Analyse et représentation de la langue

Damien Nouvel

Plan

Correction orthographique

- **Recherche dans une base lexicale**
  - Liste des mots pour une langue donnée
  - Comparaison mots du texte vs lexique
  - Disposer de toutes les formes possibles : **couverture**

  $\Rightarrow$ Complexe : par exemple $1000 * 50K = 50M$ ...
- Optimisation de la recherche
  - Utilisation d'automates
  - Arbre des préfixes (TRIE)
- Besoin d'informations contextuelles

Couverture du lexique

- Difficultés de couverture
  - Langues « agglutinantes » (morphologie / composition)
  - Morphologie, dont **flexions** (déclinaisons / conjugaisons)

  ⇒ **Paradigmes** d'agglutination et de flexion
  - Variantes d'écritures
  - Noms propres (classe **ouverte**) acronymes, etc.

  ⇒ Automates de reconnaissances partielles

Applications

- Éditeurs de texte
- Saisie prédictive / autocorrect (T9)
  - Handicap (aveugle, moteurs)
  - Grand public

⇒ Logiciels **hors-ligne** (contrairement aux reconnaissances)

- Détection de nouveaux mots (néologismes, noms propres, etc.)

Plan

Lier les mots dans les énoncés

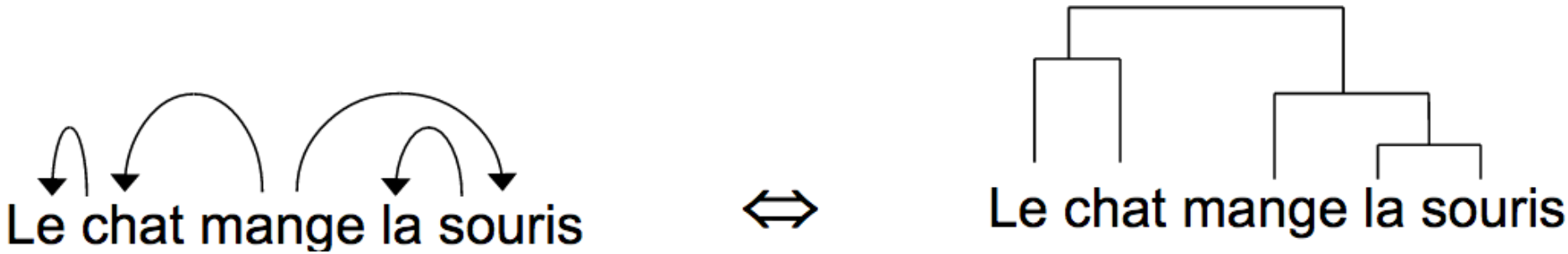
- Niveau de représentation
  - Exploite les approches morphologiques / lexicales

  ⇒ Importance des **catégories morpho-syntaxiques** (POS)
  - Détermine les liens / relations / **dépendances** entre les mots
  - Identifie des **syntagmes**
- Modélisation de la « grammaire » (SOV, VSO, etc.)
- Deux catégories d'approches
  - **Constituants** : arbres syntaxiques (projectivité)
  - **Dépendances** : relations entre mots (Tesnière, 1940)

  ⇒ Possiblité de convertir de l'un à l'autre

  ⇒ Approches : LFG, GPSG, TAG, HPSG …

Exemple



(Bourigault / SYNTEX)

Quelques difficultés

- Le rattachement prépositionnel
  - « Il voit le chat de sa sœur »
  - « Il voit le chat de sa fenêtre »
- Triple ambiguïté locale
  - « Il voit l'homme dans le parc avec des jumelles »
- Les propositions (relatives, subordonnées, etc.)
- Quelle représentation pour la coordination
- Traitement des anaphores (syntaxe et sémantique)

Démo

- Test avec **Stanford Core NLP** (Manning)
  - Analyseur : `http://corenlp.run`
  - Documentation : `https://stanfordnlp.github.io/CoreNLP/`
- Quelques essais avec **FRMG** (de la Clergerie)
  - Analyseur : `http://alpage.inria.fr/frmgwiki/frmg_main/frmg_server`
  - Portail : `http://alpage.inria.fr/frmgwiki/`

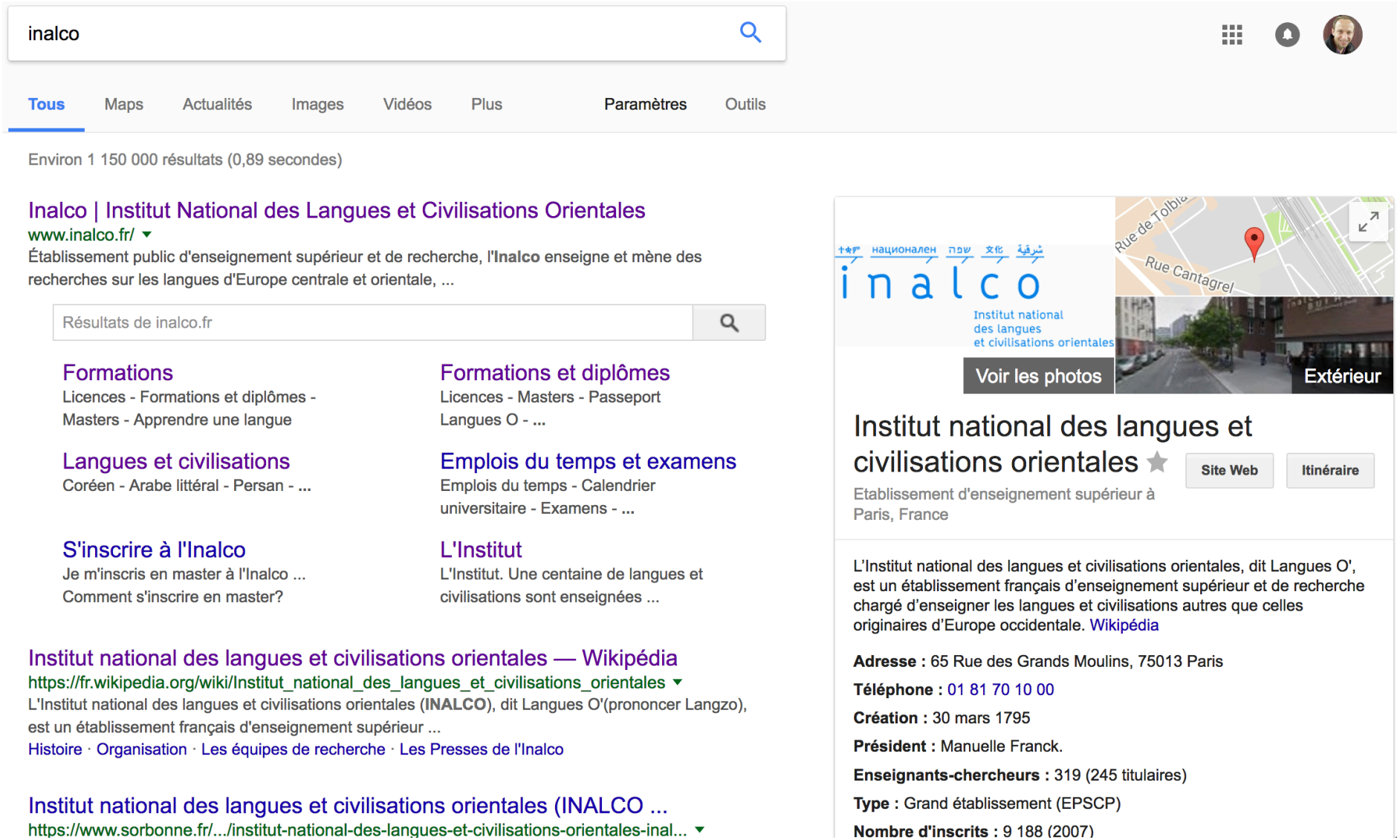
Plan

Profondeur de la syntaxe

- Analyse syntaxique
  - Morpho-syntaxe
  - Syntaxe de surface / parenthésage (syntagmes)
  - Syntaxe « grammaticale »
- Syntaxe « profonde »
  - Sujet du verbe *vs* agent de l'action (passif)
  - Objet : patient, récepteur, etc.
  - Sujet distant (anaphores / coréférences ...)

$\Rightarrow$ Interface syntaxe / sémantique

Les entités nommées



Les entités nommées

- Pas de définition stable
  - « Noms propres et quantités d'intérêt » (MUC)
  - « Entités du monde concret qui ont un nom » (ESTER)
  - « Expression qui réfère à une entité unique » (Ehrmann)
  - « Objets mentaux pour la logique » (Nouvel)
  - …
- Deux grandes catégories linguistiques
  - **Noms propres**

  ⇒ Limites : « Superman », « l'Histoire », etc.
  - **Descriptions définies** (syntagmes nominaux définis)

  ⇒ Limites : « cette voiture », « le roi des forains », « l'occident »

⇒ Tenir compte des **synonymes**, **homonymes**, **métonymies**

⇒ Beaucoup d'applications …prétraitement ?

Structurer les textes en connaissances

- Des **formats** plus ou moins **structurés**
  - Texte brut : sauts de lignes
  - PDF : graphiquement stable, mais peu structuré
  - HTML : sections, styles, liens, micro-formats
  - Wikipedia : articles et info-boxes
  - BabelNet : `http://babelnet.org/synset?word=Inalco&lang=FR`

⇒ Structuration : du texte aux **graphes** (valués)

- Quelques **réseaux sémantiques**
  - WordNet
  - Graphes conceptuels
  - Web sémantique (RDF/OWL)
  - JeuxDeMots `http://www.jeuxdemots.org`

⇒ Interactions **homme-machine** et **machine-machine**

Plongements de mots

- Sémantique distributionnelle
  - « You shall know a word by the company it keeps » (Firth, 57)
  - Analyse « distributionnelle » des mots en contexte (Harris, 60)
- Mélange de deux axes d'analyse
  - **Syntagmatique** : les successions de mots
  - **Paradigmatique** : les substitutions de mots
- Méthode non-supervisée de positionnement (sémantique)
  - Collection de corpus de texte **très volumineux**
  - Choix de **paramètres** (dimensions)
  - Recherche de **correlations entre mots et leurs contextes**

⇒ **Embeddings** (Collobert, 2011), **Word2Vec** (Mikolov, 2013), FastText (Bojanowski, 2016)

⇒ Démo : `http://embeddings.sketchengine.co.uk`