LA RECONNAISSANCE AUTOMATIQUE DES RELATIONS ANAPHORIQUES AVEC UNITEX

Xiaoqin HU, Thi Nhung Pham, Pierre André Buvet
Lexiques, Dictionnaires, Informatique (LDI) UMR-7187
Université Paris 13, Paris, France

Problématique

- **Notion d'anaphore**
  - ✓ L'anaphore est un phénomène linguistique de se référer à un élément précédemment mentionné dans le texte ; le mot ou la phrase qui se réfère à un élément précédent est appelé l'anaphore et l'élément auquel le mot ou la phrase se réfère est appelé l'antécédent (Ruslan Mitkov 2003)
  - ✓ Le processus de déterminer l'antécédent d'une anaphore est appelé la résolution d'anaphore.
- **L'importance de la résolution d'anaphore pour TAL**

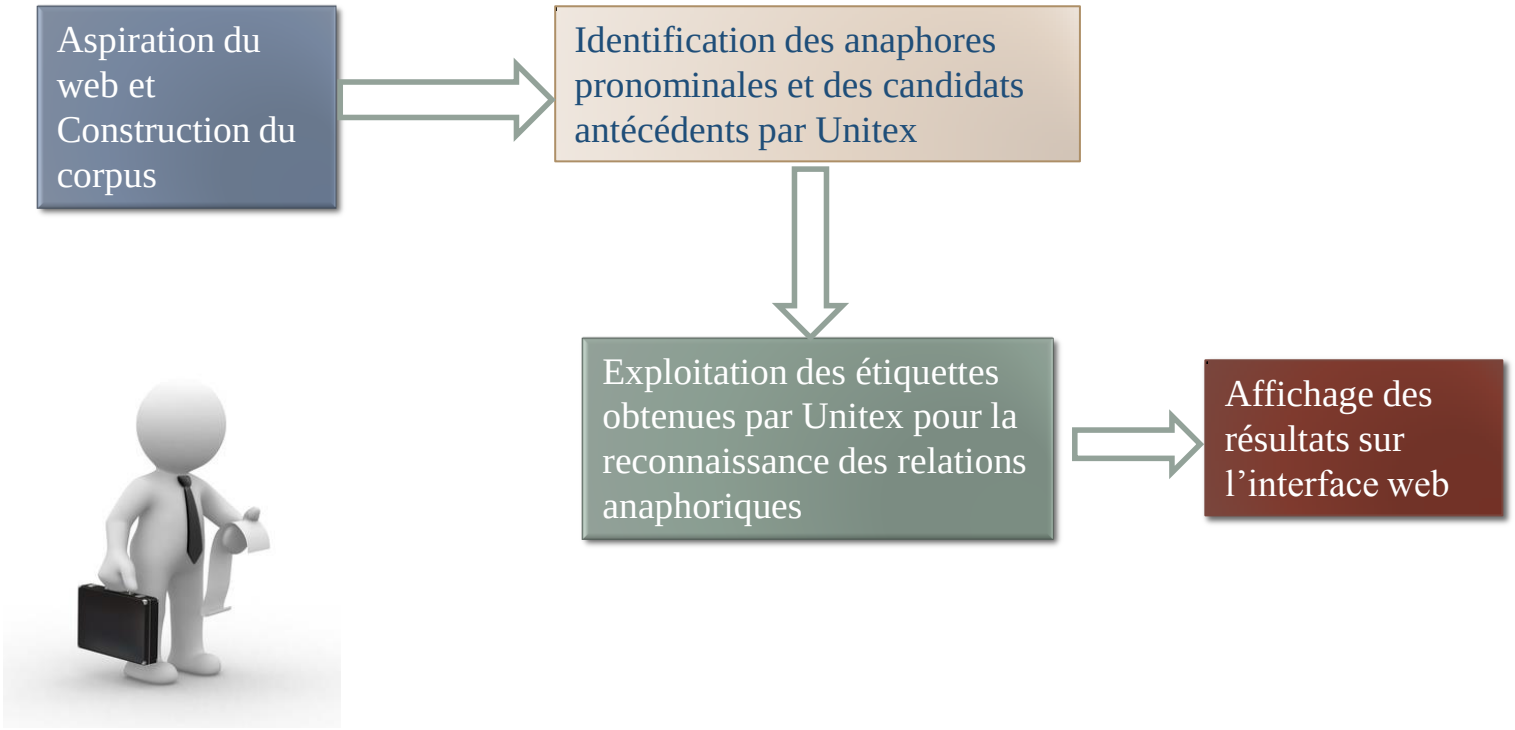
  Plusieurs tests prouvent que la résolution d'anaphore permet d'améliorer la performance de plusieurs applications de traitement automatique des langues, telles que la traduction de machine, le résumé automatique, le système de dialogue, l'extraction de l'information, etc.
- **Etat des connaissances**
  - ✓ Les méthodes basées sur des informations morpho-syntaxiques (ex, genre, nombre, etc.)
  - ✓ Les méthodes en se posant sur les règles de niveau sémantique ou pragmatique.
  - ✓ Pour toutes les méthodes, les connaissances syntaxiques sont nécessaires.

- **Unitex et objectif du projet**
  - ✓ Développer un système de résolution d'anaphore pronominale qui exploite des ressources linguistiques (des dictionnaires électroniques et des grammaires locales) pour identifier les pronoms (sujets et compléments) et les associe aux groupes nominaux qui sont leurs antécédents à l'aide de la plateforme Unitex.

Présentation générale du système de résolution d'anaphore

- Le système permet de reconnaître automatiquement les relations anaphoriques dans les commentaires sur des sites de vente en ligne.
- La résolution d'anaphore du système se limite, pour l'instant, aux anaphores pronominales.
- Le système récupère automatiquement le corpus des sites de vente donnés à l'avance
- En intégrant la plateforme Unitex, le système repère les anaphores pronominales et les candidats antécédents en les associant les informations morphosyntaxiques correspondantes. Il le fait à l'aide des grammaires locales et des dictionnaires électroniques
- Un script python sélectionne l'antécédent correspondant de chaque pronom anaphorique en calculant la distance et considérant les étiquettes morphosyntaxiques obtenues par Unitex
- Un script perl rédige automatiquement une page web pour les résultats. Cela permet d'afficher les résultats sur l'interface web

Architecture du système de résolution d'anaphore



Construction du corpus

- **Télécharger les pages web**
  - ✓ Cinq sites de vente riches de commentaires : http://www.priceminister.com, http://www.lesnumeriques.com, http://www.amazon.fr, http://www.ciao.fr et http://www.commentcamarche.net.
  - ✓ Sur chaque site web, on aspire des centaines de commentaires sur les produits de haute technique, tels que casque-audio, téléphone, lecteur mp3,…
  - ✓ Le module perl LWP ::UserAgent.
- **Extraire les textes signifiants**
  - ✓ Perl HTML::Parser
  - ✓ Parcourir les pages web téléchargées au format html
  - ✓ Repérer le texte entre les balises marquant le début et la fin que nous avons prédéfinies pour le parser
  - ✓ Récupérer les métadonnées : le titre, la date et l'auteur de commentaire
  - ✓ Encadrer respectivement le titre, la date et l'auteur par "(TITRE:", ":TITRE)", "(DATE :", " : DATE)", " (AUTEUR :" et " : AUTEUR)"
- **Nettoyer les textes extraits**

  Eliminer les retours chariot en trop, les espaces superflus, le symbole "<", ">" ou "&" dans le contenu de texte qui empêche de transformer le texte en fichier XML
- **Résoudre le problème d'encodage**
  - ✓ Transformer le fichier en utf8
  - ✓ Décoder les entités html

Identification des anaphores pronominales et des candidats antécédents par Unitex

- **Identification des anaphores pronominales**
  - ✓ Unitex: un graphe **NON_PRONOM_ANAPHORIQUE.grf** faisant appel à trois sous graphes pour repérer les pronoms non anaphoriques, tels que le pronom impersonnel (ex, *il s'agit, il faut, il y a*, etc.) et les déterminants sous la même forme que certains pronoms anaphoriques, tels que *ces, les, la, le*, etc. selon le contexte et les encadre avec "NON":
    - **NON_PRONOM_IMPERSONNEL_SUJET.grf** qui comprend toutes les formes comme *il faut, il paraît, il semble…*
    - **NON_COD.grf** qui comporte les déterminants directes comme *les, le, la…*
    - **NON_COI.grf** qui permet de distinguer les déterminants sous la même forme que les pronoms d'objet indirect comme *leur-même, lui-même, ses, son, leur, ...* Exemple
  - ✓ Unitex: un graphe **PRONOM_ANAPHORIQUE.grf** faisant appel à deux sous graphes pour d'identifier les pronoms anaphoriques et les étiqueter avec les balises sous la forme <PRO _ANAPHORIQUE genre="F/M" nombre="S/P"></ PRO _ANAPHORIQUE >:
    - **PRONOM_ANAPHORIQUE_COD.grf** qui identifie les pronoms de sujet direct
    - **PRONOM_ANAPHORIQUE_SUJET.grf** qui identifie les pronoms d'objet Exemple
  - ✓ Unitex: un autre graphe **replace_NONIL.grf** qui supprime toutes les étiquettes " NON " ajoutées par le graphe NON_PRONOM_ANAPHORIQUE.grf Exemple

Identification des anaphores pronominales et des candidats antécédents par Unitex

- **Identification des candidats antécédents**

✓ Notion de groupes nominaux
  - Les candidats antécédents peuvent être non seulement un nom (N) mais aussi un groupe nominal (GN).
  - Les groupes nominaux sont définis par la forme basique Dét+N+Modifieur dans la grammaire transformationnelle de Maurice Gross (1986).
  - Dét: déterminant défini, indéfini, numéral(ex, un, deux, trois……), adjectif(ex, ces, mon, sa, son, etc.), adverbial(ex, plus, ne…ni, ne…jamais, etc.), nominal(ex, beaucoup de)…; Modifieur: adjectif (ex, joli), adjectif modifié par un adverbe(ex, très joli), proposition relative…(ex, qui est très joli)

✓ Repérer les groupes nominaux de noms d'artefacts comme les candidats antécédents en intégrant une ressource lexicale développée par le laboratoire LDI (Lexique, Dictionnaire, Informatique).
  - Cette ressource lexicale comprend 13400 noms d'artefacts
  - Chaque entrée est associée à une série d'informations syntacico-sémantiques : le nombre, le genre, la classe sémantique et le domaine.

✓ **graphe_GN.grf** qui fait appel à deux sous graphes pour identifier les candidats antécédents:
  - **graphe_GN_artefact.grf** qui reconnaît les groupes nominaux de noms d'artefacts
  - **gaphe_GN_marque.grf** qui reconnaît les noms de marques

✓ Sortir les informations syntaxiques de chaque candidat antécédent dans les étiquettes sous forme xml
  - Chaque groupe nominal de nom d'artefact reconnu est étiqueté en même temps avec les informations syntaxiques : nombre et genre.
  - Ces informations sont enregistrées comme attributs dans les balises xml. Exemple

Exploitation des étiquettes obtenues par Unitex pour la reconnaissance automatique des relations anaphoriques pronominales

- **Sélection de l'antécédent correspondant pour chaque pronom anaphorique parmi les candidats antécédents**
  - ✓ La sélection est exécutée en fonction de la distance entre le candidat antécédent et l'anaphore.
  - ✓ Un script perl numérote les antécédents et les pronoms anaphoriques ensemble et ajoute l'identifiant comme attribut dans chaque balise correspondante
  - ✓ Le script perl numérote aussi les phrases selon le symbole {S} obtenu dans le prétraitement d'Unitex
    - Unitex segmente les textes en phrases dans le prétraitement et marque le début de chaque phrase par un symbole {S}
  - ✓ Transformation en fichier XML par le script perl
    - ajouter les balises <texte></texte> pour chaque texte dans le corpus, dont le début est marqué par (TITRE :…… : TITRE)
    - transformer le symbole (TITRE :…… : TITRE) en balise xml <titre>…</titre>.
    - étiqueter chaque phrase par les balises xml contenant l'information de numérotation (<phrase id="…">…</phrase>).
    - ajouter une balise <doc> tout au début du document et une balise </doc> tout à la fin Exemple

Exploitation des étiquettes obtenues par Unitex pour la reconnaissance automatique des relations anaphoriques pronominales

- **Sélection de l'antécédent correspondant pour chaque pronom anaphorique parmi les candidats antécédents.**

  ✓ L'algorithme de calcul de distance :

  a) Pour chaque anaphore pronominale, on collecte tous les candidats antécédents dans les trois phrases les plus proches de la phrase où se trouve cette anaphore pronominale ;

  b) On assigne à chaque candidat antécédent un score dont la valeur est identique à sa valeur de numérotation. Plus proche du pronom, plus haut est son score ; si le candidat antécédent se trouve dans le titre, on rajoute 2 à son score original ;

  c) On sélectionne le candidat antécédent qui a le score le plus haut comme l'antécédent de l'anaphore pronominale et on enregistre respectivement le numéro de l'antécédent sélectionné et celui de l'anaphore pronominale correspondante. On enregistre les paires de (N° d'anaphore, N° d'antécédent) dans un fichier ;

  d) S'il n'y pas de candidats antécédents reconnus dans les trois phrases les plus proches devant l'anaphore pronominale, par défaut, on considère que cette anaphore se réfère au titre et on écrit la paire comme (N° d'anaphore, titre).

  ✓ Cet algorithme est réalisé par un module python lxml

  ✓ On enregistre les paires de numéros (identifiant de l'anaphore et identifiant de son antécédent correspondant) dans le fichier **resultat_raap.txt**

Affichage des résultats sur l'interface web

- **Rédaction de page web**
  - ✓ Un script affichage.pl encadre les paires anaphore-antécédent en fonction des résultats calculés (N° d'anaphore, N° d'antécédent) par les balises html …
  - ✓ On met une étiquette [[GN= N°]] pour l'antécédent et [[Ref:GN= N°]] pour l'anaphore entre les balises… .
  - ✓ On ajoute aussi une étiquette [[TITRE]] pour le titre de chaque texte.
  - ✓ On écrit la tête du fichier html, les balises ou <p> pour chaque bloc ou chaque paragraphe correspondant.
- **Interface web**
  - L'interface web à travers laquelle l'utilisateur communique avec la machine est écrite dans le fichier **raap.php**
  - Elle contient un champ d'options permettant de sélectionner un corpus à traiter
    - Le champ d'options comporte cinq choix : Priceminister , Lesnumeriques, Amazon, Ciao, Commentcamarche qui sont liés à cinq sites web : http://www.priceminister.com, http://www.lesnumeriques.com, http://www.amazon.fr, http://www.ciao.fr et http://www.commentcamarche.net à partir desquels on télécharge le corpus.
  - Le bouton "valider" permet de lancer tous les scripts derrière
    - Les scripts de chaque étape sont liés par les fichiers .bat sous Windows ou .sh sous linux.

Evaluation

- On calcule respectivement le taux de précision/rappel de la reconnaissance des candidats antécédents, le taux de précision/rappel de la reconnaissance des anaphores pronominales et le taux de précision/rappel de l'identification des relations anaphoriques

Taux de précision

	GN	AP	Paires (GN, AP)
Amazon	85.12%	86.76%	76.33%
Priceminister	96.51%	76.23%	66.33%
Ciao	94.92%	86.67%	77.67%
Commentcamarche	93.11%	83.89%	69.33%
Lesnumeriques	87.46%	73.33%	65.75%

Taux de rappel

	GN	AP	Paires (GN, AP)
Amazon	68.72%	82.33%	69.43%
Priceminister	86.86%	70.50%	65.67%
Ciao	78.14%	86.21%	76.33%
Commentcamarche	76.60%	78.33%	71.97%
Lesnumeriques	63.85%	68.21%	65.33%

Conclusion et perspectives

- **Système de résolution d'anaphore**
  - ✓ Aspire le corpus du site web donné et le nettoie automatiquement
  - ✓ Intègre la plateforme Unitex dans la chaîne de traitements pour l'identification des candidats antécédents et les pronoms anaphoriques
  - ✓ Exploite les étiquettes obtenues par Unitex et calcule la distance entre les candidats antécédents et les anaphores pronominales pour repérer la relation anaphorique
  - ✓ Affiche le résultat sur l'interface web en mettant les paires (son antécédent, l'anaphore) en couleur
- **La pertinence de reconnaissance des pronoms anaphoriques et des candidats antécédents peut être influencée par la pertinence d'étiquetage morphosyntaxique.**
  - ✓ Dans Unitex, les dictionnaires Dela et Delac enregistrent toutes les possibilités morphosyntaxiques pour chaque entrée, par exemple, *a* peut être un nom qui indique la lettre alphabétique ou être la conjugaison du verbe *avoir* au présent de la troisième personne singulière.
  - ✓ On enlève les possibilités morphosyntaxiques moins fréquentes pour certaines entrées souvent utilisées, mais cette méthode sans considérer les contextes donne des erreurs des fois
- **Les fautes d'orthographes (par ex., omission d'accent), langages SMS, des emprunts et des néologismes**
  - ✓ Qui ne peuvent pas être reconnus par le dictionnaire d'Unitex baissent aussi le taux de précision et de rappel des graphes établis pour repérer les pronoms anaphoriques et des candidats antécédents.
- **On peut aussi ajouter les informations sémantiques dans l'outil pour une sélection de candidat antécédent plus précise.**
- **De plus, la résolution d'anaphore de cet outil peut également être étendue à tous les types d'anaphores et tous les types de candidats antécédents.**

Merci pour votre attention.