Détermination du caractère anaphorique du pronom personnel « *il* »

Mais qui est ce « il » ?

Agathe Mollé

Université de Nantes

3 mai 2013

agathe.molle@etu.univ-nantes.fr

Encadrant : Nicolas Hernandez

Plan

Plan

Résolution d'anaphores

Anaphore

Expression désignant une entité : « *il* », « *cette maison* »
Possède un ou plusieurs antécédent(s).

Ex : « *Vendredi,* ***M. Edmond Maire*** *a réuni la presse pour apporter* ***son*** *soutien à* ***sa*** *fédération des cheminots* »

Problématique : trouver l'antécédent.

Caractère co-référentiel d'une expression

L'une des étapes de la résolution d'anaphore est de déterminer le caractère co-référentiel (anaphorique) d'une expression.

Intérêt des pronoms : fréquents et facilement identifiables.

On s'intéresse au pronom personnel « *il* ».
Celui-ci peut s'avérer :

- anaphorique « *il ne pouvait entrer ainsi dans un conflit ouvert* »
- impersonnel « *il est temps de mettre un terme à la grève* »

Objectif

Actuellement :

- De nombreuses approches traitant du « *it* » anglais : à base de règles (Lappin et Leass, 1994) ou bien d'apprentissage automatique (Li et al., 2009 - Bergsma et Yarowski, 2011).
  $\Longrightarrow$ s'appuient sur la structure syntaxique de l'anglais donc pas aisément portables
- A notre connaissance, une seule étude pour le français : à base de règles (Danlos, 2005)

Objectif : déterminer le caractère anaphorique du « *il* » à l'aide d'apprentissage automatique supervisé.

Plan

ILIMP - Danlos 2005

- ILIMP : outil à base de règles
- Pour établir les règles : lister exhaustivement toutes les constructions impersonnelles possibles (lexique-grammaire, Gross 1994)
- Règles transformées en patrons linguistiques pour UNITEX :
  - Ex : `Il[IMP] <être.V:3s> <Adj1:ms> de <V:W>`
    reconnaissant « *Il est difficile de résoudre ce problème.* »
- Prétraitement des données : tokénisation + toutes les étiquettes morpho-syntaxiques possibles (dictionnaire DELAF)

ILIMP - Danlos 2005 (2)

- Occurrences étiquetées `[ANA]` par défaut. Si elles correspondent à un patron : `[IMP]`. Pour les cas ambigus : `[AMB]`.
- Évaluation : sur *Le Monde 1994* (+3 000 000 tokens), précision de 97,5%
- Source des erreurs :
  - Pas de désambiguïsation pour l'étiquetage morpho-syntaxique
  - Toutes les constructions impersonnelles ne sont pas répertoriées
  - Heuristiques brutales pour éviter la balise `[AMB]`

NADA - Bergsma et Yarowski 2011

- NADA : Outil à base d'apprentissage automatique supervisé et d'observations statistiques.
- Traite du « *it* » anglais
- 2 types de traits :
  - lexicaux-syntaxiques (extraits sur la phrase entière)
  - tirés du corpus n-gramme de Google
- Évalué sur 3 corpus : BBN, WSJ-2 et ItBank $\Longrightarrow$ entre 85,1% et 86,2%

Plan

Corpus de travail

- Portion du corpus annoté par Tutin et al. (2000) : *Le Monde*, rubrique économie
- + de 200 000 mots dont 1176 pronoms « *il* » (537 anaphoriques / 631 impersonnels)
- Seules les expressions anaphoriques sont annotées $\Longrightarrow$ on ne garde que les annotations du pronom « *il* » et on considère tous les « *il* » non annotés comme impersonnels
- Division du corpus en 3 parties :
  - 10% : corpus de développement
  - 10% : corpus de test (que l'on se réserve pour de futurs travaux)
  - le reste : corpus d'entraînement

Plan

Traits recensés

D'après L. Danlos, les constructions impersonnelles reposent sur des conditions tant lexicales que syntaxiques :

- « *Il est* ***violet*** » [1]
- « *Il* ***pleut*** »
- « *Il est* ***probable*** *que Fred viendra* »
- « *Il est* ***difficile*** *de résoudre ce problème* »
- « *Il est* ***difficile*** *à résoudre (ce problème)* »

$\Longrightarrow$ Traits extraits des propriétés sémantiques et syntaxiques de la tête lexicale + de la nature syntaxique du complément.

Problèmes :

- Comment distinguer la tête lexicale et le complément ?
- Comment construire des lexiques suffisamment riches ?

1. Les exemples mentionnés sont tirés de l'article de L. Danlos (2005)

Traits (2)

Autres observations :

- des expressions typiquement impersonnelles (« *quoi qu'il en soit* », « *il était une fois* », etc).
- la présence de déterminants possessifs à la 3$^{\text{ème}}$ personne du singulier (« *son* », « *sa* »), du pronom « *lui* » , d'autres occurrences de « *il* », etc.
- des données extrinsèques au corpus (corpus n-gramme Google)

Plan

Approche

Approche expérimentée :

- Étiquetage de séquences (*sequence labeling*)
- Outil Wapiti
- Classes IMP et ANA
- 2 expériences
- Prétraitement avec les analyseurs d'Apache OpenNLP

Baseline :

- Tout étiqueter selon la classe majoritaire, à savoir IMP (53,66%)

Expériences

Première expérience :

- Traits classiques utilisés pour la reconnaissance de segments syntaxiques (*chunks*)
- Observation du contexte local (formes de surface, étiquettes morpho-syntaxiques, bigrammes, majuscules, etc. dans une fenêtre de 5 tokens)

Seconde expérience :

- Ajout de traits spécifiques à cette tâche
- Simplification du problème : au lieu de chercher la tête lexicale, on observe le premier verbe, adjectif ou adverbe suivant l'occurrence
- Lexiques non implémentés
- Statistiques sur la phrase (présence de possessifs, etc.)

Évaluation

- Validation croisée, 10 partitions
- Sur le corpus d'entraînement
- Mesures utilisées :
  - Taux d'occurrences correctement classées (exactitude/*accuracy*)
  - Pour chaque classe :

$$Precision = \frac{Nombre\ d'occurences\ correctement\ classees\ ANA}{Nombre\ d'occurences\ classees\ ANA}$$

$$Rappel = \frac{Nombre\ d'occurences\ correctement\ classees\ ANA}{Nombre\ d'occurences\ annotees\ comme\ anaphoriques}$$

$$F - mesure = \frac{2 * (Precision * Rappel)}{(Precision + Rappel)}$$

Plan

Résultats

Expérience avec les traits classiques de reconnaissance des segments syntaxiques :

	ANA			IMP		
Exactitude	Précision	Rappel	F-mes.	Précision	Rappel	F-mes.
0,79	0,78	0,76	0,77	0,81	0,83	0,82

Série d'expériences avec incorporation de nouveaux traits :

	ANA			IMP		
Exactitude	Précision	Rappel	F-mes.	Précision	Rappel	F-mes.
0,8	0.79	0.77	0.78	0.81	0.83	0.83

Interprétation

- La seule observation du contexte local suffit à obtenir des résultats intéressants (79% d'instances correctement classées)
- Les traits ajoutés lors de la seconde expérience ne sont pas concluants : ils n'améliorent presque pas les scores.
- Ceci s'explique par une heuristique trop naïve pour détecter les têtes lexicales
- Les traits relatifs à la présence d'entités spécifiques (possessifs,. . .) n'améliorent pas les scores car dans ces cas, les traits classiques de la reconnaissance de chunks suffisent

Plan

Perspectives

- Aborder le problème comme une tâche de classification simple
- Exploiter la totalité du corpus, afin d'obtenir des proportions plus raisonnables (actuellement 1/5)
- Détecter correctement la tête lexicale d'une construction, ainsi que son complément (chunking ?)
- Établir des lexiques d'entités ancrant des phrases impersonnelles ou anaphoriques
- Utiliser le corpus n-gramme de Google, à l'instar des travaux de Bergsma et Yarowski sur le « *it* » anglais

Conclusion

- Résultats encourageants pour une approche à base d'apprentissage automatique supervisé.
- La seule observation du contexte local dans une fenêtre de 5 tokens donne 80% de réussite
- Il reste des pistes à explorer + des implémentations à améliorer

Questions

Des questions ?