Identification automatique de chaînes de coréférences : vers une analyse des erreurs pour mieux cibler l'apprentissage

Frédéric Landragin[1], Bruno Oberle[2]
(1) Lattice, CNRS, ENS Paris, PSL University Research, Univ. Sorbonne Nouvelle, USPC
1 rue Maurice Arnoux, 92120 Montrouge, France,
(2) LiLPa, Univ. de Strasbourg, 14 rue René Descartes, 67084 Strasbourg Cedex, France
frederic.landragin@ens.fr, oberleb@unistra.fr

RESUME

Nous présentons une étude qualitative préliminaire concernant l'analyse linguistique des erreurs commises par des systèmes de détection automatique de chaînes de coréférences. Nous soulignons plusieurs cas de bruit et de silence, caractérisés par des gravités différentes, ainsi que des types d'erreurs spécifiques, notamment la construction de chaînes « fourre-tout » regroupant des expressions référentielles inexploitées par ailleurs. Dans le but de définir une méthodologie généralisable, nous proposons une première typologie d'erreurs et quelques pistes de réflexion pour leur prise en compte à terme dans les processus d'apprentissage, ce qui passe par des considérations sur les types d'hybridation à envisager pour ces processus.

ABSTRACT

Automatic identification of coreference chains: Towards a linguistic analysis of errors in order to improve machine learning features.

We present a preliminary qualitative study dealing with the linguistic analysis of the errors made by NLP systems dedicated to the automatic detection of coreference chains. We describe several cases of noise and silence, characterized with different degrees of importance, as well as coreference-specific types of errors, for instance the construction of "catch-all" chains that group non-used referring expressions. In order to further define a generalizable methodology, we propose a first typology of errors, and some guidelines for their consideration within the machine learning process. This research implies considerations on the possible types of hybrid systems.

MOTS-CLES : Expressions référentielles, coréférences, apprentissage artificiel, traits d'apprentissage, SVM, analyse linguistique des erreurs, diagnostic.

KEYWORDS: Referring expressions, coreferences, machine learning, machine learning features, SVM, linguistic analysis of errors, diagnosis.

1 Introduction : définitions, enjeux et approches

Cet article de réflexion reflète les préoccupations actuelles du projet ANR Democrat (Description et modélisation des chaînes de référence : outils pour l'annotation et le traitement automatique) quant à l'analyse linguistique des erreurs commises par les systèmes d'identification automatique d'expressions référentielles et de chaînes de coréférences, dans des textes tout-venants écrits en langue française contemporaine. Ce projet ANR regroupe des spécialistes de linguistique et de TAL

(Landragin & Schnedecker, 2014 ; Schnedecker *et al.*, 2017), avec l'objectif de collaborations comme celle traitée ici, et nous utiliserons indifféremment la terminologie « chaîne de référence » (utilisée plutôt en linguistique) ou « chaîne de coréférences » (utilisée plutôt en TAL).

Une expression référentielle est un *token* ou suite de *tokens* consécutifs qui permet au lecteur d'avoir accès à une entité extralinguistique ou disponible dans ce que certains appellent « mémoire discursive » (Groupe de Fribourg, 2012), qu'il s'agisse d'un référent humain, animal, concret ou abstrait. Noms propres, groupes nominaux (avec ou sans expansions) et pronoms anaphoriques sont des expressions référentielles typiques. La détection automatique des expressions référentielles d'un texte est donc une tâche qui implique sans s'y réduire la reconnaissance des entités nommées (Nouvel *et al*., 2015), la détection des *chunks* nominaux et celle des pronoms impersonnels – pour les exclure. Enfin, une chaîne de coréférences est un ensemble d'expressions référentielles, qui ont comme propriété commune de toutes référer au même référent extralinguistique (Corblin, 1995 ; Schnedecker, 1997). La détection automatique des chaînes de coréférences est donc une tâche qui implique sans s'y réduire celle de résolution des anaphores pronominales (Mitkov, 2002).

Historiquement, les techniques mises en œuvre pour réaliser de tels logiciels de TAL résident dans des systèmes à base de règles, avec une, deux voire beaucoup plus de phases, par exemple : 1. exclusion des pronoms impersonnels ; 2. repérage des expressions référentielles (en retenant notamment tous les autres pronoms) ; 3. résolution des anaphores pronominales ; 4. construction des chaînes. On peut même envisager une phase par type d'expression référentielle, en reprenant les typologies proposées en linguistique comme celle de (Charolles, 2002). En caricaturant un peu, notons qu'une telle approche est réalisée par des linguistes ou des linguistes-informaticiens, et que la conception des règles se fait empiriquement, en observant à chaque étape les résultats obtenus : *grosso modo*, l'analyse linguistique des erreurs fait pleinement partie du travail des concepteurs.

Ce n'est pas le cas pour les techniques plus récentes fondées sur l'apprentissage artificiel. Typiquement, cette approche peut être scindée en deux sous-approches : la première consiste à distinguer deux grandes phases, à savoir le repérage des expressions référentielles puis la construction des chaînes ; la seconde consiste à tout faire en une fois. La justification de la première sous-approche réside dans la différence de nature des phases : l'identification des expressions référentielles est une tâche d'étiquetage, pour laquelle des techniques à base de CRF semblent bien adaptées ; la construction des chaînes passe par l'appariement incrémental de couples d'expressions référentielles, ce qui en fait une tâche de classification binaire, pour laquelle des techniques à base de SVM (entre autres) semblent bien adaptées (Denis, 2007 ; Recasens, 2010 ; Lassalle, 2015). Quant à la deuxième sous-approche, c'est le terrain de prédilection des techniques à base de réseaux neuronaux, sans pour autant que ces réseaux relèvent (pour l'instant) de l'apprentissage profond : beaucoup de réalisations actuelles se contentent d'une ou deux couches cachées. Là aussi, en caricaturant quelque peu, notons d'une part que de telles approches sont réalisées essentiellement par des informaticiens, d'autre part qu'il s'agit d'approches de type « boîte noire », et que – par conséquent – l'analyse linguistique des erreurs est souvent reléguée à plus tard.

Le constat est immédiat : en parallèle à l'augmentation constante des performances pures des systèmes à base d'apprentissage, il devient urgent de combler le manque d'analyse linguistique des erreurs. Les enjeux d'un tel travail sont multiples : 1. identifier les comportements et erreurs typiques impliqués par telle ou telle technique ; 2. diagnostiquer les systèmes correspondants, si possible en fonction d'autres systèmes, donc *via* des sortes de *test suites* multilingues dédiées à l'analyse d'erreurs (et ne comportant pas que des cas extrêmes tels que des schémas de Winograd) ; 3. identifier les ressources linguistiques utiles qui permettraient de compenser les erreurs ; 4. identifier la méthode de prise en compte de ces ressources dans le système lui-même (en amont,

en aval, ou par forçage de nouveaux *features* d'apprentissage). Le point commun à ces enjeux est clair : la collaboration entre spécialistes de linguistique et de TAL s'avère nécessaire. Le constat par rapport à l'état de l'art est sans appel : pour la tâche d'identification de chaînes qui nous concerne, il n'existe ni méthodologie claire, ni *test suites*, ni *baselines* quant à la compensation d'erreurs, ni repères d'aucune sorte. De nombreux travaux ont été réalisés dans ce sens, mais ils restent cantonnés soit à une sous-tâche précise, soit à une langue particulière. Par exemple, le 4e enjeu ci-dessus a fait l'objet de réflexions et d'expérimentations pour un segmenteur-étiqueteur du français (Constant *et al.*, 2011) mais, à notre connaissance, jamais pour l'identification des chaînes de coréférences en français. Autre exemple, le travail de thèse de (Uryupina, 2007) avait ouvert la voie à plusieurs tentatives concernant quelques-uns des 4 enjeux mentionnés (pas pour le français, ceci dit), mais sans véritable répercussion perceptible sur la méthodologie de réalisation des systèmes plus récents. Pour terminer cet état de l'art forcément réducteur, notons que le 4e enjeu enchaîne sur un 5e qui consiste à étudier les types d'hybridation, de manière à envisager la réalisation de systèmes hybrides, non seulement à l'image de (Godbert et Favre, 2017) qui, pour la langue française, utilise apprentissage artificiel pour la détection des expressions référentielles et système de règles pour celle de chaînes, mais – plus généralement – afin d'envisager de nouvelles techniques de « contrôle » de l'apprentissage pour optimiser les voies explorées dans les calculs et mieux exploiter des ressources linguistiques à l'intérieur même des fameuses « boîtes noires ».

Dans cet article, nous présentons quelques pistes explorées actuellement dans le cadre du projet Democrat. Il s'agit d'une étude tout à fait préliminaire, et exclusivement qualitative : nous n'avons encore aucune mesure de performance, aucun taux d'amélioration de système à présenter, et nous considérons d'ailleurs qu'à cette étape du travail, ce n'est pas forcément rédhibitoire. En effet, les chiffres ont tendance à orienter nos efforts dans des voies parfois biaisées, surtout quand ces chiffres en question reflètent mal la gravité des erreurs telle que perçue par les linguistes – à l'instar des mesures de l'accord inter-annotateurs concernant les expressions référentielles et les chaînes (Mathet, 2017 ; Mathet & Widlöcher, 2016). Par ailleurs, nous soulignons qu'aucun réseau neuronal n'a encore été testé dans le cadre de cette étude préliminaire : nous nous fondons essentiellement sur des expérimentations autour du système CROC (Désoyer *et al.*, 2014), c'est-à-dire concernant la phase d'appariement d'expressions référentielles coréférentes, avec des techniques plus « classiques » – arbres de décision, SVM, *Naive Bayes*. Même si nous soupçonnons que des réseaux neuronaux pourraient aboutir à des erreurs similaires ou comparables, nous n'avons pour l'instant aucun argument dans ce sens.

2 *Features* pour l'identification de chaînes de coréférences

Typiquement, les *features* exploités par l'apprentissage artificiel relèvent d'interprétations faciles à obtenir sur du texte brut, donc : des informations morphosyntaxiques, des catégories linguistiques fondées sur des marqueurs fiables (notamment la détermination : défini, démonstratif, ou indéfini) et des mesures bas niveau d'ordre typographique : distance textuelle entre deux expressions référentielles ; distance de Hamming ou de Levenshtein entre les formes de surface de deux expressions. Sont parfois ajoutés quelques aspects sémantiques, notamment pour les langues comme l'anglais qui bénéficient de ressources telles que WordNet. (Soon *et al.*, 2001) propose ainsi un ensemble de 12 *features*, largement repris dans de nombreux projets sur la coréférence, par exemple dans celui portant sur la langue polonaise (Ogrodniczuk *et al.*, 2015). Ensemble focalisé sur l'appariement de paires d'expressions coréférentielles, on y trouve la distance en nombre de phrases entre l'expression *i* et l'expression *j*, ainsi que 11 booléens qui répondent à des questions telles que : *i* est-il un pronom ? *j* en est-il un ? *j* est-il un groupe nominal défini ? *i* et *j* s'accordent-ils en genre ? en nombre ? *i* et *j* sont-ils tous les deux des noms propres ? etc.

La conception de CROC (Désoyer *et al.*, 2014) est partie de cet ensemble de référence. Pour l'apprentissage, nous nous sommes fondés sur le corpus d'oral transcrit annoté ANCOR (Muzerelle *et al.*, 2014). Il s'agit d'un corpus de parole conversationnelle (standards téléphoniques, entretiens sociolinguistiques) qui comprend 488.000 mots, 116.000 mentions et 51.000 relations anaphoriques annotées. Parmi les *features* de (Soon *et al.*, 2001), nous avons dû ignorer ceux qui n'étaient pas calculables automatiquement pour la langue française, notamment ceux fondés sur WordNet. Et nous avons ajouté des *features* spécifiques de l'oral, notamment l'identité du locuteur, ainsi que des spécificités des annotations d'ANCOR, par exemple l'attribut « *new* » qui dit si une expression référentielle est une première mention ou non – élément d'importance pour l'identification automatique de chaînes. Notons au passage que, comme tout corpus, ANCOR présente des particularités telles que ce fameux « *new* », et qu'il est toujours difficile de « quantifier » la dépendance au corpus : en l'absence de métriques dédiées, les résultats obtenus par apprentissage sur ce corpus restent liés à ce corpus et non généralisables. (Désoyer *et al.*, 2014) présente en détails l'ensemble de 30 *features* retenu, les techniques utilisées (SVM) et les résultats obtenus par CROC.

Le déroulement dans Democrat de nouvelles expérimentations conduit également à l'extension des ensembles de *features* envisagés, avec par exemple des informations syntaxiques (Grobol *et al.*, 2018). Notons que la taille de l'ensemble de *features* n'admet pas de limite autre que celle autorisée par le temps de calcul. Plutôt que de tester expérimentalement l'ajout des *features* un par un, il semble que les concepteurs commencent par identifier le maximum de *features* « faciles à calculer », avec le postulat implicite que plus le système disposera de *features*, plus il pourra trouver des arguments pertinents pour l'aider dans sa tâche. C'est d'ailleurs un aspect discuté dans (Uryupina, 2007), qui est allée jusqu'à envisager pas moins de 351 *features* ! Cet aspect est important pour notre propos, car l'identification d'erreurs conduira souvent à ajouter des *features* supplémentaires, plutôt qu'à remettre en cause des *features* existants voire l'approche elle-même.

3 Première analyse des erreurs commises par le système CROC

Les quatre enjeux présentés dans l'introduction constituent un tout premier pas vers la mise en place à plus long terme d'une méthodologie d'analyse linguistique des erreurs pour la coréférence. Afin de contribuer à établir des repères pour une telle méthodologie, nous proposons d'explorer les catégories suivantes : 1. les erreurs de frontières et 2. de type, c'est-à-dire les types d'erreurs classiquement observés pour des tâches de TAL plus simples que celle relevant de la coréférence ; 3. les problèmes de bruit spécifiques de cette tâche impliquant des difficultés de résolution de la référence ; 4. les problèmes de silence spécifiques à la détection des coréférences ; 5. les biais tels que la construction par les systèmes de chaînes « fourre-tout » quand les données ne suffisent plus. Nous nous appuyons sur des exemples issus du système CROC, sachant que celui-ci présente une limitation de taille : il détecte des paires coréférentes (et donc des chaînes par transitivité) à partir d'un corpus déjà annoté en mentions. Un autre outil doit être utilisé pour détecter ces dernières.

3.1 Erreurs de frontières

Les erreurs de frontières concernent la délimitation des marquables, c'est-à-dire, dans le cas des chaînes de coréférences, la délimitation des expressions référentielles. Il faut d'abord noter que délimiter correctement constitue une tâche d'une grande complexité, très délicate pour un humain, y compris un linguiste. À titre d'illustration, le manuel d'annotation de Democrat comporte environ 30 pages, dont 29 fourmillent de conseils, d'exemples et de précautions à prendre pour cette

délimitation (alors que les détails de l'affectation d'un référent remplissent à peine une page). Parmi les difficultés rencontrées par des humains et *a fortiori* par des systèmes : l'identification des modifieurs, subordonnées relatives, appositions qui appartiennent parfois aux expressions (relatives restrictives) et parfois non (relatives explicatives). Par conséquent, ce type d'erreur n'est pas le plus intéressant pour évaluer les performances d'un système de TAL, et encore moins pour lancer un travail de diagnostic et de développement supplémentaire du système. Au contraire – et ni les mesures de performances des systèmes (MUC, B3, CEAF, BLANC, voir l'article de présentation de CROC – (Désoyer *et al.*, 2014) – qui les décrit) ni les mesures d'accord inter-annotateurs n'en rendent compte correctement – il nous semble qu'il s'agit d'erreurs soit négligeables (elles n'ont aucune conséquence sur la constitution des chaînes) soit rattrapables *via* une uniformisation *a posteriori* à l'aide d'un analyseur syntaxique voire d'un *chunker*.

3.2 Erreurs de type

Les erreurs de type concernent la catégorisation des marquables selon la liste de catégories impliquée par la tâche ; ce sont, dans le cas des chaînes de coréférences, des erreurs d'« affectation » d'une expression référentielle à un référent. Ce type d'erreur est peu fréquent quand il s'agit de groupes nominaux comprenant des mots pleins, mais beaucoup plus fréquent quand il s'agit de pronoms. Par exemple, dans l'énoncé « est ce que vous avez un des informations chez des expositions qui sont à la bibliothèque », le pronom « qui » a été délimité mais n'a pas été rattaché à la chaîne des expositions. Au contraire, dans « UBS bonjour… j'ai ma fille qui a passé son DEUG… elle a passé […] », le pronom « elle » a été rattaché à UBS et non à la fille du locuteur.

Il s'agit là d'une erreur d'importance, car la constitution des chaînes en est bouleversée. Or corriger ce type d'erreur n'est pas évident : ce n'est pas en ajoutant un nouveau *feature* que l'on peut espérer compenser un taux d'erreurs trop important. Au contraire, c'est probablement le processus d'apprentissage qu'il s'agit de revoir, en modifiant par exemple les paramètres de classification pour que les différentes classes correspondent aux différents types de reprises, ou encore en apprenant des modèles spécifiques pour chaque type de reprises, à partir des mêmes *features* qui se verront attribuer des poids différents en fonction du type d'expressions pris en compte (Denis, 2007).

3.3 Bruit : les pronoms impersonnels et les groupes nominaux non référentiels

Dans la mesure où un même groupe de mots tel que « *il* existe » implique un pronom « il » qui peut être personnel (« Pierre, *il* existe ») ou impersonnel (« *il* existe des gens qui… »), il arrive que les systèmes comme CROC retiennent comme référentiels des pronoms impersonnels : « il faudrait demander à ma collègue », « il faudra monter à 'info montagne' », entre autres exemples de CROC. Prévoir en amont ou en aval de l'apprentissage un système spécifique de détection automatique des pronoms impersonnels n'est pas forcément une bonne idée, du fait du taux d'erreur non nul que présentera ce système, qu'il soit conçu à base de règles ou d'apprentissage. Corriger ce type d'erreurs peut se faire *a priori* par ajout de nouveaux *features* : appartenance ou non à une liste prédéfinie d'expressions comme « il y a », « il faut » ou « il existe » (en plus des classiques « il pleut » et « il neige ») ; prise en compte non pas du verbe seul mais de plusieurs *tokens* avant et après le pronom ; etc.

Dans le cas du système CROC, un autre type d'erreurs relevant du bruit est apparu : le système a retenu comme référentiels les groupes nominaux « une belle jambe » et « la grosse tête », alors qu'ils apparaissaient dans des expressions figées non référentielles (« ça me fait une belle jambe », « il a

pris la grosse tête »). Afin de compenser ce comportement, nous avons envisagé – comme pour celui relatif aux pronoms impersonnels – d'ajouter de nouveaux *features*, et notamment un booléen indiquant l'appartenance ou non à une liste prédéfinie d'expressions figées françaises avec la mention d'une partie du corps humain – liste facile à trouver sur le *web*. Il s'agit cependant d'une intention qui, avant même d'être implémentée, a été décriée par des spécialistes d'apprentissage artificiel : pour eux, l'apprentissage peut ignorer totalement le « signal faible » émis par l'ajout d'un *feature*, et donc ne pas augmenter ses performances face à quelques dizaines d'occurrences.

Notons que le même problème apparaît avec des locutions intégrant un nom commun potentiellement référentiel, comme les connecteurs « dans le fond », « à la fin », ou les locutions telles que « de manière à ». L'observation est identique, et la solution envisagée l'est également, avec les mêmes réserves. S'y ajoute un aspect relevant du caractère plus ou moins figé de ces expressions : gérer une liste de formes figées françaises peut entraîner des débats linguistiques complexes. Or, le but étant d'ajouter le maximum de ressources linguistiques dans de nouveaux *features*, on peut considérer que plus les listes sont étendues, mieux c'est. Éventuellement, rien n'empêche de gérer deux listes, la première avec des formes vraiment figées, la seconde avec des formes en cours de figement ou semi-figées : le système pourra alors adapter de lui-même son exploitation de la seconde en fonction des exemples rencontrés.

3.4 Silence : maillons manquants et coupures de chaîne

Le silence peut concerner aussi bien le ratage d'une expression référentielle (ce qui entraînera une chaîne incomplète), que le ratage d'un maillon dans une chaîne, c'est-à-dire la non affectation de l'expression à la chaîne qui la concerne (l'expression restant alors un singleton, similaire à un groupe nominal jamais repris). C'est souvent le cas lors de reprises par un synonyme, comme dans « vous pouvez me dire le bureau ou alors e la pièce », où la pièce n'a pas été affectée à la chaîne du bureau. Ce type d'erreur conduit parfois à séparer une chaîne en deux chaînes parallèles. Par exemple, on trouve une chaîne « Siemens » (« Siemens… Siemens… ») parallèle à une chaîne « Société Siemens… la société Siemens… », là où il ne devrait n'y avoir qu'une seule chaîne. Ces erreurs pourraient être compensées par l'exploitation de ressources sémantiques permettant de repérer la synonymie et les différentes désignations d'une même entité nommée.

Les erreurs les plus fréquentes, cependant, concernent des marqueurs peu communs : dès que le texte s'écarte des classiques noms propres, groupes nominaux et pronoms, plus de silence intervient. Ainsi, les pronoms « ça » font l'objet d'oublis, de même que les pronoms « ce » (ou « c' »), « en », « y », ou encore les relatives sans antécédent. Trouver une tactique pour compenser ce type d'erreur n'est pas évident non plus : ajouter des *features* semble bien aléatoire. Là aussi, comme dans la section 3.2, il semble raisonnable d'envisager des apprentissages séparés.

3.5 Tendance à construire des chaînes non pertinentes

Les chaînes non pertinentes, qui réunissent plusieurs référents – parfois une multitude – ou n'en reflètent au contraire aucun (quand leurs maillons sont des groupes nominaux non référentiels), sont souvent rédhibitoires pour les linguistes car ce sont les erreurs les plus saillantes pour un lecteur humain. Il s'agit parfois de la fusion de plusieurs référents relevant d'un même champ lexical : « un appel… un portable… un message… mon portable… un beau mec ». Rectifier ce comportement nécessiterait l'ajout d'un *feature* indiquant la compatibilité sémantique des termes, mais les contre-exemples de reprises par des termes inattendus existent, ce qui modère l'intérêt d'un tel *feature*.

Beaucoup de chaînes non pertinentes sont cependant des suites de pronoms de troisième personne sans aucun nom : « les diplômes… les… les… les diplômes », où les deux pronoms « les » forment une chaîne indépendante, parallèle à celle des diplômes. D'autres rattachent une série de pronoms à un nom en fin de la chaîne, comme dans « une personne qui se trouve alors attendez elle est absentée ici mais je vais vous la passer ailleurs parce qu'elle est partie à la repro donc elle s'occupe de… la repro », où la série de pronoms – « qui », « elle » et « la » – est rattachée à l'expression « la repro » plutôt qu'à « une personne », expression pourtant énoncée avant les pronoms. Dans une autre conversation, plusieurs pronoms (« elle… la… elle… elle… elle ») sont considérés par CROC comme une série de singletons, alors qu'ils désignent bien le même référent.

L'erreur est ici importante, et il nous semble que la priorité doit être donnée à l'évitement de ce biais de l'apprentissage. Or il est difficile de compenser ce comportement par l'ajout de *features* supplémentaires, et il n'est pas dit qu'un apprentissage séparé résoudra quoi que ce soit. À notre sens, c'est ainsi l'enjeu le plus important que nous ayons pu identifier, sans avoir – dans l'état actuel de cette étude préliminaire – autre chose qu'un constat à présenter.

4 Vers une méthodologie d'analyse linguistique des erreurs pour les chaînes de coréférences

4.1 Analyse des erreurs

Selon la méthodologie que nous envisageons, l'analyse des erreurs commence par la comparaison de la sortie d'un système de détection automatique de la coréférence à un corpus de référence annoté à la main, comme le corpus ANCOR pour l'oral ou le corpus Democrat pour l'écrit (ce dernier est en cours d'annotation dans le cadre du projet du même nom : il s'agit d'un corpus composé de textes de différents genres, annotés en chaînes de coréférences et d'un volume comparable à celui d'ANCOR). Nous menons deux types d'observations. D'abord une étude qualitative, en comparant les chaînes prédites par le système au corpus de référence à l'aide d'un outil *ad-hoc* – encore en développement au moment de la rédaction de cet article – qui permet de visualiser par un jeu de couleurs les erreurs de rappel (lorsque le lien de coréférence n'a pas été fait) et les erreurs de précision (lorsqu'un lien de coréférence a été fait alors qu'il aurait dû ne pas l'être). Sur l'ensemble d'ANCOR, nous avons ainsi un corpus de près de 11.000 chaînes à comparer.

Cette étude comparative implique de replacer chaque erreur dans le contexte plus large d'un phénomène linguistique ou discursif. Par exemple, une différence dans le nombre (singulier ou pluriel) d'une expression n'est pas toujours le signe d'une non-coréférence. Un nom collectif peut en effet être repris par un pluriel : « je vais regarder je peux donner le numéro de l'office tourisme d'Espagne… vous leur écrivez vous les appelez ils vous envoient gratuitement tout ce que vous désirez » ; ou encore : « pour nos candidats quoi pour notre personnel ». Ce dernier exemple montre également l'importance du genre textuel, qui doit être pris en compte puisqu'il a une influence sur la composition des chaînes (Schnedecker, 2014). Pour l'oral, il s'agit par exemple de reformulations (« il y a une formation e une licence professionnelle », « est-ce que tu sais où sont les grands plastiques e les sacs poubelles »), de présence de pronoms de première et deuxième personne (« je », « tu », « nous », « vous »), etc.

L'objectif d'une telle analyse est de repérer les erreurs et d'y apporter la meilleure réponse théorique. Elle permet de cibler les ressources linguistiques les plus pertinentes (Uryupina, 2007).

Puisqu'elle s'inscrit dans une démarche pleinement linguistique, elle vise également à apprécier la « qualité linguistique » d'un corpus annoté automatiquement, et d'aider non seulement à améliorer le nombre de liens de coréférence correctement détectés, mais aussi à créer un corpus dans lequel les types d'erreur les plus rédhibitoires pour les linguistes auront été éliminées.

Notre deuxième observation est quantitative et se fait directement sur la sortie de CROC, dont l'algorithme procède en deux temps : 1. classification de paires de mentions (coréférentes ou non coréférentes), 2. reconstruction des chaînes par transitivité. Cette analyse permettra à terme une étude chiffrée et statistique, par exemple en termes d'analyse en composantes principales, des *features* les plus susceptibles d'entraîner une erreur de classification. Il ne s'agit pas ici d'étudier les *features* les plus pertinents pour la détection de la coréférence, mais ceux qui entraînent le plus d'erreurs, notamment en fonction du type de l'expression référentielle.

4.2 Retour aux données d'apprentissage

Notre méthodologie suppose de retourner, après le diagnostic des systèmes, aux données d'apprentissage. Avec un système d'apprentissage statistique, il est possible d'augmenter sans fin le nombre de *features* (Uryupina, 2007). Mais, comme le souligne (Désoyer *et al.*, 2014), il n'est pas aisé de savoir quel est l'effet de tel ou tel *feature* : la solution qui consiste à ajouter les *features* un à un, en entraînant un modèle à chaque fois, et en ne gardant que les *features* qui améliorent les résultats de classification, est une solution difficile à mettre en œuvre car elle demande un plan d'expérimentation conséquent et des temps de calcul importants.

L'ajout de *features* fait par ailleurs débat entre les linguistes et les spécialistes de l'apprentissage. Alors qu'il s'agit de plus en plus de la seule façon de travailler (et d'être impliqués) pour les premiers (Tanguy & Fabre, 2014), les seconds n'y voient que peu d'intérêt : le système ignorera les *features* qu'il considérera non pertinents même s'ils semblent, aux yeux des linguistes, importants pour la détection de la coréférence. L'ajout de *features*, même linguistiquement motivés, n'est donc pas une garantie de l'amélioration d'un système d'apprentissage. Ainsi, même avec 351 *features* « linguistiques », (Uryupina, 2006) ne trouve qu'une amélioration « modérée » des résultats par rapport à l'approche dite « pauvre en connaissances » de (Soon *et al.*, 2001), qui ne comprend que 12 *features*.

Une solution alternative est l'ajout dans le corpus d'apprentissage d'exemples à la fois positifs et négatifs, ce qui inclut la duplication d'exemples, c'est-à-dire la manipulation « brutale » des données, à l'instar du principe de modélisation des utilisateurs (*user simulation*) exploité dans le domaine du dialogue homme-machine, cf. section 7.8 de (Rieser & Lemon, 2011). Par exemple, en intégrant massivement des occurrences d'expressions figées incluant des parties du corps humain, il serait possible de corriger le défaut de CROC, décrit plus haut, de considérer qu'avoir « la grosse tête » est référentiel. Mais la sélection des exemples à ajouter introduit un biais supplémentaire.

Notre objectif, à l'état de spéculations pour l'instant, est de tester et d'évaluer de telles méthodes pour identifier la façon la plus efficace de prendre en compte les résultats de l'analyse linguistique des erreurs dans les données d'apprentissage. Ce qui nous amène aux perspectives.

5 Perspectives

L'un des volets du projet Democrat est d'établir une méthodologie d'analyse linguistique des erreurs des systèmes de détection automatique de la coréférence, et d'utiliser cette analyse pour améliorer les performances de ces systèmes en ciblant les données d'apprentissage. Cette méthodologie devra pouvoir être utilisée quelle que soit la technique d'apprentissage mise en œuvre.

L'analyse des erreurs ouvre également la voie à d'autres méthodes de détection de la coréférence, dites *hybrides*, mêlant à la fois les systèmes symboliques, à base de règles, et les systèmes à base d'apprentissage. Outre l'alternance de méthodes symboliques et statistiques (par exemple, un système statistique pour la détection des expressions référentielles suivi d'un système symbolique pour la détection de la coréférence), nous envisageons pour l'heure trois possibilités pour intégrer des ressources linguistiques dans un système statistique. Il est d'abord possible de prendre en compte ces ressources à l'intérieur même d'un système, comme l'a montré (Constant *et al.*, 2011) pour l'étiquetage morphosyntaxique. Il est ensuite possible de *filtrer* de façon systématique les données présentées à l'algorithme. Par exemple, un ensemble de règles pour détecter les expressions figées non référentielles (comme celles impliquant les parties du corps) pourrait filtrer ces expressions avant le passage d'un algorithme statistique de détection de la coréférence. Enfin, il est possible de *corriger* les données à la sortie de l'algorithme par d'autres règles, ou même de déléguer certaines tâches à un système symbolique, par exemple la détection des anaphores liées – en termes génératives – comme les pronoms réfléchis ou les pronoms relatifs. En reprenant les méthodes de (Denis, 2007) et (Lee *et al.*, 2013), qui consistent, respectivement, à apprendre des modèles ou définir des règles spécifiques à chaque type de reprises (pronominales, fidèles, infidèles, etc.), on peut faire alterner des phases de filtrage et/ou correction avec des phases statistiques. Ces approches hybrides, élaborées à partir d'une analyse des erreurs des systèmes existants, devraient permettre de retenir le meilleur des approches symboliques et statistiques.

Remerciements

Ce travail a été réalisé avec le soutien de l'ANR dans le cadre du projet Democrat (ANR-15-CE38-0008). Il a bénéficié de réflexions et de discussions avec des chercheurs des laboratoires Lattice, LiLPa, ICAR et IHRIM : merci à eux, et en premier lieu à Isabelle Tellier (1968 – 2018).

Références

CHAROLLES M. (2002). *La référence et les expressions référentielles en français*. Paris : Ophrys.

CONSTANT M., TELLIER I., DUCHIER D., DUPONT Y., SIGOGNE A., BILLOT S. (2011). Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français. Actes de TALN, Montpellier, 321.

CORBLIN F. (1995). *Les formes de reprise dans le discours. Anaphores et chaînes de référence*. Rennes : Presses Universitaires de Rennes.

DENIS P. (2007). New Learning Models for Robust Reference Resolution. Ph.D. dissertation, Austin, University of Texas.

DESOYER A., LANDRAGIN F., TELLIER I., LEFEUVRE A., ANTOINE J.-Y. (2014). Les coréférences à l'oral : une expérience d'apprentissage automatique sur le corpus ANCOR. *Traitement Automatique des Langues* 55(2), 97-121.

GODBERT E., FAVRE B. (2017). Détection de coréférences de bout en bout en français. Actes de la *24e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2017)*, Orléans.

GROBOL L., TELLIER I., DE LA CLERGERIE É., DINARELLI M., LANDRAGIN F. (2018). ANCOR-AS: Enriching the ANCOR Corpus with Syntactic Annotations. Proceedings of the *11th Edition of the Language Resources and Evaluation Conference (LREC 2018)*, Miyazaki, Japan.

GROUPE DE FRIBOURG (2012). *Grammaire de la période*. Berne : Peter Lang.

LANDRAGIN F., SCHNEDECKER C. (2014). *Les chaînes de référence*. *Langages* 195, Paris : Larousse.

LASSALLE, E. (2015). Structured learning with latent trees: A joint approach to coreference resolution. Thèse de doctorat, Université Paris Diderot Paris 7.

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.

MATHET Y. (2017). The Agreement Measure Gamma-Cat, a complement to Gamma focused on Categorization of a Continuum. *Computational Linguistics* 43(3), 661-681.

MATHET Y., WIDLÖCHER A. (2016). Évaluation des annotations : ses principes et ses pièges. *Traitement Automatique des Langues* 57(2), 73-98.

MITKOV R. (2002), *Anaphora Resolution*, Upper Saddle River, NJ : Pearson Education.

MUZERELLE J., LEFEUVRE A., SCHANG E., ANTOINE J.-Y., PELLETIER A., MAUREL D., ESHKOL I., VILLANEAU J. (2014). ANCOR CENTRE, a large free spoken French coreference corpus: description of the resource and reliability measures. Proceedings of the *9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland.

NOUVEL, D., EHRMANN M., ROSSET S. (2015). *Les entités nommées pour le traitement automatique des langues*, Londres : ISTE éditions.

OGRODNICZUK M., GLOWINSKA K., KOPEC M., SAVARY A., ZAWISLAWSKA M. (2015). *Coreference in Polish: Annotation, Resolution and Evaluation*. Berlin : Walter De Gruyter.

RECASENS, M. (2010). Coreference: Theory, Annotation, Resolution and Evaluation. Ph.D. dissertation, Universitat de Barcelona.

RIESER V., LEMON O. (2011). *Reinforcement Learning for Adaptive Dialogue Systems. A Data-driven Methodology for Dialogue Management and Natural Language Generation*. Heidelberg : Springer.

SCHNEDECKER C. (1997). *Nom propre et chaîne de référence*. Paris : Klincksieck.

SCHNEDECKER C. (2014). « Chaînes de référence et variations selon le genre ». *Langages* 195, 23-42.

SCHNEDECKER C., GLIKMAN J., LANDRAGIN F. (2017). *Les chaînes de référence en corpus*. *Langue Française* 195, Paris : Armand Colin.

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.

TANGUY L., FABRE C. (2014). « Évolutions de la linguistique outillée. Méfaits et bienfaits du TAL ». *L'Information Grammaticale* 142, 14-22.

URYUPINA O. (2006). Coreference resolution with and without linguistic knowledge. Proceedings of the *5th Edition of the Language Resources and Evaluation Conference (LREC 2006)*, Genoa, Italy.

URYUPINA O. (2007). Knowledge acquisition for coreference resolution. Ph.D. dissertation, Saarbrücken : Universität des Saarlandes.