Apprentissage automatique d'un modèle de résolution de la coréférence à partir de données orales transcrites du français : le système CROC

Adèle Désoyer[1, 2] Frédéric Landragin[1] Isabelle Tellier[1]
(1) Lattice, CNRS, ENS et Université de Paris 3 - Sorbonne Nouvelle
(2) MoDyCo, CNRS, Université Paris Ouest - Nanterre La Défense
adele.desoyer@gmail.com, frederic.landragin@ens.fr, isabelle.tellier@univ-paris3.fr

Résumé. Cet article présente CROC [1] (*Coreference Resolution for Oral Corpus*), un premier système de résolution des coréférences en français reposant sur des techniques d'apprentissage automatique. Une des spécificités du système réside dans son apprentissage sur des données exclusivement orales, à savoir ANCOR (anaphore et coréférence dans les corpus oraux), le premier corpus de français oral transcrit annoté en relations anaphoriques. En l'état actuel, le système CROC nécessite un repérage préalable des mentions. Nous détaillons les choix des traits – issus du corpus ou calculés – utilisés par l'apprentissage, et nous présentons un ensemble d'expérimentations avec ces traits. Les scores obtenus sont très proches de ceux de l'état de l'art des systèmes conçus pour l'écrit. Nous concluons alors en donnant des perspectives sur la réalisation d'un système *end-to-end* valable à la fois pour l'oral transcrit et l'écrit.

Abstract.
Machine Learning for Coreference Resolution of Transcribed Oral French Data : the CROC System
We present CROC (Coreference Resolution for Oral Corpus), the first machine learning system for coreference resolution in French. One specific aspect of the system is that it has been trained on data that are exclusively oral, namely ANCOR (ANaphora and Coreference in ORal corpus), the first corpus in oral French with anaphorical relations annotations. In its current state, the CROC system requires pre-annotated mentions. We detail the features that we chose to be used by the learning algorithms, and we present a set of experiments with these features. The scores we obtain are close to those of state-of-the-art systems for written English. Then we give future works on the design of an end-to-end system for oral and written French.

Mots-clés : corpus de dialogues, détection de coréférences, apprentissage, paires de mentions.

Keywords: Dialogue corpus, Coreference resolution, Machine learning, Mention-pair model.

1 Introduction

Depuis les vingt dernières années, la reconnaissance automatique des chaînes de coréférence représente un objet d'étude à part entière du TAL, au cœur de grandes campagnes d'évaluation telles que celles proposées par MUC (*Message Understanding Conference* [2]), ACE (*Automatic Content Extraction* [3]), SemEval (*Semantic Evaluation* [4]) ou CoNLL (*Computationnal Natural Language Learning* [5]). Ces chaînes constituent une unité discursive complexe qui contribue à la cohésion du discours. Les identifier automatiquement oblige à prendre en compte la séquence des phrases qui le composent, et leurs relations. Ce domaine a donné lieu à de nombreux travaux, mais les données sur lesquelles ils se sont fondés (issues des campagnes précédemment citées) étaient jusqu'à présent essentiellement de l'anglais écrit. Les travaux présentés dans cet article ont la particularité de se concentrer sur la reconnaissance automatique de chaînes de coréférence présentes dans de

1. Présenté plus en détails dans (Désoyer *et al.*, 2015)
2. Voir notamment la tâche sur la coréférence dans MUC-7 en 1998, cf. http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.
3. Cf. http://www.itl.nist.gov/iad/mig//tests/ace/.
4. Voir notamment la tâche sur la coréférence dans SemEval-2 en 2010, cf. http://semeval2.fbk.eu/semeval2.php?location=tasks.
5. Voir notamment la tâche sur la coréférence en 2011 et 2012, cf. http://conll.cemantix.org/2011/ et http://conll.cemantix.org/2012/.

l'oral transcrit français. Cette modalité rend-elle les coréférences plus ou moins fréquentes, explicites et faciles à repérer ? Ce sera tout l'enjeu des expériences que nous avons menées.

Commençons par définir précisément notre objet d'étude. Les chaînes de coréférence s'appuient sur la notion plus restreinte d'anaphore. Cette dernière décrit une procédure référentielle regroupant les phénomènes de renvoi à un antécédent du discours immédiat. Les anaphores sont des relations asymétriques entre un antécédent et une expression anaphorique qui ne peut être interprétée qu'à partir de son antécédent. Dans l'exemple de la figure 1, extrait d'un dialogue du corpus OTG d'ANCOR, l'anaphorique nominal « le nom » ne peut être interprété qu'à partir de son antécédent « une grande libraire » ; il en va de même pour le pronom « elle ».

— on m'a parlé d'**une grande libraire** mais on se rappelle plus **le nom**
— **Arthaud**
— **Arthaud** peut-être
— **elle** se trouve dans le centre ville

FIGURE 1 – Mise en évidence d'une chaîne de coréférence dans un dialogue

Le phénomène plus large de la coréférence se décrit, quant à lui, comme la relation existant entre plusieurs expressions référant à une même entité. Contrairement à l'anaphore qui distinguait strictement ses deux parties, la relation de coréférence est symétrique. Dans l'exemple de la figure 1, l'ensemble des expressions en gras composent une chaîne de coréférence.

Les systèmes de résolution automatique de la référence sont encore aujourd'hui extrêmement rares s'agissant du français, et même, à notre connaissance, inexistants – à l'exception de systèmes à base de règles tels que celui décrit dans (Trouilleux, 2001) ou RefGen présenté dans (Longo, 2013). Sans recourir intégralement aux modèles statistiques, la méthode hybride de RefGen allie la performance de ceux-ci à la pertinence des modèles symboliques avec un premier module de segmentation probabiliste, puis un module linguistique repérant les marqueurs de cohésion. L'apprentissage automatique n'avait pu encore être mis en œuvre dans ce contexte, faute jusqu'à présent de corpus annotés et disponibles librement. Notre travail repose sur les données du corpus ANCOR [6], réalisé dans le cadre du projet du même nom financé par la région Centre, qui annote différentes formes de reprise dans le discours : la reprise fidèle (représentant 40 % des reprises annotées) est la relation qui unit deux groupes nominaux référant à la même entité du discours, et qui ont la même tête nominale ; la reprise infidèle est plus rare (7 % des reprises) et se distingue de la précédente par le fait que les deux mentions qu'elle relie ont des têtes nominales distinctes souvent proches sémantiquement (par synonymie, hyperonymie ou hyponymie). La relation la plus représentée dans le corpus (42 % des reprises) est l'anaphore pronominale, c'est-à-dire la reprise d'un groupe nominal par un pronom. Les deux dernières relations représentées relèvent de l'anaphore associative, c'est-à-dire que les deux expressions en relation réfèrent à des entités distinctes, qui entrent par exemple dans une relation méronymique. Parmi ces phénomènes, l'annotation d'ANCOR distingue l'anaphore associative nominale, reliant deux syntagmes nominaux (10 % des reprises) et l'anaphore associative pronominale, reliant un syntagme nominal à un pronom (1 % des reprises). Ce corpus global de 488 000 mots se décompose en fait en quatre sous-corpus issus de précédents projets de recherche : deux d'entre eux sont extraits du projet ESLO et sont essentiellement composés d'interviews, faiblement interactives (CO2, qui représente 8 % des données, et ESLO qui correspond à 85 % des données) ; les deux autres ensembles sont composés de dialogues interactifs, avec d'une part l'office du tourisme de Grenoble (corpus OTG, représentant 5 % des données), d'autre part le standard téléphonique de l'Université de Basse-Normandie (corpus UBS, représentant 2 % des données).

2 Résolution de la coréférence comme tâche de classification

2.1 État de l'art

Les premiers systèmes de résolution automatique de la coréférence traitent la tâche de façon symbolique, avec des règles écrites à la main. Dans les années 1970, la problématique est limitée à la résolution des anaphores pronominales, avec une mise en avant et des pistes pour calculer la saillance (Lappin & Leass, 1994). Elle débouche sur des approches dites *knowledge-poor*, à l'instar de celle de (Mitkov, 2002) qui, pour seul prétraitement, nécessite une analyse morphosyntaxique et un découpage en *chunks*. Ces approches ont rapidement trouvé leurs limites et l'essor de la linguistique de

6. http://tln.li.univ-tours.fr/Tln_Ancor.html.

corpus a encouragé l'exploitation de données attestées. Les efforts de recherche actuels se concentrent désormais sur des approches fondées sur l'apprentissage supervisé, ce qui nécessite deux prérequis : reformuler l'identification d'une chaîne de coréférence comme une tâche que l'on sait aborder par de telles méthodes (par exemple une tâche de classification ou d'annotation) ; disposer d'un corpus annoté servant à la fois pour l'apprentissage et le test, afin d'évaluer les performances du système ainsi construit. Différents types d'approches s'opposent quant à la façon de formuler la tâche confiée aux algorithmes d'apprentissage, parmi lesquels :

– les modèles *mention-pair* ou *pairwise* qui sont fondés sur une classification binaire comparant une anaphore à des antécédents potentiels situés dans les phrases précédentes. Concrètement, les exemples fournis au programme sont des paires de mentions (une anaphore et un antécédent potentiel) pour lesquelles l'objectif est de déterminer si elles sont coréférentes ou non. Une anaphore ne pouvant avoir qu'un unique antécédent, une deuxième phase doit déterminer quel est le véritable antécédent de l'anaphore parmi tous ceux qui sont possibles (à l'intérieur des paires de mentions classées comme coréférentes). Différents systèmes ont été implémentés pour cela, parmi lesquels ceux de (Soon *et al.*, 2001) et (Ng & Cardie, 2002) et plus récemment ceux de (Bengtson & Roth, 2008) et (Stoyanov *et al.*, 2010), régulièrement utilisés comme systèmes de référence à partir desquels les nouveaux systèmes comparent leur performance. Pour les premiers, l'antécédent sélectionné parmi un ensemble pour une anaphore donnée est celui qui en est le plus proche. Il s'agit d'un regroupement dit *Closest-First* qui, pour chaque anaphore, parcourt l'ensemble du texte vers la gauche, jusqu'à trouver un antécédent ou atteindre le début du texte. Les seconds proposent une alternative à cette approche, dit regroupement *Best-First*, qui sélectionne comme antécédent celui ayant le plus haut score de « probabilité coréférentielle » parmi l'ensemble des précédentes mentions. L'inconvénient de ce type de méthode est la réduction du problème à une série de classifications binaires indépendantes, qui ne prend pas en compte l'ensemble des différents maillons d'une même chaîne de coréférence ;
– les modèles *twin-candidate*, proposés dans (Yang *et al.*, 2003) considèrent également le problème comme une tâche de classification, mais dont les instances sont cette fois composées de trois éléments (x, y_i, y_j) où x est une anaphore et y_i et y_j deux antécédents candidats (y_i étant le plus proche de x en termes de distance). L'objectif du modèle est d'établir des critères de comparaison des deux antécédents pour cette anaphore, et de classer l'instance en FIRST si le bon antécédent est y_i et en SECOND si le bon antécédent est y_j. Cette classification alternative est intéressante car elle ne considère plus la résolution de la coréférence comme l'addition de résolutions anaphoriques indépendantes, mais prend en compte l'aspect « concurrentiel » des différents antécédents possibles pour une anaphore ;
– les modèles *mention-ranking*, tels celui décrit dans (Denis, 2007), envisagent non plus d'étiqueter chaque paire de mentions mais de classer l'ensemble des antécédents possibles pour une anaphore donnée selon un processus itératif qui compare successivement cette anaphore à deux antécédents potentiels : à chaque itération, on conserve le meilleur candidat, puis on forme une nouvelle paire de candidats avec ce « gagnant » et un nouveau candidat. L'itération s'arrête lorsqu'il n'y a plus de candidat possible. Une alternative à cette méthode propose de comparer simultanément tous les antécédents possibles pour une anaphore donnée ;
– les modèles *entity-mention* (Yang *et al.*, 2008) déterminent quant à eux la probabilité qu'une expression réfère à une entité ou à une classe d'entités précédemment considérées comme coréferentes (*i.e.* un candidat est comparé à un unique antécédent ou à un cluster contenant toutes les références à une même entité).

Enfin, parmi les travaux les plus récents, certains cherchent à concilier les avantages de chaque méthode, y compris celles à base de règles, en distinguant plusieurs strates de résolution de manière à optimiser les performances en fonction des phénomènes ciblés par chaque strate (Lee *et al.*, 2013).

2.2 Reformulation du problème en tâche de classification

CROC est fondé sur une classification binaire opérant sur des paires d'unités référentielles, pour les ranger soit dans la classe des mentions coréférentes, soit dans celle des mentions non coréférentes. La qualité d'un tel système repose sur les données qui lui sont fournies en apprentissage, et particulièrement sur les traits linguistiques (ou *attributs*) décrivant les unités à classer.

Les systèmes de l'état de l'art procédant de la sorte s'inspirent tous de l'ensemble des traits définis dans (Soon *et al.*, 2001), composé de douze attributs décrivant les propriétés d'un antécédent *i* et d'une reprise potentielle *j* : 1) Distance en nombre de phrases entre *i* et *j* ; 2) *i* est-il un pronom ? ; 3) *j* est-il un pronom ? ; 4) Les chaînes de caractères de *i* et *j* sont-elles égales ? ; 5) *j* est-il un SN défini ? ; 6) *j* est-il un SN démonstratif ? ; 7) *i* et *j* s'accordent-ils en nombre ? ; 8) *i* et *j* s'accordent-ils en genre ? ; 9) *i* et *j* appartiennent-ils à la même classe sémantique ? ; 10) *i* et *j* sont-ils tous deux des noms propres ? ; 11) *i* et *j* sont-ils alias l'un de l'autre ? ; 12) *i* et *j* sont-ils au sein d'une structure appositive ?

(Ng & Cardie, 2002) ajoutent à ces douze traits de référence quarante et un nouveaux, également repris dans les travaux plus récents. Ces attributs se répartissent dans deux familles : les traits non relationnels, qui décrivent une mention d'entité, et les traits relationnels, qui caractérisent la relation unissant les deux mentions d'une paire. Notre propre ensemble reprend ceux-ci et en ajoute certains, en s'adaptant aux spécificités des données dont nous disposons. Ainsi, les traits non relationnels que nous intégrons sont de différents types :
– morphosyntaxiques (correspondant aux traits 2, 3, 5 et 6 de l'ensemble de (Soon *et al.*, 2001)) ;
– énonciatifs (une mention initie-t-elle une chaîne de coréférence ou non ?) ;
– sémantiques (une mention est-elle une entité nommée ? Si oui, de quel type ?).
Quant aux traits relationnels, ils nous permettent de caractériser différentes distances entre un antécédent m_1 et une reprise m_2 potentielle :
– distances lexicales (m_1 et m_2 sont-elles strictement égales ? partiellement égales ?) ;
– distances morphosyntaxiques (m_1 et m_2 s'accordent-elles en genre ? en nombre ?) ;
– distances spatiales (m_1 et m_2 sont séparées de combien de caractères ? de mots ? de mentions ? de tours de paroles ?) ;
– distance syntaxique (l'une des deux mentions est-elle une partie de l'autre ? Autrement dit, si l'une des deux est un syntagme nominal complexe l'autre est-elle une partie de ce syntagme ?) ;
– distances contextuelles (m_1 et m_2 sont-elles précédées du même token ? suivies du même token ?) ;
– distance énonciative (m_1 et m_2 sont-elles produites par le même locuteur ?).
Un total de 30 traits décrivant différents niveaux linguistiques constituent l'ensemble que nous utilisons dans nos séries de test.

3 Expérimentations d'apprentissage et résultats

3.1 Plan d'expérimentations

Différents paramètres entrent en jeu dans la construction du système de détection de chaînes de coréférence ; c'est l'optimisation de la combinaison de ces paramètres qui permettra au système d'améliorer ses performances. Les expérimentations que nous menons ici consistent à générer différents modèles de classification en faisant varier d'une part la représentation des données, d'autre part l'algorithme de calcul du modèle. Afin de mesurer la qualité de chacun des modèles générés sans introduire de biais lié à la dépendance aux données d'apprentissage, nous en distinguons trois ensembles :
– un ensemble d'apprentissage (60 % des données initiales) utile au calcul des différents modèles
– un ensemble de développement (20 % des données initiales) fourni à chacun des modèles générés afin de déterminer celui qui optimise au mieux les paramètres
– un ensemble de test (20 % des données initiales) fourni au système final pour en évaluer ses performances sur de nouvelles données
Afin d'évaluer l'influence de la taille de l'ensemble d'apprentissage sur les résultats de classification, un premier paramètre consiste en la variation du nombre de données utiles au calcul du modèle (trois ensembles sont donc sélectionnés : un réduit dit *small_trainingSet*, un moyen dit *medium_trainingSet* et un grand dit *big_trainingSet*)[7]. Un second paramètre concerne l'ensemble d'attributs décrivant ces données, et de nouveau, différents sont testés : un premier ensemble contient tous les attributs, un second uniquement les attributs relationnels, et un troisième exclut les traits spécifiques de l'oral tels que la correspondance des locuteurs ou la distance en tours de parole. Enfin, nous nous inspirons de l'état de l'art pour tester trois algorithmes distincts : les arbres de décision, les SVM et Naive Bayes tels qu'implémentés dans la plate-forme Weka, avec leurs paramètres par défaut.

Le plan d'expérimentations mis en place consiste à combiner les données des différents corpus d'apprentissage avec les trois ensembles d'attributs produits, puis à fournir chacune de ces représentations de données aux trois algorithmes d'apprentissage. Chacun des modèles ainsi généré permet alors de classer de nouvelles paires (ensemble de développement) qui sont ensuite filtrées pour ne conserver qu'un unique antécédent pour une anaphore (les résultats de classification peuvent en effet associer une même reprise à différents antécédents). Cette sélection s'appuie sur celle décrite dans les travaux de (Soon *et al.*, 2001) : il s'agit de la stratégie *Closest-First* qui, lorsqu'une mention a plusieurs antécédents possibles, sélectionne le plus proche à gauche pour former une chaîne. Tous les systèmes ainsi générés seront comparés quantitativement à partir de leurs résultats chiffrés.

7. Les paires de mentions sont sélectionnées aléatoirement au sein du corpus, sans distinguer les sous-corpus dans lesquels elles apparaissent, et sont réparties telles que : 71 881 instances dont 11 908 paires coréférentes et 59 973 non coréférentes dans *small_trainingSet* ; 101 919 instances dont 17 844 paires coréférentes et 84 075 non coréférentes dans *medium_trainingSet* ; 142 498 instances dont 24 620 paires coréférentes et 117 878 non coréférentes dans *big_trainingSet*.

System	Langue	Corpus	MUC	B^3	CEAF	BLANC
		Systèmes end-to-end				
(Soon *et al.*, 2001)	ANGLAIS	MUC-7	60.4	-	-	-
(Ng & Cardie, 2002)	ANGLAIS	MUC-7	63.4	-	-	-
(Stoyanov *et al.*, 2009)	ANGLAIS	ACE-2003	67.9	65.9	-	-
(Stoyanov *et al.*, 2010)	ANGLAIS	MUC-7	62.8	79.4	-	-
(Haghighi & Klein, 2010)	ANGLAIS	ACE-2004	67.0	77.0	-	-
(Lassalle, 2015)	ANGLAIS	CoNNL-2012	68.8	54.56	50.20	-
(Longo, 2013)	FRANCAIS	MULTI-GENRES	36	69.7	55	59.5
		Systèmes pré-annotés				
(Yang *et al.*, 2003)	ANGLAIS	MUC-7	60.2	-	-	-
(Luo *et al.*, 2004)	ANGLAIS	ACE-2	80.7	77.0	73.2	77.2
(Denis & Baldridge, 2008)	ANGLAIS	ACE-2	71.6	72.7	67.0	-
(Bengtson & Roth, 2008)	ANGLAIS	ACE-2004	75.1	80.8	75.0	75.6
CROC	FRANCAIS	ANCOR	63.45	83.76	79.14	67.43

TABLE 1 – Résultats de systèmes de résolution de la coréférence

3.2 Évaluations

La tâche de la résolution de la coréférence est traditionnellement évaluée selon quatre métriques :

– MUC (dont le nom est issu de la campagne d'évaluation *Message Understanding Conference*) se concentre sur l'évaluation des liens de coréférence qui sont communs à l'ensemble des chaînes avérées et à celui des chaînes prédites par le système.
– B^3 : (Bagga & Baldwin, 1998) considèrent comme unité de base la mention plutôt que le lien.
– CEAF, développé dans les travaux de (Luo, 2005), est fondée sur l'entité, c'est-à-dire la référence commune à tous les maillons d'une chaîne de coréférence (le nom complet de la mesure est *Constrained Entity Aligned F-Measure*).
– BLANC (pour *BiLateral Assessment of Noun-phrase Coreference*) est la métrique la plus récente mise au point dans les travaux de (Recasens, 2010), dont la vocation est de considérer conjointement les liens de coréférence et de non-coréférence.

À titre de repères, le haut du tableau 1 présente les résultats de certains des systèmes *end-to-end* (c'est-à-dire sans connaître *a priori* les positions des mentions dans le corpus de test) les plus connus à ce jour pour la langue anglaise. Les résultats du système RefGen de (Longo, 2013), correspondant à la moyenne des scores obtenus pour différents genres textuels, sont également présentés dans ce tableau 1. Par ailleurs, les résultats des systèmes non *end-to-end* qui, comme le nôtre, basent leurs expérimentations sur de *vraies* mentions, sont présentés en bas de tableau 1 (à noter que tous sont conçus pour l'anglais écrit standard).

Les résultats obtenus sur l'ensemble de développement démontrent nettement que NaiveBayes n'est pas idéal pour la résolution de la coréférence. C'est dans la majorité des cas les systèmes appris par l'algorithme SVM sur les données du plus petit ensemble d'apprentissage qui présentent les meilleurs résultats sur de nouvelles données (*i.e.* l'ensemble de développement). Concernant l'ensemble de traits, le plus pertinent semble être le plus exhaustif puisque c'est en l'incluant que les modèles sont meilleurs dans la grande majorité des systèmes. Le fait de ne pas prendre en compte les deux traits spécifiques de l'oral est apparemment assez peu pénalisant. Mais ces résultats seraient certainement à nuancer si nous avions distingué les différents sous-corpus, qui varient quant à leur degré d'interactivité, lors de l'apprentissage. Notamment, la distance en tours de parole entre deux mentions successives est sans doute différente d'un sous-corpus à un autre, et utiliser ce trait en mélangeant les sous-corpus restreint son impact.

Le modèle finalement intégré au système de résolution est donc celui calculé par SVM sur l'ensemble complet d'attributs et le corpus d'apprentissage *small_trainingSet*, puisque la moyenne de ses quatre métriques obtenues lors de la phase de développement est la plus haute (71,9). Les résultats de ce système sur l'ensemble de test forment une moyenne de 73,4 et sont présentés plus en détails en fin de tableau 1.

De manière générale, il est extrêmement délicat de comparer nos résultats à ceux déjà connus, tant les données varient : la langue des corpus, la modalité (oral *vs* écrit), le codage des coréférences spécifiques au corpus, les traits disponibles, etc., sont différents. CROC se veut plutôt une base nouvelle, à laquelle pourront se comparer les futurs autres systèmes

de reconnaissance de la coréférence dédiés au français. A titre d'exemple, sur l'énoncé extrait du sous-corpus OTG d'ANCOR présenté en figure 2, l'annotation manuelle relève deux chaînes de coréférence : une première composée des maillons (2, 4, 5, 7, 8, 9) et une seconde composée des maillons (3, 6). Le système CROC détecte quant à lui trois chaînes de coréférence : une première composée des maillons (1, 8, 9), une seconde composée des maillons (2, 4, 5) et une dernière composée des maillons (3, 6). On constate que l'annotation automatique produit des erreurs de plusieurs types : elle intègre le maillon *le numéro de l'office de tourisme d'Espagne* qui fonctionne en fait comme un singleton, et le considère comme référent des mentions anaphoriques *les* et *ils*. Cette erreur implique la suivante, puisque la chaîne dont le référent est *l'office de tourisme d'Espagne* perd deux de ses maillons, qui ont été précédemment intégrés à une fausse chaîne. De plus, on observe que le maillon *leur* n'est pas considéré par le système, il s'agit donc ici d'une omission.

— je peux donner [*le numéro de* [*l'office de tourisme d'Espagne à* ***Paris***$_3$]$_2$]$_1$ **qui**$_4$ vous enverront tout ce que vous désirez **c'**$_5$ est à **Paris**$_6$ vous **leur**$_7$ écrivez vous **les**$_8$ appelez **ils**$_9$ vous envoient tout

FIGURE 2 – Exemple d'énoncé avec annotation des maillons

Ces premières observations nous amènent à envisager de mettre en place une métrique supplémentaire pour évaluer la qualité de l'annotation en chaînes de coréférence, qui prendrait en compte le nombre d'opérations à effectuer sur les chaînes automatiques pour parvenir à la véritable annotation : il s'agirait d'une sorte de calcul de distance d'édition qui relèverait les différentes substitutions, insertions et délétions de maillons dans les différentes chaînes de coréférence.

4 Conclusion et perspectives

Depuis que la tâche de résolution de la coréférence occupe une place importante dans les problématiques de traitement automatique des langues, beaucoup de travaux se sont attachés à développer des systèmes de détection de chaînes de coréférence pour l'anglais. Très peu cependant, mis à part celui décrit dans (Longo, 2013), ont étudié le phénomène et ses pistes de résolution automatique sur le français. De fait ce travail présente l'un de ces premiers systèmes appris automatiquement sur un corpus annoté. En plus de cette évolution de langue, c'est une distinction de canal qui caractérise ce travail, puisqu'à ce jour aucun modèle de résolution fondé sur l'apprentissage supervisé n'avait été développé spécifiquement pour l'oral transcrit.

Les résultats d'évaluation obtenus par notre modèle de résolution, proches de ceux de certains systèmes de l'état de l'art, suppose toutefois que le texte sur lequel on l'applique est déjà annoté en mentions et que certains attributs de ces mentions (genre, nombre, caractère nouveau ou non de l'entité référencée...) sont disponibles. Ces résultats expérimentaux nous ont néanmoins permis d'observer certaines propriétés du phénomène étudié et d'envisager des pistes de travail pour améliorer les performances du modèle de classification, étendre ses capacités à celles d'un système *end-to-end*, et en compléter l'évaluation. Pour obtenir un tel système *end-to-end* en français, il faudrait coupler CROC avec un étiqueteur POS, un reconnaisseur d'entités nommées et divers autres outils capables d'identifier les genres et nombres des mentions, notamment.

Il est difficile, en l'état actuel de nos expériences, de mesurer l'impact spécifique des différents traits utilisés. Une perspective de ce travail serait de s'attacher précisément à cette sélection d'attributs, par exemple *via* une méthode de sélection ascendante qui évaluerait un modèle appris sur un ensemble ne contenant qu'un trait, puis ajouterait de manière incrémentale un nouveau trait à l'ensemble, en ne le conservant que si les résultats de classification sont meilleurs que pour l'ensemble précédent.

Dans ce travail, la tâche de classification ne permet de ranger les instances que sous deux classes : soit coréférentes, soit non-coréférentes. En procédant ainsi, nous ne distinguons pas les différentes formes de reprises telles qu'annotées dans le corpus ANCOR, et ne prenons pas en compte le fait que chacune d'entre elles est susceptible d'avoir des propriétés qui lui sont propres. Pour l'anaphore pronominale, par exemple, on pourrait supposer que la distance entre les deux mentions d'une paire ne doit pas excéder un certain seuil. Mais ce typage, dans le corpus ANCOR, est toujours réalisé *relativement à la première mention de l'entité*, alors que CROC cherche à identifier les *mentions coréférentes successives*. Une fois la classification des paires de mentions adaptée, on pourra s'inspirer des recherches de (Denis, 2007), qui propose d'apprendre des modèles spécifiques pour chaque type de reprises (anaphore fidèle, infidèle, pronominale et associative).

Références

BAGGA A. & BALDWIN B. (1998). Entity-based cross-document coreferencing using the vector space model. In *Proceedings of ACL'98*, p. 79–85.

BENGTSON E. & ROTH D. (2008). Understanding the Value of Features for Coreference Resolution. In *Proceedings of EMNLP 2010*, p. 236–243.

DENIS P. (2007). *New Learning Models for Robust Reference Resolution*. PhD thesis, University of Texas at Austin.

DENIS P. & BALDRIDGE J. (2008). Specialized models and ranking for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, p. 660–669, Stroudsburg, PA, USA : Association for Computational Linguistics.

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*, **55**(2), 97–121.

HAGHIGHI A. & KLEIN D. (2010). Coreference resolution in a modular, entity-centered model. In *Human Language Technologies : The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, p. 385–393, Stroudsburg, PA, USA : Association for Computational Linguistics.

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

LASSALLE E. (2015). *Structured learning with latent trees : A joint approach to coreference resolution*. PhD thesis, Université Paris Diderot.

LEE H., CHANG A., PEIRSMAN Y., CHAMBERS N., SURDEANU M. & JURAFSKY D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, **39**(4), 885–916.

LONGO L. (2013). *Vers des moteurs de recherche intelligents : un outil de détection automatique de thèmes*. PhD thesis, Université de Strasbourg.

LUO X. (2005). On Coreference Resolution Performance Metrics. In *Proceedings of Human Language Technology - Empirical Methods in Natural Language Processing (EMNLP 2005)*.

LUO X., ITTYCHERIAH A., JING H., KAMBHATLA N. & ROUKOS S. (2004). A mention-synchronous coreference resolution algorithm based on the bell tree. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA : Association for Computational Linguistics.

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

NG V. & CARDIE C. (2002). Improving Machine Learning Approcahes to Corefrence Resolution. In *Proceedings of ACL'02*, p. 104–111.

RECASENS M. (2010). *Coreference : Theory, Resolution, Annotation and Evaluation*. PhD thesis, University of Barcelona.

SOON W. M., NG H. T. & LIM D. C. Y. (2001). A Machine Learning Approach to Coreference Resolution of Noun Phrases. *Computational Linguistics*, **27**(4), 521–544.

STOYANOV V., CARDIE C., GILBERT N., RILOFF E., BUTTLER D. & HYSOM D. (2010). *Reconcile : A Coreference Resolution Research Platform*. Rapport interne.

STOYANOV V., GILBERT N., CARDIE C. & RILOFF E. (2009). Conundrums in noun phrase coreference resolution : Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2 - Volume 2*, ACL '09, p. 656–664, Stroudsburg, PA, USA : Association for Computational Linguistics.

TROUILLEUX F. (2001). *Identification des reprises et interprétation automatique des expressions pronominales dans des textes en français*. PhD thesis, Université Blaise Pascal.

YANG X., SU J., LANG J., TAN C. L., LIU T. & LI S. (2008). An entity-mention model for coreference resolution with inductive logic programming. In *Proc. of ACL'08*, p. 843–851.

YANG X., ZHOU G., SU J. & TAN C. L. (2003). Coreference resolution using competition learning approach. In *Proceedings of ACL'03*, p. 176–183.