Cours sur le traitement automatique des langues

Violaine Prince
Université de Montpellier 2
LIRMM-CNRS

Introduction

- Le traitement automatique des langues (TAL) :
  - Domaine fondateur en Intelligence Artificielle
    - La langue comme système de représentation des connaissances
  - Relève également de l'Interaction Homme Machine
    - La langue comme système de communication
  - Mais aussi du traitement des données
    - Les textes comme réservoir d'information

Introduction

- Le TAL est un domaine bi-disciplinaire (informatique et linguistique)
- Son objectif est de fournir
- Des METHODES et des OUTILS
- Dans le cadre d'APPLICATIONS
- Qui aident à la résolution de TACHES utiles.

TACHES principales

- Ingénierie linguistique
  - Automatisation des tâches et des ressources textuelles
- Aux autres domaines de l 'informatique
  - Apporter une contribution aux problèmes dont les données sont des textes
- Aide à la recherche linguistique
  - Est-ce que l'informatique peut aider le linguiste dans son travail de compilation, de classification et de caractérisation ?

Ingénierie linguistique :

- Aide à la traduction automatique
- Correcteurs grammaticaux et orthographiques
- Dictionnaires
- Alignement de corpus multilingues
- Résumés automatiques

Application aux autres domaines de l 'informatique

- Moteurs de recherche d 'information
- Interrogation de bases de données
- Tuteurs intelligents
- Informatique documentaire
- Reconnaissance de la parole continue

Aide à la recherche linguistique

- Recherche de fréquences
- Aide à l 'analyse de textes
- Typage de données textuelles

Les outils

- Plate-formes complètes ou semi-complètes
  - Ressources lexicales : réseaux, dictionnaires
  - Ressources grammaticales : grammaires, corpus étiquetés
  - Analyseurs morphologiques, morphosyntaxiques
  - Outils applicatifs spécifiques : aligneurs, mémoires de traduction, algorithmes d'apprentissage
  - Outils d'évaluation : mesure des performances des autres outils

Méthodes

- Les différents types de « TAL »
  - Informatique linguistique
    - Langages formels
    - Algorithmique du texte
    - Représentation des connaissances et raisonnement
    - Systèmes à base d 'agents
    - Apprentissage automatique
  - Linguistique informatique
    - Statistiques
    - Logique

Éléments traités dans ce cours

- Terminologie :
  - la manière dont sont constitués les termes, leurs propriétés lexicales, grammaticales et sémantiques. => MATHIEU LAFOURCADE
- L'analyse automatique de phrases :
  - Comment reconnaître des termes ainsi que leur combinaison comme étant des unités de langage correctement constituées
    => CHRISTIAN RETORE et VIOLAINE PRINCE

Dimensions de l'Analyse automatique

- Morphologique
  - Les mots sont polymorphes, et ont des propriétés grammaticales.
- Syntaxique
  - Les mots ne sont pas combinés au hasard.
- Sémantique
  - Les mots ont un sens, mais la phrase, combinaison de mots, doit également en avoir un.
- Pragmatique
  - Le sens des mots et des phrases est dépendant du contexte d'énonciation.

Analyse morphologique

- Objectif :
  - Reconnaissance de mots dans un texte
  - Reconnaissance de la ponctuation
  - Affectation d 'une catégorie grammaticale au mot
- S 'appelle LEMMATISATION ou ETIQUETAGE

Exemple

- Ajouter du texte

Reconnaissance de la frontière des unités lexicales (ul)

AJOUTER

Reconnaissance de l 'ul comme « motif » présent dans un thésaurus : catégorie « verbe », forme « infinitif »

Lettre majuscule A : reconnaissance du début du texte

Quelques difficultés

- J ajoute du texte

Reconnaître une forme de « je » pronom personnel

Reconnaître une forme du motif « ajouter » ou le reconnaître comme motif : catégorie « verbe », forme « première personne du singulier ».

La multiplicité des signes

- Les signes spéciaux :
  - Qui interviennent dans une unité lexicale :
    - - , exemple : porte-manteau
    - ‘ , exemple : aujourd’hui
  - qui marquent la contraction :
    - ‘ , exemple : j ’arrive
  - Qui marquent un début ou une fin d ’unité composée :
    - « », ( ), majuscule et point, — —.

- Les signes de ponctuation :
  - , : ;
- Les signes d ’énumération :
  - 1) nombre suivi d ’une parenthèse fermante
  - °, -,*
- Le symbole du dialogue
  - —
- Les signes d ’annotation (*), (1)
- Les signes arithmétiques et les nombres inclus dans un texte

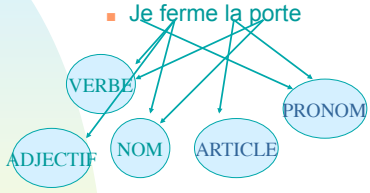
L ’ambiguïté

- Des signes :
  - l ’apostrophe, le tiret, la parenthèse fermante
- Des catégories affectables à une ul :
  - une texture ferme ⟵ adjectif
  - je ferme la porte ⟵ verbe
  - la ferme de Jean ⟵ nom

- De la majuscule : début de texte, nom propre ou emphase
  - Pierre est parti.
  - Pierre qui roule n’amasse pas mousse.
  - Oh Pierre du Savoir !
- ambiguïté de forme précise
  - je ferme la porte
    - ferme
      - catégorie : VERBE
      - forme : 1ere personne du singulier (FORME FLECHIE)
  - Il ferme la porte
    - ferme
      - catégorie : VERBE
      - forme : 3ème personne du singulier

Le côté « multiplicatif » de l ’ambiguïté de catégorie

- Je ferme la porte



La combinatoire théorique

- pronom verbe pronom verbe
- pronom verbe article verbe
- pronom verbe pronom nom
- pronom verbe article nom
- pronom nom pronom verbe
- pronom nom article verbe
- pronom nom pronom nom
- pronom nom article nom
- pronom adjectif pronom verbe
- etc. soit 12 combinaisons alors qu ’il n ’y en a qu ’une seule de bonne...

LA BONNE COMBINAISON

Les différentes techniques d'analyse morphologique

- Soit une ul U dans un texte T
  - Etiquetage
    - affectation d 'une catégorie grammaticale et/ou d 'une forme à U
  - Lemmatisation
    - étiquetage et reconnaissance de U comme élément de dictionnaire

Exemples

- Je ferme la porte
  - Etiquetage :
    - (« je », pronom personnel), (« ferme », verbe),(« la » article), (« porte »,nom)
    - étiquetage en bi-grammes
      - (« U »,C_U)
  - Lemmatisation
    - Etiquetage plus
    - (« ferme », verbe : FERMER)
      - (« U »,C_U, LEXEME)

- Etiquetage tri-gramme
  - (« U »,C_U, F_U)
    - où F est la forme prise par U (forme fléchie)
- Lemmatisation avec étiquetage tri-gramme
- (« U »,C_U, F_U, LEXEME)
- Un **lexème** est une unité lexicale signifiante.
  - Exemples : FERMER, JE, LA, PORTE, PORTER...

Quelques éléments de vocabulaire

- **Entrée lexicale ou lemme** :
  - Unité lexicale qui sert d'entrée du dictionnaire. Elle est généralement représentée par :
    - la chaîne de caractères X qui la définit
    - le lexème L auquel elle est associée
    - la catégorie grammaticale associée
    - la ou les forme(s) fléchie(s) du lexème catégorisé prise par la chaîne de caractères.
    - (X, L, C, {F_X})

Exemples

- Il existe trois entrées lexicales pour l 'ul « ferme »
  - (« ferme », FERMER,verbe,{ 1ère personne du singulier, 3ème personne du singulier})
  - (« ferme », FERME, nom commun, féminin singulier)
  - (« ferme »,FERME, adjectif qualificatif, {masculin singulier, féminin singulier})
- Remarque : les lemmes peuvent être ambigus.

Les dictionnaires

- Thesaurii lexicographiques :
  - FERMER : verbe
  - FERME-1 : nom commun
  - FERME-2 ; adjectif qualificatif
- Dictionnaires de formes fléchies : toutes les entrées lexicales de type (X, L, C, $\{F_X\}$)

- Dictionnaires sémantiques de formes fléchies:
  - on ajoute le sens du mot pour augmenter la discrimination
    - (« ferme », FERMER,verbe,{ 1ère personne du singulier, 3ème personne du singulier}, *FERMER)
      - ici, on met un pointeur sur la forme infinitive fermer, qui va elle, porter le ou les sens.
    - (« ferme », FERME-1, nom commun, féminin singulier, *bâtiment agricole*)
    - (« ferme », FERME-1b, nom commun,féminin singulier, *poutre de toit*)
    - etc.

Comment réaliser la lemmatisation

- Pour chaque ul U d 'un texte T
- Si on a un dictionnaire de forme fléchies de type (X, L, C, $\{F_X\}$) alors
  - apparier U et X
  - Récupérer toutes les sous-listes (L, C, $\{F_X\}$) correspondantes.

Qualité de la lemmatisation

- La qualité de la lemmatisation est l 'adéquation réelle entre ce que doit valoir U dans le texte T et la sous-liste (L, C, {F $_X$}) sélectionnée.
- A priori, plus il existe de listes différentes avec la même tête de liste, plus la qualité de la lemmatisation est mauvaise. Il faut donc désambiguiser.

Techiques de désambiguisation

- Par l 'analyse syntaxique
- Par apprentissage sur un corpus
- On reste dans l 'hypothèse d 'un dictionnaire de formes fléchies

Désambiguisation par l 'analyse syntaxique

- Tous types d 'analyse depuis l 'adjonction de quelques règles de syntaxe jusqu 'à l 'analyse complète.
  - Présentation de règles d 'interdiction
    - un article ne peut pas être suivi d'un verbe
      - pronom verbe article verbe
      - pronom nom article verbe
      - pronom adjectif article verbe

Je ferme la porte

à supprimer

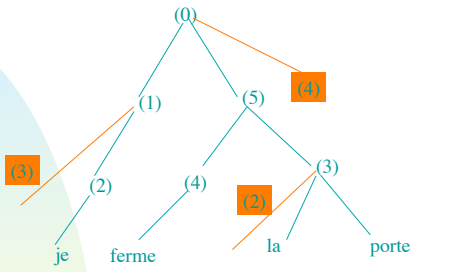
- Règles de composition autorisées (et ce qui n 'est pas autorisé est interdit)
  - Pronom verbe
  - Article nom
  - Article adjectif
    - pronom verbe pronom verbe
    - pronom verbe article verbe
    - pronom verbe pronom nom
    - pronom verbe article nom
    - pronom nom pronom verbe
    - pronom nom article verbe
    - pronom nom pronom nom
    - pronom nom article nom
    - pronom adjectif pronom verbe
    - pronom adjectif pronom nom
    - pronom adjectif article verbe
    - pronom adjectif article nom

à garder

à supprimer

Utilisation des Grammaires

- (0)proposition -> groupe sujet groupe verbal
- (1)groupe sujet -> groupe nominal
- (2)groupe nominal -> pronom
- (3)groupe nominal -> article nom
- (4)groupe verbal -> verbe
- (5)groupe verbal -> verbe groupe nominal



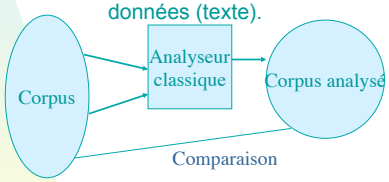
Arbre d'analyse syntaxique (descendant)

Les problèmes

- Le langage naturel n 'est pas indépendant du contexte sur le plan grammatical
- Les grammaires de la langue ne sont pas complètes
- Les textes peuvent être a-grammaticaux

L'apprentissage sur corpus

- Analyse de corpus
  - Un corpus est un ensemble de données (texte).



Rappel et bruit

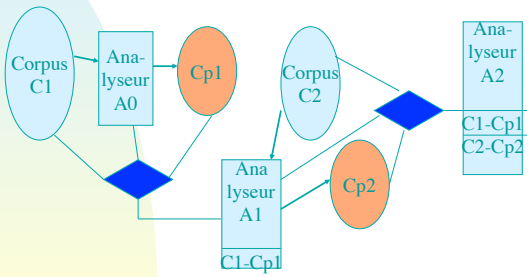
- Soit n le nombre d 'éléments du corpus d 'origine C.
- Soit m le nombre d 'éléments du corpus analysé CA.
- Soit t le nombre d 'éléments de l 'intersection de C et CA.

- Rappel : t/n
- Bruit : m-t/n
- La qualité d'une analyse dépend de ces deux variables.

Première technique d 'apprentissage

- On part d 'un analyseur qui possède un dictionnaire D et des règles R.
- On teste sur un corpus C1 et on produit Cp1. Si le rappel r1<1, on fait :
- R∪ {C1-Cp1}
- D∪ {Ui ∈ {C1-Cp1}}

Modification itérative de l 'analyseur



Problèmes et limites

- Problèmes
  - Compatibilité des ajouts ?
  - Non redondance ?
  - Mécanismes d 'abstraction non directement prévus
  - Données incomplètes en lemmatisation
- Limites
  - Le bruit n 'est pas géré.

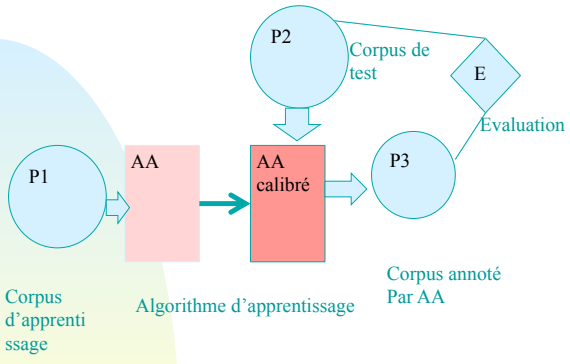
Eléments de solution

- Problèmes
  - Vérifications manuelles (PennTree), réduction de l 'absurdité
  - redondance par génération ou identité : suppression
  - Mécanismes d 'abstraction: « raisonnement »
  - Etiquetage plutôt que lemmatisation.

- Analyseur lexical de Pitrat
  - Un thésaurus et des règles de conjugaison
- Etiqueteurs
  - dictionnaires des formes fléchies simplifiés
    - A apprentissage sur corpus d 'entraînement : Brill, PennTree ❖❖
    - Grammaticaux simples (markoviens, ATN, automates, etc.)
- Analyseurs morphosyntaxiques

Deuxième technique d'apprentissage

- On part d'un corpus déjà étiqueté considéré comme bon. On le divise en deux parties : P1 et P2. On « masque » les étiquettes de P2.
- On fournit P1 en entrée à un algorithme d'apprentissage pour l'entraîner.
- Une fois l'entraînement fait, on teste l'algorithme sur P2 non étiqueté et on compare les résultats de l'algorithme avec les étiquettes de P2 (démasquées).



Problèmes et limites

- L'existence de corpus de référence
  - En partie résolu par les golden standards
- Le comportement sur un corpus nouveau
  - Est difficilement prévisible
  - N'est pas toujours évaluable humainement
- Les erreurs détectées ne peuvent pas être corrigées autrement que par ré-entraînement.

La suite du cours

- L'étiquetage morphologique seul n'est pas intéressant en TAL.
  - Plusieurs outils existants
  - La désambiguisation des catégories grammaticales est gérée par l'analyse syntaxique.
- => ANALYSE MORPHOSYNTAXIQUE !!!

Les grammaires : élément majeur

- METHODES plutôt que les outils
- Hiérarchies des langages selon Chomsky
- Grammaires indépendantes du contexte
  - Avantages, inconvénients
  - Algorithmes, exemples
  - Implémentation pour analyser quelques fragments
- Autres modèles de grammaires
  - Avec exemples et si possible démonstration