

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

Agathe MOLLÉ

30 avril 2013

Résumé

Dans cet article, nous nous intéressons au problème de la détermination du caractère co-référentiel du pronom personnel « *il* », que nous traitons comme une tâche d'apprentissage automatique supervisé. Nous présentons deux expériences utilisant d'étiquetage de séquences, réalisées sur le corpus [Tutin et al. \(2000\)](#). La première utilise les traits classiques de la détection de segments syntaxiques pour le français. Pour la seconde, nous ajoutons des traits tirés de la littérature ou observés sur le corpus de développement. Nous montrons que l'approche par apprentissage obtient des résultats encourageants.

Abstract

In this article, we deal with the problem of determining whether the french pronoun « *il* » is coreferential, using a supervised machine learning approach. We present two experiments of sequence labeling, done on the [Tutin et al. \(2000\)](#) corpus. The first one uses classic features for french chunks detection. For the second one, we add features identified in the litterature and by observation. We show that the learning approach gives encouraging results.

Mots-clés

Co-référence, Apprentissage automatique, Pronom anaphorique, Pronom impersonnel, Résolution d'anaphores

Keywords

Co-reference, Machine Learning, Anaphoric pronoun, Expletive pronoun, Anaphora resolution

Introduction

Les anaphores sont des expressions qui permettent de désigner des entités dans les textes (« *il* », « *cette maison* »). Une expression anaphorique n'a de sens que si l'on dispose de son antécédent, autrement dit une expression précédente désignant la même entité. Par exemple, dans la phrase « *Vendredi, M. Edmond Maire a réuni la presse pour apporter son soutien à sa fédération des cheminots* », les expressions anaphoriques « *son* » et « *sa* » réfèrent à la même entité que leur antécédent « *M. Edmond Maire* ».

La résolution d'anaphores consiste à déterminer quel est l'antécédent (ou les antécédents) de l'expression anaphorique. C'est une problématique fondamentale partagée par divers domaines du Traitement Automatique des Langues (TAL).

L'une des étapes de cette problématique est la détermination de l'aspect co-référentiel d'une expression.

Dans ce contexte, nous nous intéressons aux pronoms car ils ont pour avantage d'être fréquents et facilement identifiables (Danlos, 2005), et plus particulièrement au pronom personnel de la 3^{ème} personne du singulier, masculin : « *il* ». En effet, selon les cas, celui-ci peut s'avérer :

- co-référentiel (anaphorique)
« *il ne pouvait entrer ainsi dans un conflit ouvert* »
- non-coréférentiel (impersonnel)
« *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 and Leass, 1994), d'autres à base d'apprentissage (Li et al., 2009) ou d'observations statistiques (Bergsma and Yarowsky, 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. Il est en effet difficile d'énumérer de manière exhaustive toutes les constructions impersonnelles. Par ailleurs, cet outil n'effectue pas de désambiguïsation morpho-syntaxique lors de son prétraitement, il souffre donc d'erreurs liées à l'absence d'analyse syntaxique.

L'objectif de notre travail est d'établir un système déterminant le caractère anaphorique du « *il* » français. L'approche à base de règles est abandonnée au profit d'une approche à base d'apprentissage automatique supervisé. Cette tâche peut en effet être considérée comme une tâche de classification : en décrivant chaque instance de « *il* » par un jeu de traits appropriés, on peut ensuite prédire l'étiquette

d’une nouvelle instance, étant donné son vecteur de traits. 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 *sequence labeler*¹.

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

Nous expérimentons en effet deux jeux de traits pour caractériser les pronoms et leur contexte.

La première expérience consiste à utiliser les traits classiques de la détection de segments syntaxiques 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 de traits mentionnés dans la littérature.

Pour réaliser nos expériences, nous disposons du corpus [Tutin et al. \(2000\)](#), lequel est seulement annoté avec les expressions anaphoriques. Nous indiquerons dans la section 2 comment nous avons procédé pour constituer un corpus avec des instances de chaque classe.

La section 1 décrit dans un premier temps les travaux existants sur lesquels nous nous sommes appuyés. La section 2 présente notre corpus de travail, la section 3 les traits que nous avons recensé et que nous utilisons pour nos expérimentations. La section 4 présente le déroulement des expériences que nous avons réalisé sur ce corpus. Enfin, en section 5, nous détaillons les résultats obtenus ainsi qu’une discussion sur le travail effectué et sur ce qu’il pourrait être intéressant d’exploiter à l’avenir.

1 État de l’art

Pour résoudre cette tâche, nous nous sommes appuyés sur les seuls travaux à notre connaissance en français ([Danlos, 2005](#)) et sur l’un des plus récents pour l’anglais ([Bergsma and Yarowsky, 2011](#)).

En effet, les travaux de [Lappin and Leass \(1994\)](#) utilisent des règles qui s’appuient sur la structure syntaxique de l’anglais, donc ne sont pas aisément portables au

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

français. De même, les autres approches par classification ont été mises de côté, car les traits apportés aux classifieurs dépendent eux aussi de la structure de l'anglais

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.

Le travail de L. Danlos se définit en deux temps : la première démarche conduit à la construction de règles, puis la seconde amène à la reconnaissance des formes en appliquant les règles en question.

Pour élaborer les règles, Danlos observe les propriétés lexicales et syntaxiques des constructions impersonnelles, et plus particulièrement de leur tête lexicale. Elle s'appuie sur le lexique-grammaire du français qui décrit l'ensemble des têtes lexicales des phrases simples du français, avec leurs arguments syntaxiques et les alternances possibles. Elle définit alors une liste de verbes, adjectifs et expressions caractérisant des constructions impersonnelles, qu'elle divise d'ailleurs en deux catégories : les constructions intrinsèquement impersonnelles, qui ne peuvent avoir comme sujet que « *il* », et les constructions impersonnelles à sujet profond extraposé (sujet phrastique ou nominal). On peut distinguer par exemple les verbes météorologiques qui ancrent des constructions intrinsèquement impersonnelles (« *Il pleut* »), ou alors des adjectifs tels « *probable* » qui traduisent des constructions impersonnelles à sujet (phrastique) profond extraposé (« *Il est probable que Fred viendra* »).

Pour chaque construction impersonnelle, il se trouve que le contexte gauche de la tête lexicale, même s'il est complexe, est analysable sans trop de difficultés. Par exemple, on peut rencontrer les constructions suivantes :

Il est difficile de résoudre ce problème.

Il peut lui paraître très difficile de résoudre ce problème.

Il ne s'est pas avéré difficile de résoudre ce problème.

Dans ces 3 cas, la tête lexicale est l'adjectif « *difficile* » mais le contexte gauche varie. Il s'agit alors de répertorier tous les cas possibles et de les intégrer dans les règles, opération minutieuse mais qui ne pose pas de réel problème.

A l'inverse, c'est le contexte droit qui peut poser des ambiguïtés. Celles-ci peuvent être d'ordre syntaxique (une séquence peut recevoir plusieurs analyses syntaxiques), d'ordre lexical (par exemple « *Il est certain que Fred viendra.* » peut être à la fois anaphorique et impersonnelle) ou alors être dûes à des constructions impersonnelles qui ne diffèrent en surface que de façon très subtile par rapport à des constructions personnelles (« *Il reste la valise du chef* » (impersonnelle) / « *Il reste la priorité du chef (le chômage)* » (anaphorique)).

Pour éviter d'avoir trop de constructions considérées comme « ambiguës », Danlos va avoir recours à des heuristiques établies à partir d'études quantitatives

ou de ses intuitions linguistiques. Par exemple, pour une construction donnée, si on analyse plus fréquemment les phrases comme impersonnelles dans les corpus, alors cette construction sera considérée comme telle.

Afin de les incorporer à ILIMP, ces règles sont traduites en patrons linguistiques grâce à l'outil UNITEX². Par exemple, une construction typiquement impersonnelle est le verbe « être » à la 3ème personne du singulier, suivi d'un adjectif du lexique préalablement défini (il est ici la tête lexicale), suivi de la proposition « de » et d'un verbe à l'infinitif, on écrit alors le patron suivant :

Il[IMP] <être.V:3s> <Adj1:ms> de <V:W>

Ce patron correspond ainsi à des phrases comme :

Il est difficile de résoudre ce problème.

UNITEX prend en entrée des données brutes, et les prétraite en effectuant une tokenisation, et en attribuant à chaque token toutes les propriétés morpho-syntaxiques et flexionnelles que celui-ci peut avoir (obtenues grâce au dictionnaire DELAF).

Les données prétraitées et les patrons écrits, il s'agit maintenant de décorer chaque occurrence de « il » de la balise adéquate, à savoir [IMP] (impersonnelle), [ANA] (anaphorique) ou [AMB] (ambiguë).

Pour ce faire, toutes les occurrences de « il » reçoivent par défaut la balise [ANA]. A chaque fois que la construction correspond à un des patrons prédéfinis, celle-ci devient alors [IMP]. Si le cas est ambigu, et que les heuristiques n'ont pas suffi à déterminer le caractère impersonnel de la construction, la balise [AMB] est alors employée.

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. Par exemple, l'inversion du sujet n'a pas été correctement traitée, le lexique des adjectifs impersonnels est incomplet ou encore l'impasse a été faite sur la coordination.

Il y a peu d'erreurs [IMP] à la place de [ANA] malgré les heuristiques brutales.

Pour finir, un étiquetage morpho-syntaxique aurait pu éviter certaines erreurs de prédiction.

L'approche pour l'anglais proposée par Bergsma and Yarowsky (2011), va allier l'apprentissage automatique supervisé et des observations statistiques. Ce système NADA : *Non-Anaphoric Detection Algorithm* prend en entrée des données

2. <http://www-igm.univ-mlv.fr/unitex/#>

tokénisées et détecte les occurrences de « *it* » non-coréférentielles. Les données en question n'ont pas été analysées morpho-syntaxiquement, car le postulat est fait 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-gramme de Google (Brants and Franz, 2006), 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 observer 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 occurrences 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 (Weischedel and Brunstein (2005), WSJ-2 (tiré du Penn Treebank (Marcus et al. (1993))) et ItBank (Bergsma et al. (2008)). 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 effectué nos expérimentations sur une partie du corpus ELRA annoté par 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 à son(s) 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.

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 202 091 mots. Ceci nous permet d'établir rapidement notre processus d'extraction de traits.

Nous transformons cette portion en un corpus au format BIO. Ce format 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.

Pour cela, on ne garde que les balises associées aux pronoms « *il* », et on les convertit en étiquettes B-ANA (début d'une entité de « *il* » anaphorique). On considère ensuite tous les « *il* » non annotés comme impersonnels, et on les décore de l'étiquette B-IMP. Toutes les autres entités qui ne sont pas des « *il* » sont annotées O puisqu'elles ne nous intéressent pas.

On obtient par exemple :

Il[B-ANA] *ne*[O] *changera*[O] *pas*[O] *de*[O] *politique*[O] *salariale*[O] .

Il[B-IMP] *y*[O] *a*[O] *même*[O] *des*[O] *rats*[O] *à*[O] *proximité*[O] .

Nous disposons à ce stade d'un corpus contenant 1176 occurrences du pronom « *il* », dont 537 sont anaphoriques, et 631 impersonnelles.

Nous divisons ensuite ce corpus en 3 parties. Un dixième constitue le corpus de développement, un autre dixième celui de test, et tout le reste est consacré à l'entraînement de l'étiqueteur de séquences syntaxiques.

C'est sur cette dernière partie que nous travaillons ensuite. Le corpus de test demeure quant-à lui inchangé et non consulté durant toute cette phase de tests, puisque nous réservons son usage à de futurs travaux.

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.

Notre première expérience consiste à tester les traits classiques utilisés pour la reconnaissance de chunks. Les chunks sont des unités lexicales qui définissent la structure syntaxique superficielle des phrases (groupes nominaux, groupes verbaux, etc.). Nous nous sommes appuyés sur le travail d’Isabelle Tellier qui a mis en place un chunker à l’aide des CRF et de Wapiti [Constant et al. \(2011\)](#). Ces traits 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 certains traits additionnels 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 : l’expérience consistait à demander 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’occurrence :
 - Premier verbe/préposition/adverbe/adjectif/clitique suivant l’occurrence
 - Premier verbe/préposition/adverbe/adjectif/conjonction précédent l’occurrence
 - Première combinaison du type verbe+préposition suivant l’occurrence
 - Signes de ponctuation
- La présence ou non de certains tokens (booléens) :

- Présence de déterminant possessif à la 3ème personne du singulier (ses, son, sa, ...)
 - Présence d'un acronyme
 - Présence du pronom « lui »
 - Présence d'une autre occurrence de « il »
- Les statistiques sur la phrase (entiers) :
 - Taille de la phrase
 - Nombre de « il » dans la phrase
- Certains patrons fréquents (booléens) :
 - verbes météorologiques (pleuvoir, neiger,...)
 - autres verbes impersonnels (advenir, s'agir,...)
 - structures impersonnelles de temps (il est temps de/que, il est l'heure de/que)
 - patrons fréquents (Il arrive que, il paraît que, ...)
 - expressions typiquement impersonnelles (il était une fois, il y a, quoi qu'il en soit,...)
 - verbes déclaratifs (dire, affirmer, ...) sauf forme passive
 - verbes de sentiment (admirer, craindre, s'étonner,...)
 - verbes de volonté (vouloir, exiger,...) sauf forme passive
 - verbes d'opinion (croire, imaginer, penser, ...)

De plus, pour observer le contexte local, les traits classiques pour la reconnaissance de chunks utilisent une fenêtre de 5 tokens (2 avant et 2 après). Nous avons donc envisagé d'élargir celle-ci à 5 tokens avant et après, et de prendre en compte les trigrammes au sein de cette nouvelle fenêtre.

4 Cadre expérimental

Nous avons opté pour une expérimentation par étiquetage de séquences. En apprentissage automatique, ce type de tâche consiste à assigner une étiquette à chaque élément d'une séquence, compte tenu du contexte proche.

Nous utilisons pour cela l'outil Wapiti³ proposé par Thomas Lavergne, implémentant les algorithmes de CRF (Conditional Random Fields) linéaires.

Le principe du modèle CRF appliqué à l'étiquetage de séquences est le suivant : on observe dans un premier temps des comportements similaires dans le voisinage

3. <http://wapiti.limsi.fr>

de l'élément, puis on apprend à l'aide de probabilités distributionnelles les répétitions de contextes locaux, afin d'obtenir des « règles ». Celles-ci permettent ensuite de prédire les étiquettes de nouveaux éléments.

4.1 Mise en place d'une chaîne de traitement

Wapiti prend en entier des fichiers tabulés décrivant chaque token : une ligne correspond à un token, et chaque colonne contient une information telle que la forme de surface, l'étiquetage morpho-syntaxique, les flexions, ... La dernière colonne contient la classe que l'on recherche, ici l'annotation au format BIO .

Nous prétraitions en premier lieu les données avec les analyseurs Apache OpenNLP⁴ (Boudin and Hernandez, 2012). Ce prétraitement nous permet d'obtenir les étiquettes morpho-syntaxiques ainsi que les traits flexionnels. Puis, à l'aide de scripts Bash, nous mettons en place nos fichiers tabulés contenant les informations nécessaires à notre deuxième série d'expériences, à savoir la taille de la phrase, le nombre d'occurrences de « il » dans celle-ci, etc.

Au final, notre fichier tabulé contient les colonnes suivantes : forme de surface du token, son étiquette morpho-syntaxique, ses traits flexionnels, la taille de la phrase, le nombre d'occurrences de « il » dans celle-ci, idem pour « lui » ainsi que les possessifs à la 3ème personne du singulier, les verbes, prépositions, adjectifs, adverbes, conjonctions et clitiques précédent ou suivant le token, et pour finir l'étiquette BIO représentant la classe (O, ANA ou IMP).

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 ajoutons par la suite nos propres patrons.

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

4.2 Baseline

Nous considérons que la Baseline correspond à un étiquetage systématique en B-non-coref, celle-ci étant la classe majoritaire. Le taux d'occurrences correctement classées est donc de 53,66%.

4. <https://opennlp.apache.org/>

4.3 Evaluation

Nous évaluons le système obtenu par validation croisée avec 10 partitions. Les mesures utilisées sont le taux d'occurrences correctement classées (exactitude) et pour chaque classe : la précision, le rappel et la F-mesure.

Par exemple, pour la classe B-coref, les mesures sont données par les formules suivantes :

$$Precision = \frac{\text{Nombre d'occurrences correctement classées COREF}}{\text{Nombre d'occurrences classées COREF}}$$

$$Rappel = \frac{\text{Nombre d'occurrences classées COREF}}{\text{Nombre d'occurrences annotées comme référentielles}}$$

$$Fmesure = \frac{2 * (Precision * Rappel)}{(Precision + Rappel)}$$

5 Résultats et discussion

Nous détaillons à présent les résultats obtenus lors de nos expérimentations.

Expérience avec les traits classiques de reconnaissance des chunks :

Exactitude	B-coref			B-non-coref		
	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
0,79	0,78	0,76	0,77	0,81	0,83	0,82

On constate que l'observation du contexte proche suffit à obtenir des résultats satisfaisants, avec 82% d'occurrences correctement classées.

Série d'expériences avec incorporation de nouveaux traits :

Exactitude	B-coref			B-non-coref		
	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure

La plupart des traits ajoutés n'ont pas été concluents : seuls le nombre de pronoms « il » dans la phrase, le verbe suivant le token ainsi que la préposition suivante ont amélioré les scores. Nous arrivons au final à un taux de prédictions correctes de N.

Nous n'avons pour l'instant pas implémenté les traits relatifs au lexique, mais c'est une piste à envisager pour des travaux futurs.

De plus, seule l'approche par étiquetage de séquence a été abordée à travers nos expériences. Il serait judicieux de tester avant tout la résolution de ce problème en tant que tâche de classification. Comme expliqué dans l'introduction, nous avons opté pour le sequence labeling car ces expériences faisaient écho à d'autres travaux, mais ce n'est pas pour autant qu'il faille rejeter l'approche par classification.

On peut aussi noter que 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.

Par ailleurs, nous n'avons entraîné le sequence labeler que sur la première partition lors des étapes intermédiaires de nos expériences, ce qui ne reflète pas forcément 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

Cette étude s'est focalisée sur la détermination du caractère co-référentiel des occurrences du pronom « il ». Les seuls travaux à notre connaissance existants pour le français utilisant un système à base de règles, nous avons ici opté pour un système d'apprentissage automatique supervisé.

A travers cet article, nous avons décrit les expériences réalisées avec un sequence labeler. Dans un premier temps, nous avons utilisé les traits classiques employés pour de la recherche d'entités lexicales. Ces traits observent le contexte local, et donnent des résultats plutôt satisfaisants.

Nous avons ensuite apporté nos propres traits, tirés de la littérature ainsi que de l'observation de notre corpus de développement. Certains d'entre eux ont amélioré les résultats, même si tous n'ont pas été implémentés à l'heure actuelle.

Il sera donc intéressant pour la suite de tester les autres traits envisagés, ainsi que d'aborder cette tâche comme une tâche de classification.

En définitive, les pistes que nous avons exploré nous donnent des résultats relativement encourageants vis-à-vis de l'apprentissage automatique. Certaines pistes restent cependant à approfondir, afin d'obtenir un outil suffisamment performant pour être incorporé dans un système plus vaste de résolution d'anaphores.

Références

- Bergsma, S., Lin, D., Inc, G., and Goebel, R. (2008). Distributional identification of nonreferential pronouns. In *In ACL*, pages 10–18.
- Bergsma, S. and Yarowsky, D. (2011). Nada : A robust system for non-referential pronoun detection. In *Proc. DAARC*, Faro, Portugal.
- Boudin, F. and Hernandez, N. (2012). Détection et correction automatique d’erreurs d’annotation morpho-syntaxique du French TreeBank. In *Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles (TALN’2012)*, pages 281–291, Grenoble, France.
- Brants, T. and Franz, A. (2006). The google web 1t 5-gram corpus version 1.1. ldc2006t13.
- Constant, M., Tellier, I., Duchier, D., Dupont, Y., Sigogne, A., and Billot, S. (2011). Intégrer des connaissances linguistiques dans un CRF : application à l’apprentissage d’un segmenteur-étiqueteur du français. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles (TALN’2011)*, Montpellier, France.
- Danlos, L. (2005). Ilimp : Outil pour repérer les occurrences du pronom impersonnel il. In *Proceedings of TALN 2005*, Dourdan, France.
- Lappin, S. and Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Comput. Linguist.*, 20(4) :535–561.
- Li, Y., Musilek, P., Reformat, M., and Wyard-Scott, L. (2009). Identification of pleonastic it using the web. *J. Artif. Int. Res.*, 34(1) :339–389.
- Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a large annotated corpus of english : the penn treebank. *Comput. Linguist.*, 19(2) :313–330.
- Tutin, A., Trouilleux, F., Clouzot, C., Gaussier, É., Zaenen, A., Rayot, S., and Antoniadis, G. (2000). Annotating a large corpus with anaphoric links. In *Proceedings of the Third International Conference on Discourse Anaphora and Anaphor Resolution (DAARC2000)*, page 2, Royaume-Uni.
- Weischedel, R. and Brunstein, A. (2005). BBN Pronoun Coreference and Entity Type Corpus. Technical report, Linguistic Data Consortium, Philadelphia.