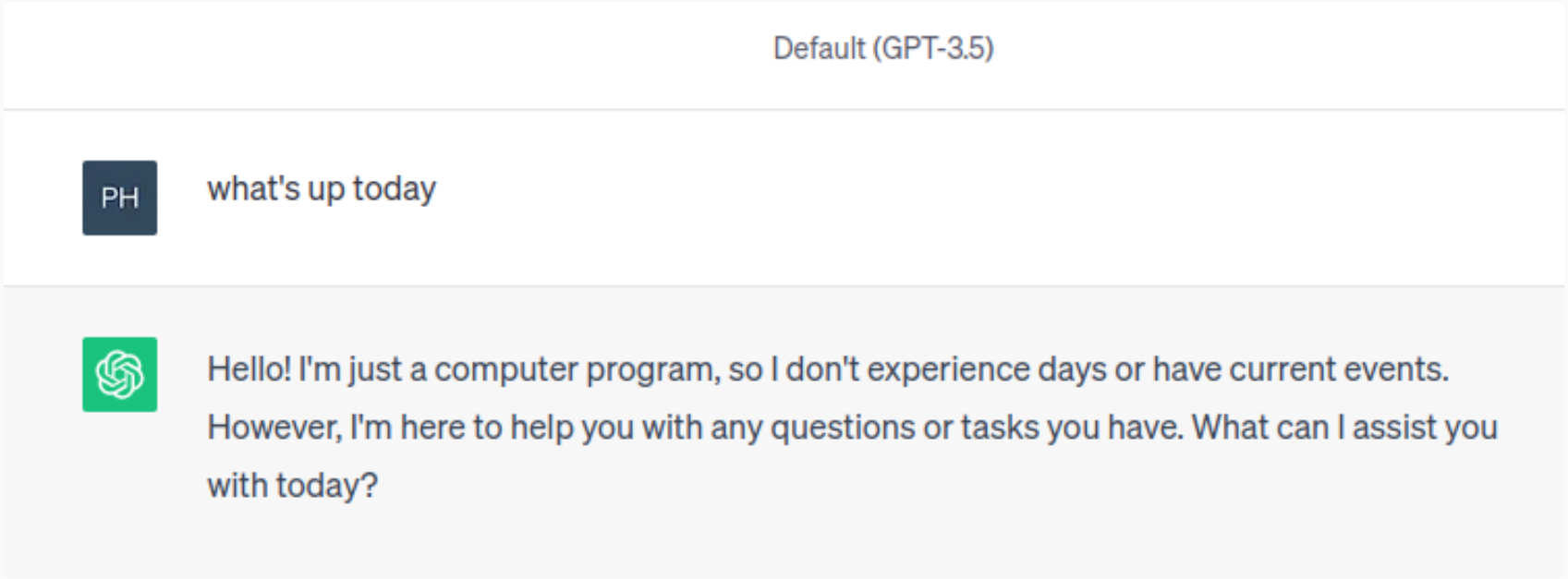
Hello! I'm just a computer program, so I don't experience days or have current events. However, I'm here to help you with any questions or tasks you have. What can I assist you with today?

One model to rule them all ?



En fait des dizaines de modèles similaires: Gemini, Llama, Mistral, Bloom, Flan, etc

One model to rule them all ?



On verra aussi comment ces modèles sont construits ... et pourquoi ça ne règle pas tout

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

Pourquoi c'est dur ?

on parle souvent d'information "non structurée" $\rightarrow$ plutôt semi-structurée

- si on veut partir de données textuelles pour obtenir une représentation

  L'aspirine traite le mal de tête.
  $\rightarrow$treatment(D001241,D006261)

$\rightarrow$ la représentation ciblée doit être "normalisée"

Pourquoi c'est dur ?

on parle souvent d'information "non structurée" $\rightarrow$ plutôt semi-structurée

- si on veut partir de données textuelles pour obtenir une représentation

  L'aspirine traite le mal de tête.
  $\rightarrow$treatment(D001241,D006261)

- problème de base du TAL: formes équivalentes

  L'acide acétylsalicylique soigne les céphalées.
  $\rightarrow$ treatment(D001241,D006261)

$\rightarrow$ la représentation ciblée doit être "normalisée"

Pourquoi c'est dur ?

on parle souvent d'information "non structurée" $\rightarrow$ plutôt semi-structurée

- si on veut partir de données textuelles pour obtenir une représentation

  L'aspirine traite le mal de tête.
  $\rightarrow$treatment(D001241,D006261)

- problème de base du TAL: formes équivalentes

  L'acide acétylsalicylique soigne les céphalées.
  $\rightarrow$ treatment(D001241,D006261)

- problème de base du TAL: ambiguités

  Le médecin traite le patient de tous les noms. ??

$\rightarrow$ la représentation ciblée doit être "normalisée"

Pourquoi c'est dur ?

on parle souvent d'information "non structurée" → plutôt semi-structurée

- si on veut partir de données textuelles pour obtenir une représentation

  L'aspirine traite le mal de tête.
  →treatment(D001241,D006261)

- problème de base du TAL: formes équivalentes

  L'acide acétylsalicylique soigne les céphalées.
  → treatment(D001241,D006261)

- problème de base du TAL: ambiguités

  Le médecin traite le patient de tous les noms. ??

- problème de base du TAL: rôle de l'implicite, de l'inférence

  Le café est un meilleur médicament que le chocolat.
  Il guérit le mal de tête.

  Il = café (vs chocolat) + la phrase (2) explique la phrase (1)
  + le chocolat ne guérit pas le mal de tête;

→ la représentation ciblée doit être "normalisée"

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 en unités

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 en unités

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 en unités

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.

relation vs. propriété : 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.

relation vs. propriété : connaissance du monde

Marie et Jean sont de bons amis

de qui ? problème du 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 (document, conversation, réunion)

Principales couches d'analyse

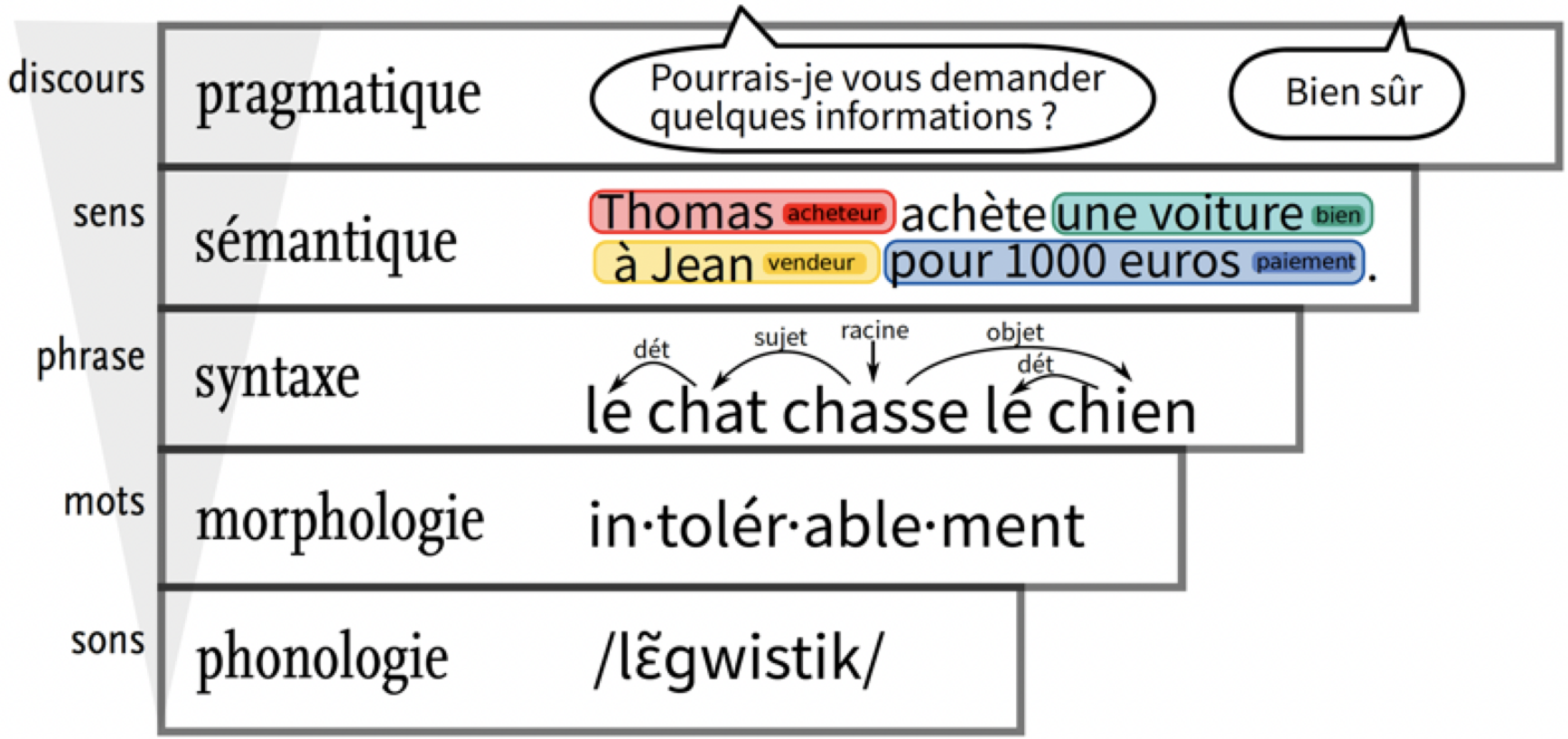


Figure réalisée par Tim Van De Cruys

Exemples d'ambiguité à tous les niveaux

Phonologie

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

Exemples d'ambiguité à tous les niveaux

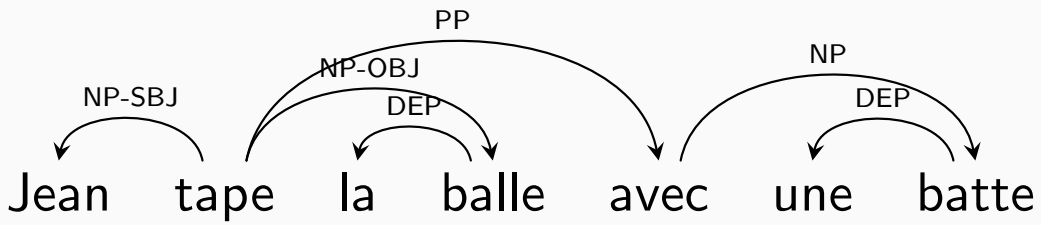
Morphologie

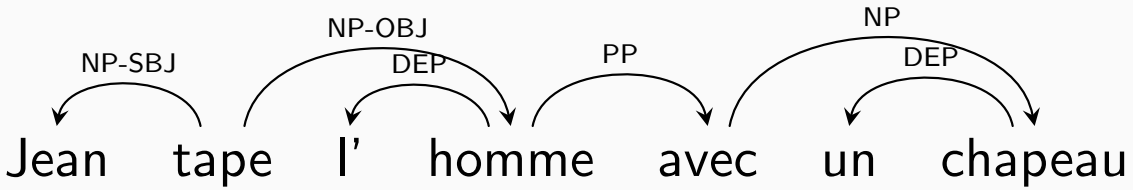
- homonymes/homographes :
  - car (conjonction) / car (nom)
  - brise (nom) / brise (verbe)
  - voler (dans le ciel) / voler (la banque)
- complique l'analyse syntaxique
- complique l'analyse sémantique
- peut se combiner pour multiplier le problème:
  La petite brise la glace

Exemples d'ambiguité à tous les niveaux

Ambiguité syntaxique

Ambiguité de structure





Exemples d'ambiguité à tous les niveaux

Ambiguité lexicale

au delà des homonyes

- La tour Eiffel est une construction solide.
  La construction a pris plus de temps que prévu
- La voiture fauche un piéton.
  Certaines parcelles sont fauchées tardivement l'été.

Exemples d'ambiguité à tous les niveaux

Ambiguité sémantique

- Aucun étudiant n'a réussi son examen.
- Toutes les 10 minutes, un homme se fait renverser par une voiture à Paris.

Exemples d'ambiguité à tous les niveaux

Ambiguité sémantique

- Aucun étudiant n'a réussi son examen.
- Toutes les 10 minutes, un homme se fait renverser par une voiture à Paris.

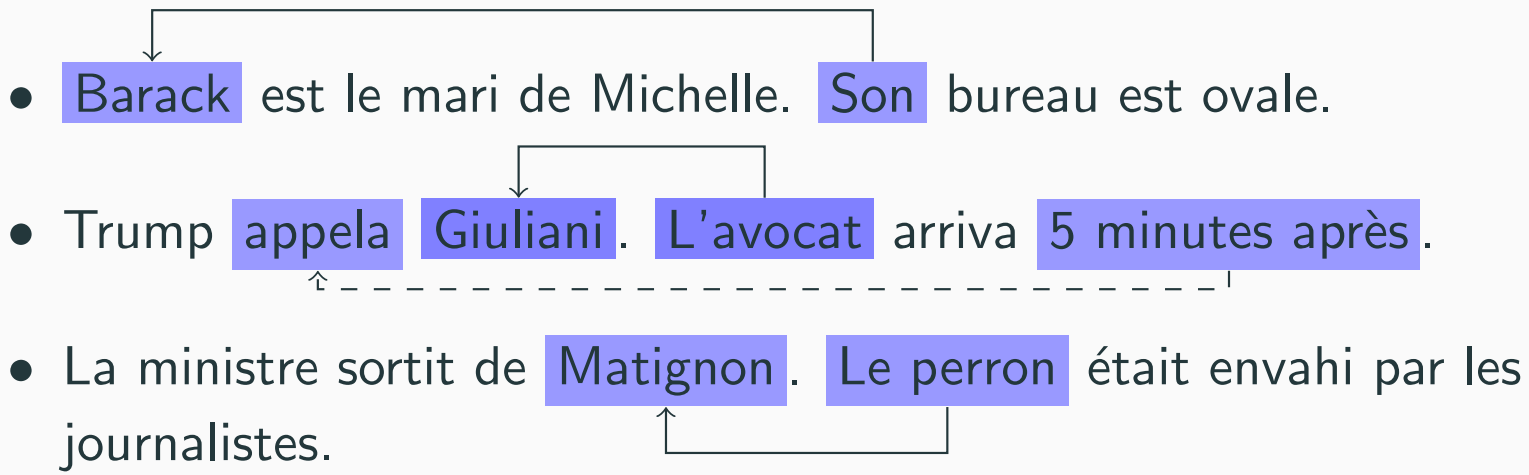
Exemples d'ambiguité à tous les niveaux

Ambiguité sémantique

- Aucun étudiant n'a réussi son examen.
- Toutes les 10 minutes, un homme se fait renverser par une voiture à Paris. Il commence à en avoir assez.

Exemples d'ambiguité à tous les niveaux

Pragmatique



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)

Comment ?

- connaissances sur le langage
- connaissances “du monde”
- combinaison des 2
- **importance des probabilités** : le TAL est un champ important d’application du Machine Learning

Ce qu’on verra dans ce cours

- Des modèles de différents aspects du langage (depuis le mot jusqu’au texte)
- Des techniques appliquées au TAL
- Des applications au fur et à mesure, certaines plus en détail: résumé, extraction de connaissances
- La méthodologie de recherche
- Les limites actuelles du domaine

Le TAL, une discipline expérimentale

- une science des données, dont beaucoup sont ouvertes
- beaucoup d'outils existant en open-source
- $\rightarrow$ ce cours aura une dimension pratique, **même en cours**

Idéalement, vous devriez disposer d'une machine et avoir accès à

- un "vrai" terminal : Linux/MacOs ou bien installer Cygwin sous Windows
- python avec outils scientifiques (par ex via Anaconda) : numpy, pandas,
- librairies ML : scikit-learn, pytorch, transformers
- librairies TAL : NLTK, Spacy et/ou Stanza

Une discipline empirique

- les modèles sont développés à partir de données linguistiques, observées ou bien créées
  - "corpus" / "dataset" : quelle représentativité ?
  - annotations humaines nécessaires pour de nombreux aspects, et pour l'évaluation
- les résultats sont évalués expérimentalement: problèmes de mesure du succès
- nécessité de l'analyse qualitative des résultats
- la diversité des langues du monde complique la possibilité d'approches "universelles" : attention aux pièges des "langues majoritaires" ou dominantes pour des raisons culturelles, économiques ou politiques.
  $\rightarrow$ The world Atlas of language structures
  `https://wals.info/`

Une discipline avec des impacts sociétaux importants

- une forte empreinte carbone cf **How to shrink AI's ballooning carbon footprint**
- des enjeux éthiques
  - exclusion de certaines populations
  - accentuation de biais sociaux
  - usages détournés : surveillance, données personnelles, ciblage publicitaire ou politique

  cf
  - **The Social Impact of Natural Language Processing**
  - **Cartography of Natural Language Processing for Social Good**