Université François-Rabelais Tours

Laboratoire d'Informatique
EA 6300

cnrs
dépasser les frontières

Istex-Entités nommées

Denis MAUREL, Anubhav GUPTA
Université François-Rabelais de Tours
Enza MORALE, Patrice RINGOT
Inist-CNRS, Vandœuvre-lès-Nancy
Gilles VOLLANT
Société ERGONOTICS, Lille

INTRODUCTION

Enrichir la base Istex

Enrichir la base Istex

- "le projet Istex entend offrir tous les moyens accessibles de consultation"
- Pour répondre à ce défi, nous souhaitons *fouiller* différemment cette immense base de connaissance
  - Les requêtes classiques permettent l'accès au document,
    - soit à partir du titre, des auteurs ou des mots-clés
    - soit en texte plein

Enrichir la base Istex

- "le projet Istex entend offrir tous les moyens accessibles de consultation"
- Pour répondre à ce défi, nous souhaitons *fouiller* différemment cette immense base de connaissance
  - Les requêtes classiques permettent l'accès au document,
    - soit à partir du titre, des auteurs ou des mots-clés (trop restreint)
    - soit en texte plein (trop large)

Enrichir la base Istex

- "le projet Istex entend offrir tous les moyens accessibles de consultation"
- Pour répondre à ce défi, nous souhaitons *fouiller* différemment cette immense base de connaissance
  - Une valeur ajoutée
    - interroger la base via les entités nommées
      - noms propres
      - dates
      - références

Les différentes tâches

Formelles	Analytiques	Logicielles
• Analyse • Définition • Préparation • Sortie	• Unitex • CasSys • Cascades • Évaluation	• Développement • Passage à l'échelle

Laboratoire d'Informatique
EA 6300

LES TÂCHES FORMELLES

Analyse

Définition

Préparation

Sortie

Analyse

- La notion d'*entité nommée* est apparue en 1996 dans le cadre des conférences américaines Muc pour l'évaluation de la recherche d'information
- Depuis chaque conférence d'évaluation ou chaque projet a défini ses propres objets à identifier

Analyse

- Quelques remarques
  1. Les noms d'universités, de centres de recherche, de laboratoires ne figurent pas dans les mots-clés, même si les affiliations des auteurs sont dans les signatures
  2. De même pour les noms de projets qui apparaissent parfois en note ou en remerciement

Analyse

- Quelques remarques
  3. Le lieu où est réalisé une expérience n'est pas forcément l'adresse du laboratoire
  4. Les dates des expériences ne correspondent pas à celle de parution de l'article

Analyse

- Quelques remarques
  5. Les noms de chercheurs cités ont une importance, alors que souvent la bibliographie indique plusieurs personnes comme signataires d'un article
  6. En SHS, des lieux, des institutions, des personnes (avec leur titre ou profession) ou des dates sont cités, indépendamment du rattachement des auteurs

Définition

- Dix entités choisies (balises TEI):
  - personnes
    - <persName>
  - lieux
    - administratifs: <placeName>
    - géographiques: <geogName>
  - organisations
    - <orgName>
    - financeurs / projets <orgName type="funder">
    - hébergeurs de ressources <orgName type="provider">

Définition

- Dix entités choisies (balises TEI):
  - temps
    - années, décennies, siècles, millénaires: <date>
  - références
    - URL: <ref type="url">
    - citations: <ref type="bibl">
    - dans le corps du texte: <bibl>

Définition

- Réalisation d' un guide d'annotation en deux parties
  - l'historique des décisions prises
  - les consignes et un grand nombre d'exemples

Définition

- Exemple (extrait de la base Istex)
  - Il a donc 30 ans lorsqu'il est invité à partir en <placeName>Hongrie</placeName>. Le baron <persName>Podmanicky</persName>, ambassadeur à <placeName>Paris</placeName>, lui fournit de multiples informations.

Définition

- Exemple (extrait de la base Istex)
  - Il naît à <placeName>Paris</placeName> le 5 septembre <date>1787</date>, d'une famille paternelle venue de l'<geogName>Ardenne</geogName>.

Définition

- Exemple (extrait de la base Istex)
  - Cette étude a été réalisée grâce à l'aide d'<orgName type="funder">AGIRA</orgName> (<orgName type="funder">Alsace Gérontologie Information Recherche</orgName>) et des médecins de la <orgName>Société de gérontologie de l'Est</orgName>…

Définition

- Interrogation de données typées
  - Washington
  - <persName>Washington</persName>
  - <placeName>Washington</placeName>

Préparation

- Préparation
  - Différents types de textes
    - Des textes récents dans différents formats XML (Elsevier, OUP, IOP, Nature, RSC)
      - analyse de ces formats
      - utilisation des balises existantes dans notre stratégie de reconnaissance des entités nommées

Préparation

- Préparation
  - Différents types de textes
    - Des textes récents dans différents formats XML (Elsevier, OUP, IOP, Nature, RSC)
      - In particular, limpets are renowned for the adhesive strength they achieve using mucus as a glue (<xref rid="I1540-7063-042-06-1164-GRENON3">Grenon and Walker, 1981</xref>; <xref rid="I1540-7063-042-06-1164-SMITH5">Smith <italic>et al.</italic>, 1999<italic>a</italic></xref>).

Préparation

- Préparation
  - Différents types de textes
    - Des textes récents dans différents formats XML (Elsevier, OUP, IOP, Nature, RSC)
      - In particular, limpets are renowned for the adhesive strength they achieve using mucus as a glue (<ref type="bibl">Grenon and Walker, 1981</ref>; <ref type="bibl">Smith et al., 1999a</ref>).

Préparation

- Préparation
  - Différents types de textes
    - Dans des textes plus anciens (PDF)
      - extraction du texte par l'Inist sous un format TEI
      - stratégies de délimitation du texte
        - récupération du corps du texte
        - isolement des entêtes, des pieds de page, des légendes, des tableaux
        - standardisation des espaces, tirets, apostrophes…

Sortie

- Annotation du texte lui-même
  - Comment combiner des annotations différentes provenant des différents services à valeur ajoutée ?
  - Comment stoker et gérer plusieurs versions du texte de l'article sur la plateforme et combiner les recherches des internautes ?

Sortie

- Déportation des annotations
  - Création d'un fichier contenant
    - les mots clés typés
    - leur nombre d'occurrences
  - Au format TEI "standOff"

Sortie

```
<standOff>
  <teiHeader>
    <sourceDesc> … </sourceDesc>
    <encodingDesc> … </encodingDesc>
    <fileDesc> … </fileDesc>
    <revisionDesc> … </revisionDesc>
  </teiHeader>
  <listAnnotation type=placeName xml:lang="en">
    <annotationBlock corresp="text" xmls="https://wwww.tei-c.org/ns/1.0">
      <placeName change="#Unitex-3.2.0-alpha" resp="istex-rd"
                                    scheme="http://placename-entity.lod.istex.fr">
        <term>Finland</term>
        <fs type="statistics">
          <f name="frequency">
            <numeric value="2"/>
          </f>
        </fs>
      </placename>
    <annotationBlock>
  </listAnnotation>
          …
</standOff>
```

Laboratoire d'Informatique
EA 6300

LES TÂCHES ANALYTIQUES

Unitex

CasSys

Cascades

Évaluation

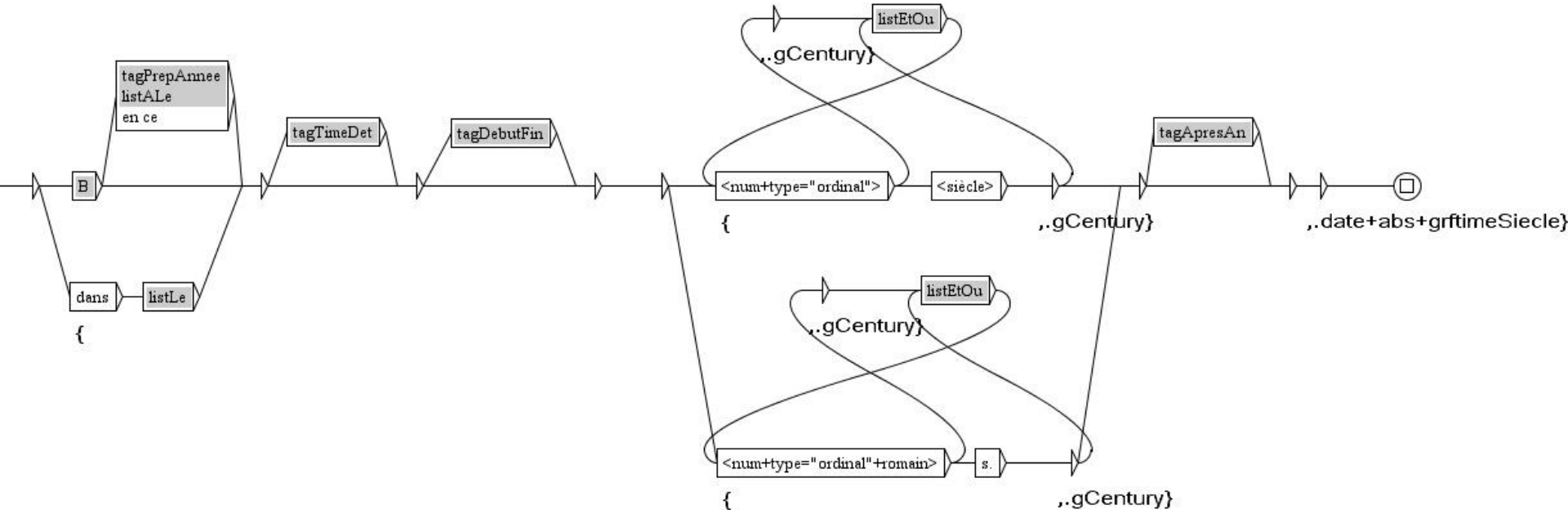
Unitex

- Unitex est un logiciel libre d'analyse lexicale automatique
- Unitex allie un système informatique performant
  - *des réseaux de transitions "augmentées"*
    - *opérations sur le texte*
    - *utilisation de variables*
    - *compilation*
- et une interface conviviale
  - *des graphes*

Laboratoire d'Informatique
EA 6300

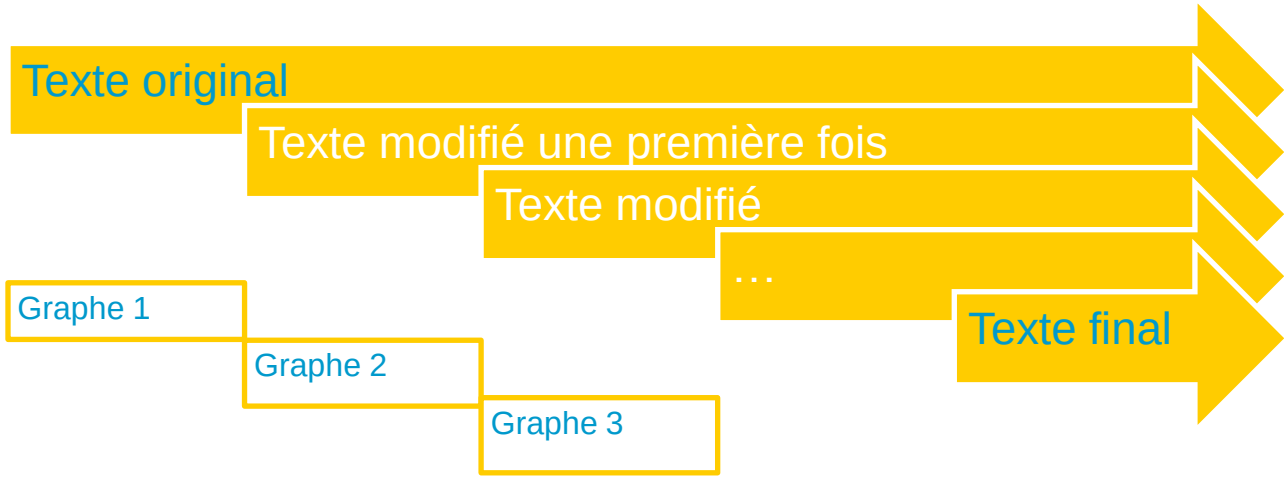
Unitex

- Exemple de graphe
  - *À peu près à la fin du XIXème siècle*



CasSys

- Un module pour la constitution et l'utilisation de cascades de graphes intégré à Unitex



CasSys

- L'ordre de passage est un paramètre important
  - Un graphe de la cascade peut
    - utiliser les motifs déjà détectés
      - Centre Georges Pompidou
      - Rue du 11 novembre 1918
    - éviter un étiquetage non souhaité pour un motif déjà reconnu

Cascades

- Organisation en 4 sous-cascades
  - Prétraitement
    - Sélection, normalisation et nettoyage
    - Application de quatre dictionnaires
  - Analyse
    - Recherche des entités nommées
  - Synthèse
    - Balisage TEI suivant le guide
  - Sortie
    - Création du fichier standOff

Cascades

- 3 cascades d'analyse
  - Anglais : 55 graphes
  - Français (1ère version) : 130 graphes
  - Français (2ème version) : 48 graphes

Évaluation

- Évaluation
  - Les deux premières cascades ont été évaluées fin 2015
    - Anglais : tests réalisés sur 49 documents contenant 5 414 entités nommées
    - Français : tests réalisés sur 40 documents contenant 4 695 entités nommées
  - La dernière le sera fin juin

Évaluation

	Anglais	Français
SER	38,6%	34,9%
Rappel	55,7%	71,5%
Précision	91,5%	87,1%

Laboratoire d'Informatique
EA 6300

LES TÂCHES LOGICIELLES

Développement

Passage à l'échelle

Tâches logicielles

- Développement
  - Normalisation et dénormalisation
  - Graphes de généralisation d'étiquetage
  - Système de débogage des cascades
  - Système de partage entre notre plateforme de développement et celle d'Istex
  - Création du fichier "standOff"

Tâches logicielles

- Passage à l'échelle
  - Optimisation du code Unitex
    - écriture d'un seul fichier en sortie
    - amélioration de la vitesse de traitement
    - scripts de lancement sur plusieurs cœurs de la plateforme Istex
  - Robustesse du passage des cascades
    - Saut d'un graphe si problème (1/1 000 000)
    - une moyenne de 0,17 s par document

Perspectives (à court terme)

Tâches analytiques

Amélioration de la cascade sur le français

Tâches logicielles

Ajout d'une fonctionnalité de priorité

Consultation étendue des dictionnaires

Laboratoire d'Informatique
EA 6300

Perspectives (à long terme)

Tâches formelles

- Affinement de la notion de date

Tâches analytiques

- Complétion des cascades
  - sur l'anglais
  - sur le français

Tâches logicielles

- Amélioration
  - du prétraitement
  - du choix des dictionnaires
  - des graphes de généralisation d'étiquetage

Laboratoire d'Informatique
EA 6300

Merci !

Évaluation

- Décomptes
  - #I = entités détectées par erreur
  - #D = entités totalement manquées
  - #T = typage incorrect
  - #E = balises mal placées
  - #S = entités détectées
  - #R = entités réelles

Évaluation

		Anglais	Français
SER	$\frac{\#I + \#D + 0{,}5 * \#T + 0{,}5 * \#E + \#TE}{\#R}$	38,6%	34,9%
Rappel	$\frac{\#S - \#I}{\#R}$	55,7%	71,5%
Précision	$\frac{\#S - \#I}{\#S}$	91,5%	87,1%
Précision du typage	$\frac{\#S - (\#I + \#T + \#TE)}{\#S}$	85,6%	84,8%
Précision du balisage	$\frac{\#S - (\#I + \#E + \#TE)}{\#S}$	79,5%	80,2%