Introduction générale

Extraction 1

Damien Nouvel

Inalco

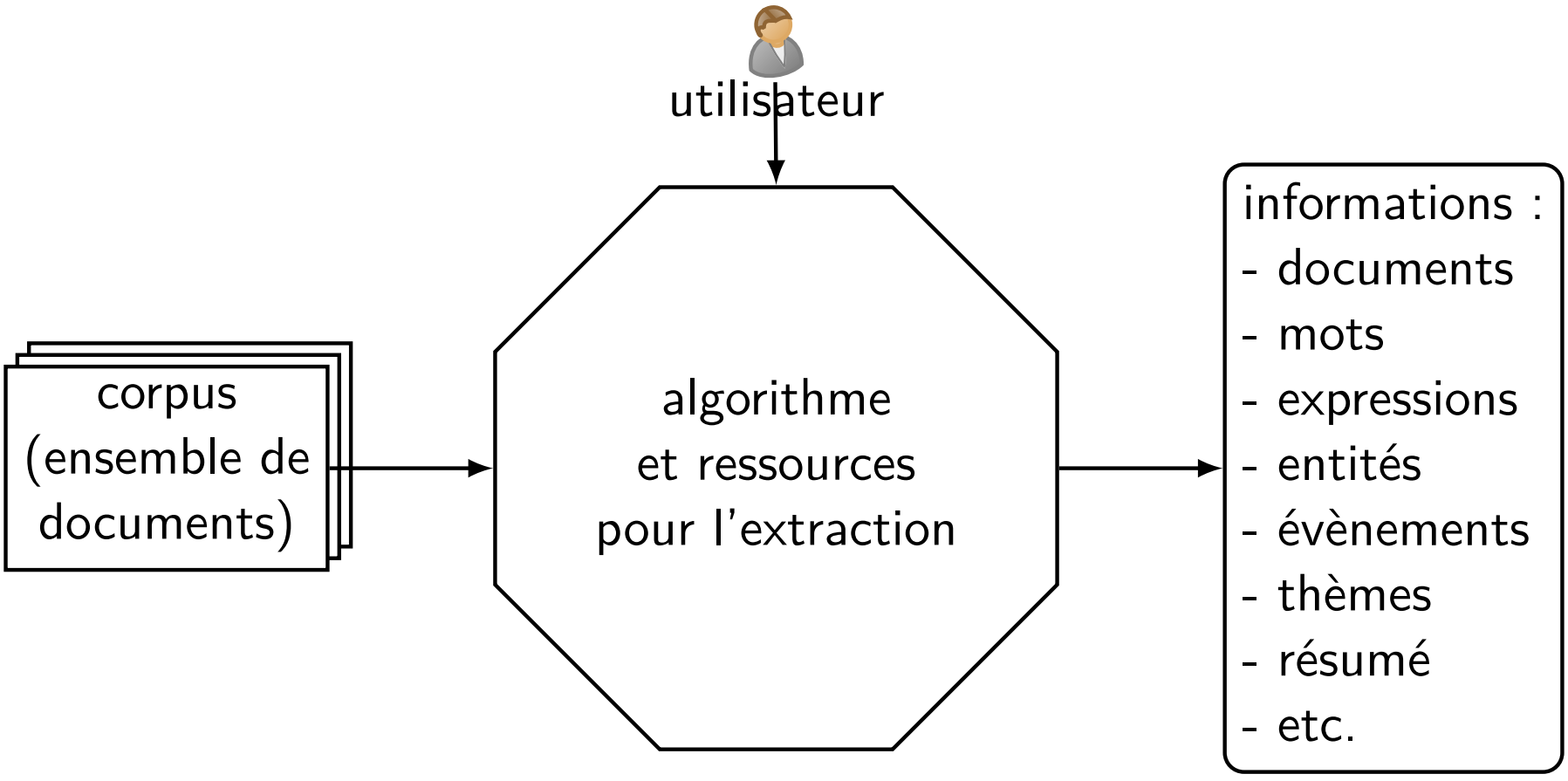
Extraction d'informations : définitions

Proposition initiale : « Méthodes permettant d'identifier et de représenter de manière explicite une ou plusieurs informations qui peuvent être présentes dans un ensemble de documents. »

Wikipedia (traduction de l'article en anglais) : « L'extraction d'informations est la tâche qui consiste à extraire automatiquement des informations structurées à partir de documents non structurés et/ou semi-structurés lisibles par machine et d'autres sources représentées électroniquement. »

ChatGPT : « L'extraction d'informations est un domaine de l'intelligence artificielle et du traitement automatique du langage naturel. Elle consiste à identifier automatiquement des données précises et structurées à partir de textes non structurés (comme des articles, des emails, des rapports, des pages web, etc.). Il s'agit de transformer un texte libre en informations organisées qu'on peut stocker et exploiter dans une base de données. »

Processus d'extraction



Notes sur la notion d'information

Notion d'**information** liées aux **connaissances**

- une information comme nouvelle connaissance (i.e. journaux)
- objets du monde réel (personnes ou objets)

Informations véhiculées par la **langue naturelle**

- message plus ou moins **informatif**
- action sur le récepteur (intégration de l'information)
- degré de besoin en **compréhension**

Quantifier et **structurer** les informations

- recherche de l'unité d'information « atomique »
- **granularité** (documents, paragraphes, phrases, mots)
- représentation **sémantique** de l'information
- problème de la redondance (entropie en théorie de l'information)

Problématique de l'extraction

Prise en compte des **attentes de l'utilisateur**

- **représentativité** des sources
- **filtrages** des contenus extraits
- **formats** de sortie
  - documents complets
  - extraits des documents (avec références)
  - représentations explicites et structurées
  - représentations encodées
  - reformulation en langue naturelle générée

Importance de la **complétude** (silence mesuré par le rappel)

Degrés de **compréhension**

- localisation de motifs linguistiques
- recherche d'entités nommées
- réponse à des questions
- résumé de textes

Données et applications

Éléments exploités pour l'extraction (en plus des sources)

- algorithmes
- lexiques (langue, entités)
- corpus de référence
- modèles de langue

Quelques exemples d'**applications**

- moteurs de recherche (requêtes)
- extraction d'évènements
- alimentation de base de données
- génération de résumés
- analyses linguistiques

Plan du cours

- Introduction générale
- Fouille lexicale
- Annotation de documents