Classification par paires de mention pour la résolution des coréférences en français parlé interactif

Maëlle Brassier[1], Alexis Puret[2] , Augustin Voisin-Marras[2], Loïc Grobol[3]
(1) LIFAT, Université de Tours, 3 place Jean Jaurès, 41000 Blois, France
(2) LIFO, Université d'Orléans, 6 Rue Léonard de Vinci, 45067 Orléans, France
(3) LATTICE-CNRS, ENS, Université Paris 3, Université Sorbonne Paris Cité.

maelle.brassier@yahoo.fr, {alexis.puret,augustin.voisin-marras}@etu.univ-orleans.fr, loic.grobol@gmail.com

RÉSUMÉ

Cet article présente et analyse les premiers résultats obtenus par notre laboratoire pour la construction d'un modèle de résolution des coréférences en français à l'aide de techniques de classifications parmi lesquelles les arbres de décision et les séparateurs à vaste marge. Ce système a été entraîné sur le corpus ANCOR et s'inspire de travaux antérieurs réalisés au laboratoire LATTICE (système CROC). Nous présentons les expérimentations que nous avons menées pour améliorer le système en passant par des classifieurs spécifiques à chaque type de situation interactive, puis chaque type de relation de coréférence.

ABSTRACT

Mention-pair classification for corefence resolution on spontaneous spoken French.
This paper presents the first experiments conducted by our laboratory (LIFAT) on the question of the resolution of coreference on spontaneous spoken French. We have developed a mention-pair classifier, trained on the ANCOR French coreference corpus, which is based on various classification techniques among which support vector machines (SVM). The paper details several experimental studies that investigate several factors (classification model, interactivity degree, nature of the coreference…) that should affect the performances of the system.

MOTS-CLÉS : détection de coréférence, corpus, apprentissage automatique, classification
KEYWORDS: coreference resolution, corpus, machine learning, classification

1. Introduction

Depuis sa création, le TAL a engendré de nombreuses technologies qui ont permis le développement d'applications comme la traduction et la compréhension de textes pour lesquelles la question de la coréférence, est un enjeu essentiel. Afin de s'assurer de l'interprétation correcte des textes étudiés, il est en effet important de relier les expressions linguistiques aux entités du discours auxquelles elles réfèrent. Une simple erreur au cours de cette tâche peut entraîner un contresens radical d'une phrase. La détection de la coréférence, qui nous intéresse ici, consiste à relier les expressions

(appelées mentions) qui réfèrent la même entité du discours. Considérons l'énoncé (A) ci-dessous : *Le roi de la pop* et *Michael Jackson* font référence à la même entité du discours (en l'occurrence une personne) tout en ayant des dénominations différentes.

A. **Michael Jackson** est mort en 2009. **Le roi de la pop** aura influencé de nombreux artistes. (*coréférence indirecte*)

On distingue la coréférence de l'anaphore, qui décrit seulement le fait que l'interprétation d'une mention dépend de celle d'un autre élément linguistique, son antécédent. Les notions d'anaphore et de coréférence sont étroitement liées, mais sont néanmoins distinctes. En effet, une anaphore ne manifeste pas forcément une relation coréférente puisqu'elle peut reprendre une expression déjà introduite dans le discours sans pour autant désigner la même entité. Dans l'exemple (B), le groupe nominal *Sa toiture* est anaphorique de *La maison* bien que les deux mentions soient non-coréférentes.

B. **La maison** est délabrée. **Sa toiture** tombe en ruine. *(Anaphore nominale*)

À ce jour, peu de systèmes de résolution des coréférences adaptés à la langue française ont été mis en place, malgré l'existence d'un corpus français annoté qui permet de développer des systèmes par apprentissage automatique. Ce papier présente les premiers développements d'un nouveau système de résolution par apprentissage automatique s'inspirant des travaux de (Désoyer et al., 2015) sur le système CROC (Adèle Désoyer et al., 2015). Nous commencerons par placer notre sujet dans son contexte en présentant son état de l'art du domaine. Nous présenterons ensuite en section 3 le corpus ANCOR qui nous a servi de support tout le long de notre travail. Le cadre expérimental de notre étude est détaillé en section 4. La section 5 énumère les traits linguistiques qui jouent un rôle prépondérant dans nos expérimentations, qui seront elles développées ensuite dans la section 6. Cette section 6 correspond à la part la plus original de notre travail : nous y étudions l'impact du degré d'interactivité des différents corpus sur le système ainsi qu'une stratégie de résolution à l'aide d'un multi-classifieur.

2. État de l'art

Les premières techniques de résolution de coréférence se sont tout d'abord basées sur des systèmes à base de règles (rule-based approach). Certains travaux comme ceux de (Hobbs, 1978) et (Lappin & Leass, 1994) se sont notamment concentrés sur le cas des coréférences pronominales. Ces approches se basaient sur une analyse syntaxique profonde des énoncés qui pose souvent des problèmes de robustesse, ce qui a conduit dans un premier temps à la proposition alternative d'approches purement heuristiques (Mitkov, 2002). L'analyse se limite dans ce cas à évaluer une fonction heuristique suivant la présence ou l'absence de traits le plus souvent locaux.

Ces traits se retrouvent, au tournant du millénaire, dans des travaux pionniers (Soon et al., 2001) et (Ng & Cardie, 2002) relevant de l'apprentissage supervisé sur corpus. Cette approche centrée sur les données, qui a été rapidement dominante, consiste à réaliser une classification binaire (coréférent / non coréférent) sur toutes les paires de mentions présentes dans le texte. Outre la mise à disposition

de corpus annotés de taille suffisante, elle nécessite toutefois un travail fin d'ingénierie sur les traits linguistiques d'apprentissage, travail dont permettent de s'affranchir les techniques neuronales d'apprentissage profond. À la suite de (Clark & Manning, 2016) et (Wiseman et al., 2016), puis (Lee et al., 2017) pour un traitement de bout en bout (end-to-end), les techniques connexionnistes connaissent ainsi un développement rapide sur cette problématique, où elles présentent de bonnes performances sur les données de campagnes d'évaluation telle que la *shared task* CoNNL'2012.

La communauté française a eu longtemps peu d'opportunité de conduire des travaux par apprentissage automatique du fait d'un manque crucial de ressources annotées en coréférence pour le français. C'est seulement en 2014, avec la création du corpus ANCOR – présenté ultérieurement – que les recherches sur le français ont pu prendre une nouvelle envergure. En dehors de deux systèmes à base de règles (Longo & Todirascu 2009 ; Godbert & Favre, 2017), nous pouvons aussi citer le projet européen SENSEI (Kabadjov & Stepano, 2016) où un système de résolution des coréférences a été développé par apprentissage sur le corpus ANCOR et intervient dans un processus d'analyse du discours. Notre travail se situe dans la continuité des travaux de (Désoyer et al., 2015) sur le système CROC : nous reprenons en effet l'idée d'un apprentissage supervisé à base de SVM, une approche de résolution *mention-pair* et enfin une reprise des traits descriptifs (*features*) retenus dans le système CROC. C'est à partir de cette base que nous avons proposé une nouvelle stratégie d'analyse reposant sur un multi-classifieur. Un de nos objectifs à terme est de poursuivre ce travail d'optimisation afin de définir une *baseline* robuste pour étudier l'apport réel des techniques d'apprentissage profond par rapport aux classifieurs de l'état de l'art. Il est à noter qu'à l'instar du travail de Désoyer, nous n'avons pas développé pour l'heure un système bout-en-bout. On suppose ici que les mentions sont déjà identifiées : ce sont celles qui sont présentes dans le corpus ANCOR.

3. Corpus ANCOR

Afin de palier au manque de ressource francophone annotée en coréférence, les laboratoires LIFAT et LLL ont réalisé le corpus ANCOR (Muzerelle et al., 2014 ou Desoyer et al., 2015), premier corpus français d'envergure répertoriant les relations de coréférence et relations anaphoriques. Il se compose de trois sous-corpus de parole transcrite qui présentent autant de situations de production orale, entre entretiens et dialogues interactifs (cf. tableau 1). Leur degré d'interactivité est par ailleurs variable et dépend de la situation discursive.

Corpus	Type de dialogue	Interactivité	Durée	Taille (mots)
ESLO_ANCOR	Interview	Faible	25 heures	417 000
OTG	Dialogue en face à face	Forte	2 heures	26 000
Accueil_UBS	Dialogue téléphonique	Forte	1 heure	10 000

TABLE 1 : Description des sous-corpus du corpus ANCOR

L'intérêt du corpus est de proposer une annotation riche qui peut servir à la fois à la conduite d'études linguistiques et à l'apprentissage automatique. Chaque mention, i.e. chaque entité linguistique référant à un objet du discours, et chaque relation sont décrites par un ensemble de propriétés. Une mention est ainsi décrite par l'ensemble de propriétés linguistiques suivant :

- Le Genre (masculin/féminin) et le Nombre (singulieur/pluriel) : GENRE et NB
- L'inclusion ou non dans un groupe prépositionnel : GP
- Type d'entité nommée (toponyme, anthroponyme…) : EN
- La définitude (Défini, indéfini, démonstratif ou explétif) : DEF
- Caractère générique ou spécifique de la référence de l'item considéré : GEN_REF
- Nouvelle entité du discours ou non : NEW

L'annotation a par ailleurs consisté à caractériser l'ensemble des relations de référence entre les mentions présentes dans le corpus. Une relation est quant à elle décrite par des indications sur l'accord en nombre et en genre des deux mentions reliées, mais également le type de la reprise. Cinq classes de relations sont ainsi définies, qui se partagent entre deux types principaux :

- Les **coréférences**, où les deux mentions réfèrent à la même entité du discours : on fait la distinction entre coréférence directe, indirecte, pronominale
- Les **anaphores associatives**, où les deux mentions ne sont pas coréférentes mais partagent un lien référentiel : nominale ou pronominale

Une relation directe correspond à une coréférence où la reprise et l'antécédent ont la même tête nominale ("*le petit garçon … ce gentil garçon*"). À l'inverse, une relation indirecte associera deux expressions aux têtes nominales disjointes, rapprochées généralement par un phénomène de synonymie, hyponymie ou bien d'hyperonymie ("*la maison … le bâtiment … la demeure*"). Enfin, la coréférence pronominale décrit une reprise par un pronom ("*le chien … il*").

Les relations associatives nominales relient deux mentions qui ne sont pas coréférentes. Néanmoins, ces mentions partagent une relation ontologique, c'est-à-dire que l'interprétation de l'une dépend de l'autre, par exemple suivant une relation méronymique ("*le gâteau … sa dernière part*"). La relation associative pronominale répond à cette définition mais met en jeu une reprise par un pronom ("*la France … ils ont gagné*"). Les travaux présentés ici ne concernent que la coréférence.

4. Cadre expérimental

Le travail qui est présenté dans cet article est une première tentative des laboratoires LIFAT et LIFO de construire un système de résolution des coréférences qui, à terme, sera utilisé entre autres pour la détection et le suivi de nominations[1], ceci dans le cadre du projet ANR TALAD. Comme base de réflexion, nous avons souhaité reproduire les expérimentations faites au laboratoire LATTICE

[1] Les nominations sont des formes de désignation émergentes, promues généralement par une communauté sociale donnée pour décrire un nouveau concept, et qui n'ont pas encore été figées dans la langue. Elles sont étudiées en particulier par la communauté d'analyse du discours.

(Desoyer et al. 2015) avec le système CROC. Dans un premier temps, nous avons donc reproduit une *baseline* assez proche de ces travaux. Nous nous poserons ensuite la question d'une amélioration des performances. Nous présenterons ici le cadre expérimental dans lequel a été conduite cette étude.

4.1. Techniques d'apprentissage

Les expérimentations que nous avons conduites ont été réalisées sur la plateforme Weka (Witten et al., 2011) qui fournit de nombreux algorithmes d'apprentissage automatique. Trois types de classifieurs ont été considérés : les arbres de décision (retenus pour le caractère explicatif du modèle de classification obtenu), les séparateurs à vaste marge (SVM, retenus pour le niveau reconnu de performance) et un classifieur bayésien naïf (Naive Bayes) comme *baseline*. Il est à noter que nous avons utilisé les paramètres par défaut de Weka pour chacun d'entre eux.

Le modèle d'arbre de décision j48 intégré dans Weka est une implémentation de l'algorithme C4.5 (Quinlan, 1993). Il construit par apprentissage supervisé un arbre où chaque noeud de décision (dit aussi nœud interne) représente un test sur un unique attribut. La sélection d'un test se fait par le choix de l'attribut qui discrimine au mieux les données et ainsi aura une meilleure qualité de séparation.

SVM est une méthode de classification binaire introduite par Vapnik et Chervonenkis et développée par (Boser et al., 1992). Elle consiste à déterminer une séparation par hyperplan de marge optimale dans l'espace multidimensionnel qui décrit les données d'apprentissage. Afin de traiter des problèmes non linéairement séparables, des fonctions noyaux permettent de transformer l'espace de représentation en un espace de plus grande dimension où l'on cherche une séparation linéaire. Nous avons étudié ici un classifieur linéaire, implémenté par la librairie LibSVM, et un classifieur polynomial avec la libraire SMO qui implémente l'algorithme d'optimisation de (Platt, 1998).

Enfin, Naive Bayes est un algorithme d'apprentissage statistique qui s'inspire du théorème de Bayes. Il repose sur l'hypothèse que chaque variable d'apprentissage est soumise à une indépendance conditionnelle. Cette indépendance supposée facilite l'estimation du modèle de classification, qui peut donner des performances correctes avec assez peu de données d'apprentissage.

4.2. Constitution de l'ensemble des corpus d'apprentissage et de test

Nous avons divisé ANCOR en différents sous-corpus équilibrés dans l'idée d'étudier l'influence de plusieurs paramètres sur la résolution de la coréférence. Nous avons évoqué précédemment le degré d'interactivité, variant d'un corpus à un autre. Nous avons considéré ce facteur de discrimination car cette interactivité semble avoir un fort impact sur la réalisation des chaînes de coréférences. Nous distinguerons donc des sous-corpus d'entraînement spécifiques aux sous-corpus ESLO (peu interactif) et OTG (interactif), en maintenant un équilibrage (50%/50%) entre les deux situations d'interaction. Le corpus UBS n'est pas inclus dans les ensembles d'entraînement du fait de sa taille réduite.

Corpus d'entraînement		ESLO	OTG	Total
SmallTrainingSet	COREF	1500	1500	3000
	NOT_COREF	1075	1075	2150
MediumTrainingSet	COREF	1500	1500	3000
	NOT_COREF	1917	1917	3834
BigTrainingSet	COREF	1500	1500	3000
	NOT_COREF	2617	2617	5234

TABLE 2 : Constitution des ensembles d'apprentissage à partir des sous-corpus ANCOR

L'autre facteur concerne l'équilibre des instances positives/négatives. On connaît l'influence de la prévalence d'une classe sur l'apprentissage des classifieurs. Suivant (Désoyer et al., 2015), nous avons défini trois corpus d'apprentissage (*Small*, *Medium* et *BigTrainingSet* dans le tableau 2) répondant à des ratios respectifs de une, deux et trois instances négatives pour une positive.

		ESLO	OTG	UBS	Total
TestSet_*i*	COREF	500	222	197	919
	NOT_COREF	700	311	276	1287

TABLE 3 : Constitution des ensembles de test à partir des sous-corpus ANCOR

Les corpus d'entraînement contiennent la même proportion d'instances provenant d'ESLO et OTG. Nous avons partagé de même notre corpus de test en plusieurs sous-corpus répondant à des degrés d'interactivité différents, afin d'étudier l'impact de cette dimension dialogique. Le corpus de test a été créé à partir des relations restantes pour chaque corpus, en prenant soin que tous les sous-corpus présentent le même ration d'instances positives (coréférence) et négatives. Compte tenu des relations à disposition, ce ratio a été fixé à 1 positive pour 1,4 négatives. Cette fois, c'est le corpus UBS qui est retenu comme corpus interactif. Ce corpus de test est par ailleurs partagé en 3 sous-corpus (TestSet_*i* dans le tableau 3 ci-dessous) pour permettre des études statistiques en significativité sur les résultats. Afin d'éviter tout sur-apprentissage, l'apprentissage a été réalisé par validation croisée avec 10 plis.

5. Traits linguistiques

Afin de s'assurer d'un bon niveau de performance, il est important de fournir au classifieur des traits linguistiques d'apprentissage pertinents. Pour nos travaux, nous avons choisi de reprendre les traits définis par (Désoyer et al., 2015), en excluant néanmoins ceux se rapportant à l'introduction d'un nouvel élément dans le discours (m1_NEW, m2_NEW et id_NEW). Ces attributs, présents dans le corpus annoté ANCOR, sont difficilement identifiables automatiquement sur du texte brut : leur estimation automatique reste un challenge aussi délicat que la résolution des coréférences elle-même. Nous allons décrire rapidement les attributs d'apprentissage subsistant.

5.1. Traits non-relationnels

Les traits non-relationnels servent à décrire chaque mention, indépendamment de celle à laquelle elle pourrait être liée. Chaque mention d'une paire se voit donc attribuer une étiquette, respectivement *m1* pour la première et *m2* pour la seconde, à qui on associe certains traits non relationnels. Grâce à l'annotation du corpus ANCOR, nous obtenons la catégorie syntaxique (`TYPE`), la détermination (`DEF`), le genre (`GENRE`), le nombre (`NOMBRE`) et le type d'entité nommée (`EN`) d'une mention.

5.2. Traits relationnels

Les traits relationnels ont pour but de comparer deux mentions d'une paire en observant leur forme, leurs attributs non-relationnels ou leur distance. Deux types de traits peuvent être distingués.

Les traits booléens vérifient généralement l'identité entre les valeurs de traits non relationnels des mentions en question. On peut par exemple citer le trait vérifiant l'accord en genre entre les deux mentions concernées, ou d'autres types d'information tels que l'identité de forme des chaînes de caractères (trait `ID_FORM`), l'inclusion strictement complète et contiguë de la plus petite chaîne de caractère dans la plus grande (`ID_SUBFORM`) ou bien l'inclusion au sens large des items d'une mention dans ceux de l'autre (`EMBEDDED`). Ainsi, si l'on considère l'exemple suivant *m1* = "*le tigre féroce et menaçant*" et *m2* = "*le tigre menaçant*", nous obtiendrons respectivement `ID_FORM = FALSE, ID_SUBFORM = FALSE et EMBEDDED = TRUE`.

Les autres traits sont décrits par un nombre entier ou réel. Il s'agit pour la plupart de distances spatiales ou lexicales telles que les distances dans le texte entre deux mentions en termes de nombres de mots (`DISTANCE_WORD`), de caractères (`DISTANCE_CHAR`), de mentions (`DISTANCE_MENTION`) ou bien même de tours de paroles (`DISTANCE_TURN`). Mais il peut également concerner des attributs qui apportent de nouvelles informations lexicales à savoir `INCL_RATE` et `COM_RATE` qui indiquent respectivement le taux d'inclusion de tokens et le taux de tokens communs. Appliqués à l'exemple précédent, nous obtenons `INCL_RATE = 1` et `COM_RATE = 2/5`.

6. Expérimentations

6.1. Recherche du meilleur classifieur

Les premières expériences que nous avons menées ont consisté à étudier la complexité du problème considéré en comparant le niveau de performances obtenues, après entraînement sur l'ensemble du corpus d'apprentissage, par un classifieur linéaire et un classifieur polynomial. Ces expérimentations ont été réalisées avec l'ensemble d'apprentissage *MediumTrainingSet* et les trois ensembles de test. Le tableau 4 contient la moyenne des F-mesures obtenues sur les trois sous-corpus. Nous évaluons ici la classification brute, c'est-à-dire la qualité de la classification `COREF/NOT_COREF` de chaque paire de mention, et non pas l'identification des chaînes (ou ensembles) de coréférence finales.

Techniques d'apprentissage	Moyenne de la f-mesure sur les trois TestSet_*i*
SVM polynomial (librairie SMO)	0,924
SVM linéaire (librairie LibSVM)	0,917

TABLE 4 : F-mesure en classification pure des SVM polynomial (SMO) et linéaire (LibSVM)

Le SVM polynomial dépasse légèrement SVM linéaire, avec un écart de 0,007 de f-mesure. Cette différence est toutefois significative d'un point de vue statistique (test de Wilcoxon-Mann-Withney : $Z_{inv} = 0{,}0633 < 0{,}1$). La faible amplitude de cette amélioration des performances pourrait laisser à penser que le problème de classification que nous considérons reste relativement simple. Il faut toutefois noter que ces résultats ont été obtenus sans recherche d'une optimisation du modèle polynomial construit. Par ailleurs, nous étudions ici la classification pure, qui donnera a priori des différences plus sensibles en termes d'identification finale des ensembles de coréférence complets. Nous avons donc considéré les performances du SVM polynomial suffisamment élevées pour le choisir comme représentant des séparateurs à vaste marge pour la suite.

	Small	Medium	Big
j48 (arbres de décision)	0,938	**0,946**	0,943
Naïve Bayes	**0,890**	0,887	0,873
SMO (SVM polynomial)	0,920	0,924	**0,929**
Moyenne	*0,916*	*0,919*	*0,915*

TABLE 5 : Influence du corpus d'apprentissage sur les performances en classification pure

Dans un second temps, nous avons étudié l'impact de l'équilibre entre instances positives et négatives lors de l'apprentissage, ceci avec les trois types classifieurs. La comparaison a donc porté sur les trois corpus d'apprentissage *Small-, Medium- et BigTrainingSet*). Le tableau 5 donne les performances, toujours en F-mesure de classification pure, de chaque classifieur. On constate que le ratio entre exemples positifs et négatifs a une influence modérée. Le corpus *MediumTrainingSet* est celui qui permet les meilleures performances en moyenne, mais ce résultat varie suivant le classifieur. C'est cet équilibrage moyen que nous conserverons dans le reste de notre étude, tout en relevant le peu d'impact de cette variable. Par ailleurs, nous constatons que j48 semble être le meilleur modèle avec une F-mesure de 0,946. À l'inverse, le faible résultat de Naive Bayes nous contraint à le délaisser, au profit du SVM qui reste relativement proche de j48 avec 0,929 de F-mesure maximale. Ces résultats confirment nos intuitions, fondées à la fois sur (Desoyer et al., 2015) et les limites connues des classifieurs bayésiens. Ils nous laissent penser que le corpus d'ANCOR présente une taille suffisante pour répondre aux besoins de l'apprentissage sur cette tâche, un classifieur bayésien étant connu pour être moins sensible au manque de données.

Nous avons enfin cherché à confirmer ces résultats sur les performances de résolution des chaînes (ou ensembles) complètes de coréférence. Nous avons pour cela utilisé les métriques MUC (Vilain

et al., 1995) et B^3 ((Bagga and Baldwin, 1998). Les ensembles de mentions coréférentes sont obtenus, dans toutes nos expériences, suivant une stratégie de type *best-first* à partir des résultats de la classification par paire : si une mention a plusieurs antécédents potentiels, on la place dans le premier ensemble de coréférence donné. La table 6 détaille les performances obtenues. On retrouve les résultats précédents sur la hiérarchie des classifieurs. De même, l'impact de l'équilibrage positif/ négatif reste limité.

Métrique	Classifieur	Small	Medium	Big
MUC	j48	0,881	0,880	**0,891**
	NB	0,845	*0,854*	0,847
	SVM	0,854	0,859	*0,869*
	Moyenne	*0,860*	*0,864*	*0,869*
B^3	j48	**0,884**	0,874	0,876
	NB	0,855	0,847	*0,857*
	SVM	0,857	0,860	*0,864*
	Moyenne	*0,865*	*0,860*	*0,866*

TABLE 6 : Influence du corpus d'apprentissage sur les performances de j48 en résolution

6.2. Influence de l'interactivité des corpus

Le degré interactivité est très variable entre différentes situations de dialogue spontané, et peut conduire à des manifestations différentes de coréférence. Il nous est apparu important de vérifier dans quelle mesure les performances étaient influencées par ce degré d'interactivité. Pour cela, nous avons distingué dans les corpus d'apprentissage et de test des sous-corpus spécifiques à un degré d'interactivité donné (fort : OTG ou UBS ou faible : ESLO). Nous avons alors cherché à étudier si l'apprentissage de modèles spécifiques à chaque degré d'interactivité (suivant une approche par adaptation de modèles) ne pourrait pas conduire à de meilleurs performances qu'un modèle générique appris sur tout le corpus *MediumTrainingSet*,. Par exemple, pour un test en situation très interactive, nous distinguons :

- Modèle général - Apprentissage sur l'ensemble du corpus *MediumTrainingSet*,
- Adaptation de modèle - Apprentissage sur la sous-partie OTG de *MediumTrainingSet*

Nous avons étudié l'ensemble des différentes combinaisons *train*/*test* possibles. Les tables 7 et 8 décrivent les résultats obtenus respectivement par j48 et SMO dans quatre situations prototypiques. Nous avons rejoué 4 fois les expériences en changeant aléatoirement les instances négatives d'apprentissage, à fin d'étude en significativité statistique. L'évaluation porte ici sur la résolution complète (MUC et B^3), les résultats présentés étant cohérents avec ceux en classification pure.

Métrique de test	MUC		B^3	
Test (interactivité)	**UBS+OTG** (forte)	**ELSO** (faible)	**UBS+OTG** (forte)	**ELSO** (faible)
Modèle général	0,864 (σ = 0,007)	0,734 (σ = 0,010)	0,882 (σ = 0,006)	0,897 (σ = 0,004)
Modèle adapté	0,860 (σ = 0,003) (*train : OTG*)	0,742 (σ = 0,023) (*train : ESLO*)	0,884 (σ = 0,003) (*train : OTG*)	0,902 (σ = 0,009) (*train : ESLO*)

TABLE 7 : Influence du corpus d'apprentissage sur les performances en classification pure

Les observations montrent que l'impact du degré d'interactivité est réel. On observe ainsi une baisse de performances avec la mesure MUC entre les situations de faible interactivité et celles de forte interactivité. Avec j48, cette différence est statistiquement significative pour le modèle général ($|T| = 1,995 > T(0,1) = 1,983$) mais pas pour le modèle adapté ($|T| = 1,464 < T(0,1) = 1,983$). On remarque que la baisse observée avec MUC sur le corpus faiblement (ESLO) interactif est à l'opposé des résultats obtenus avec la mesure B^3 (où les différences ne sont pas significatives). On sait qu'une limitation de la mesure MUC est de ne pas prendre en considération les singletons (mentions ne faisant pas partie d'une chaîne de coréférence), contrairement à B^3. Une étude qualitative du comportement des modèles sur les singletons devra être menée pour expliquer ces résultats. Retenons pour l'heure que le degré d'interactivité peut influencer sur le comportement des systèmes.

On note ensuite que les modèles généraux entraînés sur l'ensemble du corpus *MediumTrainingSet* atteignent un niveau de performance équivalent à celui des modèles adaptés. Prenons l'exemple de j48. Dans le cas du différentiel de performances maximal observé avec ce classifieur (mesure MUC avec test sur ESLO : 0,734 contre 0,742), un test de Student donne une absence de toute différence statistiquement significative entre le modèle général et le modèle adapté : $|T| = 0,089 << T(0,1) = 1,983$. Les classifieurs généraux appris sur *MediumTrainingSet* semblent donc s'adapter par eux-mêmes à la diversité du degré d'interaction. Les gains de performance obtenus avec les modèles adaptés sont trop restreints pour justifier la construction de classifieurs spécifiques à chaque situation interactive. Pour cette raison, la suite de nos travaux concernera donc toujours des modèles génériques appris.

Métrique de test	MUC		B^3	
Test (interactivité)	**UBS+OTG** (forte)	**ELSO** (faible)	**UBS+OTG** (forte)	**ELSO** (faible)
Modèle général	0,847 (σ = 0,012)	0,729 (σ = 0,031)	0,868 (σ = 0,011)	0,992 (σ = 0,006)

Modèle adapté	0,845 (σ = 0,033)	0,717 (σ = 0,017)	0,893 (σ = 0,013)	0,897 (σ = 0,007)

TABLE 8 : Influence du corpus d'apprentissage sur les performances en résolution (j48)

6.3. Classifieurs spécifiques et multi-classifieur

Afin d'améliorer le niveau de performance de notre système, nous sommes partis de l'idée que les traits exploités n'étaient pas les mêmes en fonction du type de relation (e.g directe, indirecte et anaphore pronominale). Par exemple, dans un arbre de décision j48, le premier trait mis en valeur pour une relation directe sera `INCL_RATE` et `m2_TYPE` pour une relation indirecte. Nous avons alors cherché à savoir si la construction de classifieurs pour chaque type de coréférence ne pourrait pas conduire à une adaptation optimale des modèles sur chaque ensemble de traits d'apprentissage.

	Directe	**Indirecte**	**Anaphore**
F-mesure	0,973	0,969	0,965

TABLE 9 : F-mesure en classification pure (j48) des classifieurs spécifiques par relation

L'idée est donc de créer un classifieur spécifique pour chaque type de relation (i.e un classifieur de relation directe répondra uniquement `DIRECTE` ou `NOT_DIRECTE`). La table 9 semble indiquer que cette spécialisation est bénéfique, puisque les niveaux de performances en classification pure de chaque classifieur (expérience menée avec j48) sont très satisfaisants. Nous construisons ensuite un multi-classifieur, c'est-à-dire un système qui utilise les réponses de chaque classifieur comme vote pour la décision finale coréférente/non coréférente. Le système de vote n'est pas majoritaire : si l'un d'eux renvoie une réponse positive (i.e DIRECTE, INDIRECTE ou ANAPHORE) alors la relation est considérée coréférente. À l'inverse, elle sera jugée non-coréférente si tous renvoient une réponse négative (i.e NOT_DIRECTE, NOT_INDIRECTE, NOT_ANAPHORE). Les f-mesures obtenues par j48 de ce multi-classifieur sont comparées dans la table 10 avec celles du classifieur général original. Bien que parfois proches, les résultats de notre multi-classifieur demeurent inférieurs à ceux de celui de base. Il s'agit ici de travaux préliminaires qui demandent à être poursuivies avec le SVM.

	Classifieur de base	**Multi-classifieur**
Moyenne du testSet 1	0,9383	0,9333
Moyenne du testSet 2	0,9406	0,935
Moyenne du testSet 3	0,9353	0,933

TABLE 10 : F-mesure (classification pure) du classifieur originel et du multi-classifieur

7. Conclusion et perspectives

Dans cet article, nous avons présenté nos premières recherches visant à construire un système de résolution des coréférences basé sur des techniques de classification binaire (coréférent/non coréférent). Nous avons étudié une stratégie de résolution par multi-classifieur qui n'a pas été expérimentée à notre connaissance sur le français. Les résultats obtenus restent perfectibles, en particulier, nous poursuivons actuellement nos expérimentations sur l'ensemble des facteurs (choix des échantillons négatifs, optimisation des paramètres des classifications) pouvant influencer les paramètres. De même, il nous reste à élaborer un système de bout en bout travaillant sur des corpus bruts, en intégrant un détecteur de mentions développé au LATTICE (Grobol et al. 2017). Nous comptons par ailleurs poursuivre un travail d'ingénierie fine sur les traits d'apprentissage, tout en évaluant notre système sur les métriques CEAF (Luo 2005) et BLANC (Recassens & Hovy 2011).

Un de nos objectifs est de développer une *baseline* solide pour challenger les techniques d'apprentissage profond. Nous nous demandons en effet si, sur une tâche complexe comme la coréférence, l'intérêt des techniques neuronales réside sur leur niveau réel de performances ou sur le fait qu'elles dédouanent le chercheur d'un travail fastidieux d'ingénierie sur les traits linguistiques. Des études récentes montrent en effet que les techniques d'apprentissage présentent des performances très perfectibles (Durett & Klein, 2013) sur des tâches complexes telles que la résolution des coréférences pronominales ambigües ou les schémas de Winograd (Morgenstern et al., 2016), et que les approches neuronales n'ont pas surmonté cette difficulté.

Par ailleurs, l'objectif applicatif de nos travaux est de se focaliser sur la coréférence indirecte, dans une perspective de détection des variantes de nomination en analyse du discours. Il est à craindre qu'une approche neuronale puisse manquer de données d'apprentissage dans le cas des nominations, expressions en émergence et encore non figées dans la langue. Ceci reste à vérifier.

Remerciements

Ce travail s'inscrit dans le cadre de différents stages de fin de licence, encadrés par Jean-Yves Antoine, Anaïs Lefeuvre-Halftermeyer, Nicolas Labroche, Sylvie Billot et Marcilio de Souto que les auteurs tiennent à remercier. Cette recherche s'insère également dans le programme « Investissements d'Avenir » géré par l'Agence Nationale de la Recherche ANR-10-LABX-0083 (Labex EFL).

Références

BAGGA A., BALDWIN B. (1998). Algorithms for scoring coreference chains. *Proc. of the LREC Workshop on Linguistic Coreference*, pp. 563–566, Granada, Spain.

BOSER B. E., GUYON I., VAPNIK V. (1992) A training algorithm for optimal margin classifiers. *Proc. of the Fifth Annual Workshop on Computational Learning Theory*, ACM Press.

CLARK K., Manning C. D. (2016b). Improving coreference resolution by learning entity level distributed representations. In Association for Computational Linguistics (ACL).

DÉSOYER A., LANDRAGIN F., TELLIER I., LEFEUVRE A., ANTOINE J-Y. (2015). Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR. *Traitement Automatique des Langues*, *TAL*, vol. 55(2), pp.97-121.

DURRET G., KLEIN D. (2013) Easy victories and uphill battles in coreference resolution. *Proc. EMNLP'2013*.

GODBERT E., FAVRE B. (2017). Détection de coréférences de bout en bout en français. *Actes TALN'2017*, Orléans, Juin 2017.

GROBOL L., TELLIER I., DE LA CLERGERIE E., DINARELLI M., LANDRAGIN F. (2017) Apports des analyses syntaxiques pour la détection automatique de mentions dans un corpus de français oral. *Actes TALN 2017*, Orléans, France.

KABDJOV M., STEPANOV, J. (Eds.) (2016) The SENSEI Discourse Analysis Tools. *Rapport Technique SENSEI D4.2*.

LAPPIN S., LEASS H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20, pp. 535–561.

LEE K., HE L., LEWIS M., ZETTLEMOYER L. (2017) End-to-end neural coreference resolution. *Proc. EMNLP'2017*.

LONGO L., TODIRASCU A. (2009). Une étude de corpus pour la détection automatique des thèmes, *Actes des 6èmes journées de linguistique de corpus*, Lorient. 10-12 septembre 2009.

LUO X. (2005). On coreference resolution performance metrics. *Proc. HLT-EMNLP 2005*, pp. 25–32, Vancouver, Canada.

MITKOV R. (2002). *Anaphora resolution*. Longman.

MORGENSTERN L., DAVIS E., ORTIZ C.L. (2016) Planning, executing and evaluating the Winograd Schema Challenge. *AI Magazine*, 37(1). Pp. 50-54.

MUZERELLE J., LEFEUVRE A., SCHANG E., ANTOINE J-Y., PELLETIER A., MAUREL D., IESHKOL I., VILLANEAU J. (2014). ANCOR_CENTRE, a large free spoken French coreference corpus : description of the resource and reliability measures. *Proc. LREC'2014*, Reykjavik, Islande.

NG V., CARDIE C. (2002). Improving machine learning approaches to coreference resolution. In Proceedings of the ACL. *Proc.* of the *ACL'02*. pp. 104-111.

PLATT J. (1998) Fast Training of Support Vector Machines using Sequential Minimal Optimization. In B. Schoelkopf and C. Burges and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*.

QUINLAN J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.

RECASENS, M. HOVY, E. (2011). BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(04), pp. 485 - 510

SOON W., NG H., LIM D. (2001). A Machine Learning Approach to Coreference of Noun Phrases. *Computational Linguistics*, 27(4), 521-554.

VILAIN M., BURGER J., ABERDEEN J., CONNOLLY D., HIRSCHMAN L. (1995). A model-theoretic coreference scoring scheme. *Proc. MUC-6 Conference*, pp. 45–52

WIDLÖCHER A., MATHET Y. (2009). La plate-forme Glozz : environnement d'annotation et d'exploration de corpus. *Actes TALN'2009*. Senlis, France.

WISEMAN S., RUSH A.M., SHIEBER S.M. (2016) Learning global features for coreference resolution. *Proc. Human Language Technology and North American Association for Computational Linguistics, HLT-NAACL'2016*.

WITTEN I.H., EIBE F., HALL M.A. (2011). *Data Mining: Practical machine learning tools and techniques*,3e édition, Morgan Kaufmann.