Détermination du caractère co-référentiel du pronom personnel « *il* »

Agathe Mollé

20 avril 2013

Résumé

Introduction

Les anaphores sont des expressions qui permettent de désigner des entités dans les textes (« *il* », « *cette maison* »), et qui possèdent un ou plusieurs antécédent(s), autrement dit une autre expression désignant la même entité. La résolution d'anaphores est une problématique fondamentale partagée par divers domaines du Traitement Automatique des Langues (TAL). Il s'agit de déterminer quel est l'antécédent (ou les antécédents) de l'expression anaphorique.

Dans ce contexte, nous avons travaillé sur un problème plus particulier : la détermination de l'aspect co-référentiel d'une expression.

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 : « *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* »)

Lorsqu'un pronom n'est pas co-référent, on le qualifie d'impersonnel.

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). Celui-ci obtient de très bons résultats, mais les outils à base de règles manquent généralement de robustesse. Par ailleurs, il est dépendant d'un outil tiers (UNITEX), rendant plus difficile une désolidarisation du prétraitement.

L'objectif de notre travail est donc d'établir un système déterminant le caractère co-référentiel du « *il* » français. L'approche à base de règles est abandonnée au profit d'une approche à base d'apprentissage automatique supervisé, cette tâche pouvant être considérée comme une classification (prédiction d'étiquettes). Par ailleurs, on peut aussi chercher à définir la tâche comme un problème d'étiquetage de séquence (Sequence labeling), de par sa sensibilité au contexte local.

Notre tâche pouvant être définie selon ces deux approches, il convient de les expérimenter toutes deux. Lors de la planification d'expériences, il s'est avéré plus simple de réaliser l'expérience avec le séquence labeler [1].

En l'état de notre avancement, nous ne rapportons pour l'instant que les deux expériences à base de sequence labeling.

Pour réaliser celles-ci, nous disposons du corpus Tutin et al. (2005).

La première expérience consiste à utiliser les traits classiques de la détection de segments pour le français. Pour la deuxième, nous dégageons de nouveaux traits par observation du corpus de développement (un dixième du corpus total) et par recensement des traits de la littérature.

Cet article décrit dans un premier temps les travaux existants sur lesquels nous nous sommes appuyés. Nous présentons ensuite notre corpus de travail, puis nous développons les traits que nous avons recensé à travers la littérature et l'observation du corpus.

Dans un second temps, nous présentons le déroulement des expériences que nous avons réalisé sur ce corpus ainsi que leurs résultats.

Nous terminons cet article par une discussion sur le travail effectué et sur ce qui pourrait être intéressant d'exploiter à l'avenir.

1 Etat de l'art

Pour résoudre cette tâche, nous nous sommes appuyés sur les seuls travaux à notre connaissance en français et sur l'un des plus récents pour l'anglais.

1. L'expérience avec le sequence labeler consistait en une extension d'un TP réalisé sur la reconnaissance d'entités nommées

En effet, les travaux de Lappin et Leass utilisent des règles qui s'appuient sur la structure syntaxique de l'anglais, donc ne sont pas aisément portables au français. De même, les autres approches par classification ont été mises de côté, toujours par manque de portabilité.

L'approche à base de règles développée pour le français est l'outil ILIMP, conçu pour classer les occurrences 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 ainsi que des traits flexionnels).

Les règles établies permettent de trouver toutes les occurrences 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 ambiguïtés. Par exemple, deux constructions peuvent différer de façon très subtile : « *Il manque du poivre (dans cette maison)* »/« *Il manque de poivre, ce rôti* » (ici la première construction est impersonnelle et la seconde anaphorique alors qu'elles se distinguent seulement par *du*/*de*) ou encore des constructions comme « *Il est certain que Fred viendra* » qui peuvent être à la fois impersonnelles et anaphoriques. 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, comportant 3.782.613 tokens, dont 13.611 occurrences de « *il* », celles-ci ayant été annotées à la main. Il ressort de cette évaluation un taux de précision de 97.5%. 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.

L'approche pour l'anglais proposée par S. Bergsma et D. Yarowski (2011), va allier l'apprentissage automatique supervisé et des observations statistiques. Ce système (NADA : Non-Anaphoric Detection Algorithm) prend en entrée des données tokénisées 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'entités 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 est 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 Corpus de travail

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

Ce corpus annote 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).

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). Cette portion contient N mots. 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 transformons cette portion en un corpus au format BIO [2]. Pour cela, on ne garde que les balises associées aux pronoms « *il* », et on les convertit en étiquettes `B-coref` (début d'une entité de « *il* » coréférentielle). On considère ensuite tous les « *il* » non annotés comme impersonnels, et on les décore de l'étiquette `B-non-coref`.

Nous disposons à ce stade d'un corpus contenant N occurrences du pronom « *il* », dont N sont coréférentielles, et N impersonnelles.

Nous divisons ensuite ce corpus en 3 parties. Un dixième est consacré au développement, un autre dixième au test, et tout le reste à l'entraînement du sequence labeler. 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.

3 Compilation des traits recensés dans la littérature et enrichis par l'observation

Nos données mises en place, nous cherchons maintenant à recenser les traits pouvant être utiles à notre tâche.

2. 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.

Notre première expérience consiste à tester les traits classiques utilisés pour la reconnaissance de chunks (TODO : explication chunk). Nous nous sommes appuyés sur le travail d'Isabelle Tellier(ref + explications). Ceux-ci permettent d'observer le contexte local des occurrences à savoir : les formes de surface et étiquettes morpho-syntaxiques dans une fenêtre de 5 tokens (unigrammes et bigrammes), les préfixes et suffixes entre 1 et 4 caractères, la présence de majuscules, de ponctuation ou de chiffres dans une fenêtre de 3 tokens, si le token étudié est composé uniquement de majuscules/ponctuation/chiffres et enfin s'il commence par une majuscule.

Pour notre seconde expérience, nous apportons nos propres traits au sequence labeler. Ceux-ci proviennent de la littérature ainsi que de l'observation des données.

Nous avons choisi d'extraire les traits au sein de la phrase, à l'instar des travaux de S. Bergsma et D. Yarowski. En effet, ceux-ci se sont basés sur les procédés humains : ils ont fait l'expérience en demandant à un sujet d'annoter 200 occurrences du « it » anglais. Celui-ci obtient un taux d'exactitude de 85% s'il dispose des 4 tokens de chaque côté de l'occurrence, contre 95% s'il dispose de la phrase entière. Nous faisons l'hypothèse que cette fenêtre est aussi plus judicieuse pour le texte français.

Voici les traits retenus :

- L'observation des formes de surface autour de l'occurence :
  - 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 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) :
  - verbes météorologiques (brouillasser, bruiner, brumasser, brumer, crachiner, gouttiner, grêler, neigeoter, neiger, pleuvasser, pleuviner, pleuvioter,

pleuvoir, pleuvoter, pluviner, reneiger, repleuvoir, surventer, tonner, venter, verglacer)
- autres verbes impersonnels (advenir, s'agir, s'ensuivre, incomber, résulter ?, falloir, apparoir, barder, bichoter, dracher, s'en falloir, se pouvoir, y avoir)
- structures impersonnelles de temps (il est temps de/que, il est l'heure de/que)
- patrons fréquents (Il arrive que, il paraît que, il semble que, il se peut que, il me semble que, il vaut mieux que (ou il vaut mieux + infinitif), il suffit que (ou il suffit que + infinitif), etc.)
- expressions avec il impersonnel (il était une fois, il y a, quoi qu'il en soit,...)
- verbes déclaratifs (dire, affirmer, annoncer, croire, prétendre,...) sauf forme passive
- verbes de sentiment (admirer, craindre, s'étonner,...)
- verbes de volonté (vouloir, exiger, désirer, souhaiter, supplier, prier,...) sauf forme passive
- verbes d'opinion (croire, douter, imaginer, penser, estimer, juger,...)

4 Cadre expérimental

Une fois les traits définis, nous pouvons construire une modélisation et entraîner notre sequence labeler. Cette modélisation permettra par la suite de prédire le caractère co-référentiel pour les nouvelles occurrences de « *il* ».

4.1 Mise en place d'une chaîne de traitement

Les traits sont définis à l'aide de l'outil Wapiti (ref). 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 .

Nous avons préalablement prétraité les données avec les analyseurs Apache OpenNLP (ref). 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 Isabelle Tellier pour ses travaux sur la reconnaissance de chunks, auquel nous avons ajouté nos propres patrons.

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

4.2 Baseline

4.3 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.

4.3.1 Expérience avec les traits classiques de reconnaissance des chunks

4.3.2 Série d'expériences avec incorporation de nouveaux traits

5 Discussion

Phrase d'entrée.

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 sequence labeler 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

Remerciements

Nicolas Hernandez