Classification d'occurences du pronom personnel « *il* » selon s'il est co-référentiel ou non

8 avril 2013

Résumé

Introduction

La résolution d'anaphores est une problématique fondamentale partagée par divers domaines de Traitement Automatique des Langues (TAL). Pour ce faire, l'une des étapes consiste à déterminer le caractère co-référentiel d'une expression, autrement dit : y'a-t-il une autre expression dans le texte qui réfère à la même entité ?

Au sein de cette problématique, les pronoms sont beaucoup traités car ils ont pour avantage d'être fréquents et facilement identifiables (Danlos, 2005). Parmi ceux-ci, on peut distinguer le pronom personnel de la 3ème personne du singulier, masculin : le « *il* ». En effet, selon les cas, celui-ci peut s'avérer :

- co-référentiel (« *il ne pouvait entrer ainsi dans un conflit ouvert* »)
- non co-référentiel (« *il est temps de mettre un terme à la grève* »)

Il existe actuellement plusieurs techniques traitant du pronom personnel « *it* » en anglais. Certaines sont à base de règles lexico-syntaxiques (Lappin et al. 1994), d'autres à base d'apprentissage (Li et al. 2009) ou d'observations statistiques (Bergsma et al. 2011). Pour le français, il n'existe qu'un outil à base de règles (Danlos 2005).

L'objectif est donc d'établir un système déterminant le caractère co-référentiel du « *il* » français adaptant les travaux réalisés pour l'anglais. Une approche à base d'apprentissage automatique et d'observations statistiques sera empruntée. L'idée étant de classifier chaque occurence comme co-référentielle ou non co-référentielle. Pour entrâiner ce classifieur, on dispose du corpus Tutin et al. (2005). L'observation du corpus de développement (un dixième du corpus total) ainsi que la littérature nous permettront de dégager les traits à apporter au classifieur.

Plan

1 Etat de l'art

Résumé Lappin

L'approche à base de règles à été adaptée au français à travers l'outil ILIMP, conçu pour classer les occurences du pronom « *il* » selon si elles sont anaphoriques ou impersonnelles (Danlos 2005). Cet outil travaille sur des textes bruts, non annotés. La démarche est la suivante : dans un premier temps, on observe des constructions de phrases particulières qui permettent d'établir des patrons (en partant du principe que les entités lexicales employées avec le « *il* » sont déterminantes). Ces patrons sont écrits grâce à l'outil UNITEX (ref) (effectuant une tokénisation, un étiquetage morpho-syntaxique + traits flexionnels).

Les règles établies vont permettre de trouver toutes les occurences de « *il* » impersonnelles, et vont les décorer de la balise [IMP] (la balise [ANA] étant définie par défaut). Il se trouve que pour chaque phrase, le contexte gauche, même s'il est complexe, est analysable sans trop de difficultés, c'est le contexte droit qui peut poser des ambiguités. Pour les cas ambigus, il y a une balise [AMB], mais dans l'optique de l'utiliser le moins possible, l'outil va essayer de déterminer si le pronom est impersonnel à l'aide d'heuristiques basées sur les fréquences.

Pour évaluer ILIMP, le corpus utilisé est Le Monde et les occurences de « *il* » ont été annotées à la main. La plupart des erreurs obtenues sont des balises [ANA] à la place de [IMP] car [ANA] est la balise par défaut, et les règles utilisées n'ont pas couvert tous les cas. Il y a peu d'erreurs [IMP] à la place de [ANA] malgré les heuristiques brutales. Par ailleurs, les résultats sont meilleurs sur un corpus journalistique que littéraire.

Résumé Li

Une approche parallèle, proposée par S. Bergsma et D. Yarowski (2011), va allier l'apprentissage automatique supervisé et des observations statistiques. Ce système (NADA) prend en entrée des données tokénisées (en anglais) et détecte les occurences de « *it* » non-référentielles. Les données en question n'ont pas été analysées morpho-syntaxiquement, car ils font le postulat que l'ambiguïté repose sur la présence d'entité lexicales spécifiques, et non sur les informations morpho-syntaxiques.

Deux types de traits sont alors apportés au classifieur : des traits lexicaux-syntaxiques mais aussi des statistiques tirées du corpus n-gram de Google (ref), le classifieur en question étant fondé sur un modèle de régression.

Les traits lexicaux sont extraits sur la phrase entière. Il y a par exemple la présence ou non d'autres formes du pronom personnel (*its/itself*, la présence d'acronymes, de prépositions juste avant le nom, de « *that* » ou « *to* » après le pronom, etc.

Les traits statistiques sont établis grâce à l'énorme banque de données fournie par Google. En effet, pour chaque 4-gramme contenant le pronom « *it* », on va établir un patron (le « *it* » est remplacé par un « _ ») et regarder dans les données si d'autres mots remplissent ce patron, ou si seulement le cas de figure avec « *it* » a été rencontré. Par exemple, si on extrait « *it is able to* », on va rechercher toutes les occurences de la forme « _ *is able to* ». Les résultats obtenus permettent d'ajouter un trait (un poids) dans le classifieur. Seulement, ce corpus étant gigantesque, il a fallu le compresser au maximum. Pour ce faire, diverses techniques ont été utilisées : ne récupérer que les 4-grammes, uniquement ceux contenant « *it* », « *they* » ou « *them* », tronquer les mots à 4 caractères, les encoder, ne retenir que les changements d'un 4-gramme à l'autre, etc.

NADA a été évalué sur les corpus BBN, WSJ-2 et ItBank (refs). Les résultats sont similaires pour les 3 corpus (entre 85.1% et 86.2%). Par comparaison, l'évaluation a aussi été réalisée avec seulement les traits lexicaux et seulement les traits statistiques, obtenant de moins bons résultats indépendamment (en moyenne 80% pour les traits lexicaux et 83.5% pour les traits statistiques).

2 Approche

Nous avons opté pour une approche à base d'apprentissage automatique, la tâche étant de classifier chaque occurence de « *il* » comme COREF (coréférentielle) ou NONCOREF. Pour ce faire, nous procédons en deux phases : l'entraînement puis le test.

Pour réaliser la phase d'entraînement, nous devons dans un premier temps déterminer quels traits apporter au classifieur. Ceux-ci proviennent de la littérature ainsi que de l'observation des données. Nous retenons divers types de traits :

- L'observation des tokens autour de l'occurence (lemmes) :
  - Premier adjectif/verbe/adverbe/préposition après
  - Première combinaison telle verbe+préposition après
  - Première conjonction avant (puisqu', quand,...)
  - Signes de ponctuation
- La présence ou non de certains tokens (booléens) :
  - Présence de déterminant possessif 3e personne (ses, son, sa...)
  - Présence d'un acronyme
  - Présence du verbe falloir
  - Présence du pronom 'lui'
  - Présence d'une autre occurence de 'il'
  - Présence d'une date/time ?
- Les statistiques sur la phrase (entiers) :
  - Taille de la phrase
  - Nombre de « il » dans la phrase
- Certains patrons fréquents (booléens) :
  - il y a (pas de verbe) / il s'agit ...
  - il se verbe (pas que)

Une fois les traits définis, nous pouvons construire une modélisation et entraîner notre classifieur. Cette modélisation permettra par la suite de prédire la classe pour une nouvelle entrée, autrement dit une nouvelle occurence de « *il* ».

3 Cadre expérimental

3.1 Corpus de travail

Nous avons entraîné et testé notre classifieur sur le corpus annoté Tutin et al. 2000 composé d'un million de mots.

Ce corpus anote une grande partie des expressions anaphoriques, cependant certaines sont rejetées car trop complexes. Les expressions retenues appartiennent alors à des classes fermées.

Chacune de ces expressions est annotée dans le but d'indiquer les éléments mis en jeu ainsi que la relation entre l'expression anaphorique et son(ses)

antécédent(s). Cette relation peut appartenir à l'une des 5 classes suivantes : coréférence, membre de, description, phrase ou indéfinie.

L'annotation a été faite à la main par deux linguistes (le processus n'est pas exclusivement manuel puisque les expressions sont pré-annotées automatiquement et des outils d'édition sont utilisés pour simplifier la tâche). Afin de mesurer la fiabilité de l'annotation, une évaluation sur 5% du corpus a été effectuée.

L'annotation est effectuée en XML. Des balises permettent de distinguer les sections, les paragraphes (<p>), les phrases (<s>) ainsi que les expressions (<exp>). Chaque expression a un ID, et une balise <ptr> permet de lier une expression anaphorique a son(ses) antécédent(s), et renseigner le type de liaison. Quelques autres balises ou attributs permettent de distinguer les cas spéciaux, comme la balise <seg> pour délimiter un segment d'un antécédent spécifique à celui-ci.

Contrairement à d'autres schémas d'annotation (MUC et MATE de base (ref)), ce schéma ne se restreint pas aux coréférences. Il traite d'un ensemble fini de relations, similaire au schéma UCREL (ref).

3.2 Déroulement de l'expérience

Afin de mettre en place une chaîne de traitement, nous travaillons sur la portion du corpus contenant les articles du journal *Le Monde* (rubrique économie). Ceci nous permet d'établir rapidement notre processus d'extraction de traits, même si dans l'idéal, il aurait été préférable de travailler sur une plus grande portion du corpus.

Nous divisons ensuite cette portion en 3 parties. Un dixième est consacré au développement, un autre dixième au test, et tout le reste à l'entraînement du classifieur. C'est sur cette dernière partie que nous travaillons ensuite. Pour l'évaluation par validation croisée (voir ci-dessous), nous partitionnons le corpus d'entraînement en 10.

Les traits sont définis à l'aide de l'outil Wapiti. Celui-ci prend en entier des fichiers tabulés décrivant chaque token : une ligne correspond à un token, et chaque colonne contient une information telle que forme de surface, étiquetage morpho-syntaxique, flexions, ... La dernière colonne contient la classe que l'on recherche, ici l'annotation au format BIO [1].

1. Le format BIO permet d'étiqueter les tokens selon qu'ils débutent (Beginning of – B), sont à l'intérieur (Inside of – I), ou sont hors (Outside – O) de la zone d'intérêt.

Nous avons préalablement prétraité les données avec les analyseurs Apache OpenNLP (ref) (? ? ? uima ?). Puis, à l'aide de scripts Bash, nous mettons en place nos fichiers tabulés contenant les informations nécessaires à l'extraction de traits

Wapiti requiert aussi un fichier de patrons, permettant d'extraire les traits à partir de notre fichier tabulé. Nous utilisons le fichier *chpattern.txt*, établi par Thomas Lavergne (Isabelle Tellier ?) et contenant les traits classiques pour la reconnaissance de chunks, auquel nous avons ajouté nos propres patrons.

A partir de ces traits, nous pouvons dorénavant entraîner notre modèle.

3.3 Evaluation

Nous évaluons le système obtenu par validation croisée avec 10 partitions. Les mesures utilisées seront la précision, le rappel et la F-mesure.

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

$$Rappel = \frac{Nombre\ d'occurences\ correctement\ classees\ COREF}{Nombre\ d'occurences\ annotees\ comme\ referentielles}$$

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

Wapiti étant assez long à calculer le modèle, nous évaluons dans un premier temps uniquement sur la première partition.

4 Discussion

Par un souci de manque de temps, nous n'avons pas pu mener la totalité de ce projet à bien.

Premièrement, la totalité du corpus n'a pas été exploitée. En effet, nous n'avons travaillé que sur les articles du Monde, ce qui représente un cinquième du corpus Tutin et al. Ensuite, nous n'avons évalué le classifieur que sur la première partition, ce qui ne reflète pas forcément correctement les résultats réels.

Il aurait aussi pu être intéressant de se servir de statistiques extrinsèques au corpus, au même titre que les travaux de S. Bergsma et D. Yarowski sur la langue anglaise. Effectivement, il existe un corpus Ngram fourni par Google, en français.

Par ailleurs, il pourrait être judicieux de nettoyer les données avant de les fournir à la chaîne de traitement, ou du moins normaliser certains éléments.

Conclusion