Exploitation du corpus Democrat par apprentissage artificiel

Loïc Grobol

Exploitation du corpus Democrat par apprentissage artificiel

Loïc Grobol
LLF, 8, Rue Albert Einstein 75013 Paris, France
Lattice, 1 Rue Maurice Arnoux, 92120 Montrouge, France

Résumé

La détection automatique de chaînes de coréférences pour le français est encore un domaine assez peu exploré, entre autres en raison du développement tardif de ressources annotées adaptées. Le corpus Democrat, premier corpus de français écrit de grande envergure annoté en chaînes de coréférences rend possible l'utilisation de techniques d'apprentissage artificiel pour combler ce manque. Dans ce travail, nous présentons le système DeCOFre, premier système de détection des chaînes de coréférences pour le français parlé et étudions son utilisation pour le traitement de Democrat. Nos expériences montrent que ce système n'est pas robuste au changement induits par le passage de l'oral spontané à l'écrit et suggère que les particularités de Democrat pourraient être mieux prises en compte par des architectures plus riches que celles des systèmes *end-to-end* omniprésentes dans l'état de l'art récent.

Mots-clés

Apprentissage artificiel, réseaux de neurones artificiels, détection automatique des chaînes de coréférences, français

Abstract

Automatic coreference resolution for French has a relatively recent history, due to a lack of large scale annotated resources that has only been filled in the last few years. The release Democrat, the first large scale corpus of written French with coreference annotation makes the developement of coreference resolution system for written French using machine learning techniques possible for the first time. In this work, we present DeCOFre, the firs coreference resolution system for *spoken* French and investigate its use for processing Democrat. Our experiments shows that this system is not resilient to the differences between the spoken and written genres, which suggests the need for richer architectures than those used in the recent state of the art end-to-end coreference resolution systems.

Keywords

Machine learning, artificial neural networks, coreference resolution, French

1. Introduction

Le corpus Democrat, dont le présent volume est l'objet, a été principalement conçu pour servir deux objectifs : d'une part, l'étude de la coréférence en français par les méthodes et les outils de la linguistique de corpus ; d'autre part le développement et l'évaluation d'outils de traitement automatique du langage naturel (TAL). En effet, les phénomènes de référence, qu'ils soient de nature syntaxique, pragmatique, discursive ou mettent en jeu des connaissances extralinguistiques, sont omniprésents dans le langage humain et l'interprétation correcte des références est par conséquent essentielle pour une compréhension réelle du langage naturel. De ce fait, comme l'indiquait déjà Karttunen (1976), il semble indispensable pour un

système de TAL capable de traiter des tâches dépendant du sens de documents (et non pas seulement de leurs structures ou de leur contenu lexical) de disposer d'un module[1] de traitement des phénomènes référentiels. Ainsi, un système de question-réponse comme celui imaginé par Karttunen ayant accès à *Dune* de Frank Herbert, pour répondre à la question « Quand Muad'Dib est-il né ? » devra être capable (entre autres) de reconnaître l'antécédent du pronom « il » dans la phrase suivante :

> Ainsi, pour entreprendre l'étude de la vie de Muad'Dib, plaçons le tout d'abord en son temps: **il** naquit en la cinquante-septième année de l'Empereur Padishah, Shaddam IV.

Remarquons également que l'intérêt de cette tâche ne se limite pas à l'interprétation des anaphores pronominales. Par exemple dans des perspectives d'analyse du discours automatique ou semi-automatique, repérer les liens de coréférence entre des groupes nominaux peut être une façon d'étudier les nominations d'objets ou de concepts (Longhi *et al.* 2020) ou d'alimenter des bases de connaissances.

L'importance de ces phénomènes n'a d'ailleurs pas échappé aux architectes de systèmes concrets de TAL, qui dès les travaux de Winograd (1972) ont intégré à des systèmes destinés à des applications complexes (comme des systèmes de dialogue humain-machine, des traducteurs automatiques, des interfaces à commandes vocales…) des capacités de résolution de certains des phénomènes de référence les plus simples. Cependant, bien que des algorithmes relativement simples, tels que le célèbre algorithme de Hobbs (Hobbs 1986) montrent des résultats prometteurs, des contre-exemples tels que les schémas de Winograd — formalisés par Levesque, Davis, et Morgenstern (2012) — suggèrent qu'une compréhension satisfaisante des phénomènes de référence nécessite des approches plus sophistiquées, y compris dans les cas les plus banals. Ainsi dans l'exemple suivant, proposé par Amsili et Seminck (2017) :

> La coupe n'entre pas dans la valise marron, car **elle** est trop petite.

L'interprétation correcte du pronom « elle » fait appel à des connaissances du monde et un à modèle de relations spatiales ; un système de TAL ne peut donc pas se limiter à des considérations purement linguistiques ou internes au document. Notons d'ailleurs que parmi les systèmes existants de reconnaissance automatique des chaînes de coréférences, aucun ne parvient à traiter les exemples de ce type de façon fiable.

Dans ce contexte, l'existence de corpus annotés en chaînes de coréférences apparaît comme une nécessité pour le TAL pour deux raisons :

[1] Notons qu'un tel module ne doit pas nécessairement correspondre à un programme conçu explicitement. Des travaux tels que ceux de Jawahar, Sagot, et Seddah (2019) suggèrent ainsi que des réseaux de neurones artificiels entraîné comme des modèles de langues peuvent acquérir au cours de leur entraînement des connaissances syntaxiques de façon latente sans que rien — ni dans leur architecture ni dans la façon dont ils sont entraînés — ne les y incite *a priori*. Il ne semble donc pas impossible que d'autres systèmes conçus par apprentissage artificiel puissent acquérir de la même façon un module de reconnaissance de chaînes de coréférences.

– Pour être crédible d'un point de vue quantitatif, l'évaluation des performances des systèmes automatiques doit être effectuée sur une quantité suffisante de données réelles.
– Dans le contexte actuel, il semble acquis que les systèmes recourant à des techniques d'apprentissage artificiels offrent des performances bien meilleures que celles des systèmes n'utilisant que des règles *a priori*[2]. Or, par définition, la conception de ces systèmes nécessite une quantité conséquente de données annotées.

Dans de nombreuses autres langues, et en particulier pour l'anglais, de tels corpus existent déjà depuis de nombreuses années (Hirschman et Chinchor 1998), y compris des corpus grande envergure (Poesio et Artstein 2008; Pradhan *et al.* 2011, 2012). Pour le français, leur apparition est plus tardive, le premier, ANCOR (Muzerelle *et al.* 2014) ne datant que de 2014.

Une des conséquences de ce décalage temporel est le relatif manque d'intérêt pour la reconnaissance automatique des chaînes de coréférences pour le français : les travaux existants traitent du français parlé documenté dans ANCOR (Grobol 2019, 2020) et du français écrit de Democrat dans une version simplifiée (Wilkens *et al.* 2020), mais aucun travail ne propose à ce jour d'exploitation complète du corpus Democrat. Nous nous proposons ici de faire quelques pas dans cette direction, en étudiant l'utilisation du système DeCOFre (Grobol 2020) pour le traitement de Democrat.

2. La reconnaissance automatique des chaînes de coréférences

C'est un fait établi : la prise en compte des phénomènes de réf. pour le TAL est cruciale ; toutefois, il reste encore à déterminer de quelle façon ces phénomènes peuvent être modélisés et traités par des systèmes concrets. En effet, les modèles cognitifs, linguistiques et discursifs de la référence tendent à être beaucoup plus sophistiqués que ce qu'il est habituel de rencontrer en TAL, où, le pragmatisme primant, les modèles simples (voire simplistes) mais efficaces sont souvent préférés à des modèles plus complets, mais plus difficiles à mettre en œuvre.

En ce qui concerne la coréférence, le modèle privilégié depuis les premières campagnes MUC (Hirschman et Chinchor 1998) est le suivant :

– Les seules mentions considérées sont les pronoms et les syntagmes nominaux.
– On ne tient compte que des phénomènes de coréférence stricte, en laissant de côté la coréférence floue (Delaborde et Landragin (2019), ainsi que « La coréférence floue dans le corpus Democrat » dans le présent volume), la *near identity* (Recasens, Hovy, et Martí 2011) et les anaphores non-coréférentes (telles que les anaphores associatives).

Ces simplifications ont deux objectifs principaux. Premièrement, réduire le champ des mentions et des chaînes à celles qui sont détectables en pratique et les plus aisément interprétables, aussi bien par un système automatique que par des annotateurs humains. Deuxièmement, faire de la relation de coréférence un relation d'équivalence partielle (symétrique et transitive) sur l'ensemble des mentions. La

[2] Les exceptions à ce constat, telles que le fameux système dit *Stanford sieve* (H. Lee *et al.* 2011, 2013) sont de plus en plus rares, et ne contredisent pas nécessairement notre constat : bien que ces systèmes n'aient pas recours à des techniques d'apprentissage artificiel, le développement de leurs règles par des experts reste tributaire de l'existence de données qui permettent de tester ces règles.

tâche de reconnaissance automatique des chaînes de coréférences dans un document se réduit ainsi à deux étapes[3] :

1. Identifier l'ensemble des mentions
2. Partitionner cet ensemble en classes d'équivalences pour la relation de coréférence, ces classes constituant alors les chaînes de coréférences.

L'étape 1 est alors (par exemple) modélisable comme une tâche de recherche de séquences, qui s'apparente à la tâche bien connue de reconnaissance d'entités nommées, alors que l'étape 2 est une tâche de partitionnement, classique en analyse de données. On se ramène ainsi à des tâches plus simples et mieux maîtrisées. Notons de plus que ce modèle simplifié permet de s'affranchir des considérations liées à l'existence de plusieurs modèles concurrents des phénomènes de référence, pour ne garder que ce qui serait effectivement utile à un système de TAL tel que décrit précédemment.

Il reste clair que cette simplification n'est pas sans inconvénients : elle rend par exemple particulièrement difficile la prise en compte de l'existence des référents évolutifs, et ne permet pas de modéliser l'hétérogénéité des mécanismes linguistiques et cognitifs (syntaxe, connaissance du monde, modèles discursifs) dont dépendent les phénomènes de référence. Il se pourrait ainsi que ce modèle ne se révèle à l'avenir trop simpliste, mais à l'heure actuelle, les alternatives crédibles restent rares et leurs applications peu convaincantes.

L'étape de détection des mentions, dans ce contexte simplifié, a historiquement reçu peu d'attention : pour la plupart, les travaux traitant de cette question se limitent pour la résoudre à des heuristiques d'extraction de constituants à partir d'analyses syntaxiques automatiques. Cette approche n'est pas sans défauts (elle suppose notamment que des analyses syntaxiques de qualité suffisante soient effectivement disponibles) et les travaux récents, à la suite de K. Lee *et al.* (2017) tendent à s'en détacher. C'est en particulier le choix fait par les travaux pour le français de Grobol (2020) et de Wilkens *et al.* (2020), qui, s'attachant à traiter le français parlé, ne peuvent pas se reposer sur des analyseurs syntaxiques automatiques, les performances de ces derniers étant pour l'instant clairement insuffisantes (voir par exemples les résultats obtenus pour le français parlé lors de la campagne d'évaluation CoNLL-2018 (Zeman *et al.* 2018)).

L'étape de reconnaissance des chaînes de coréférences à proprement parler, en revanche, a été traitée bien plus systématiquement, par une grande variété d'approches et de modèles durant les vingt dernières années. Plus récemment, cependant, on observe une convergence vers un modèle commun, malgré de nombreuses variations autour de son idée principale, qui est de traiter cette tâche par une *recherche d'antécédent*. Formellement, cela consiste pour chaque mention d'un document donné, à chercher parmi les mentions qui la précèdent s'il en est une (alors appelée quelque peu abusivement « antécédent ») qui lui est coréférente ou s'il s'agit de la première mention d'une nouvelle chaîne. Par ce procédé, le partitionnement de l'ensemble des mentions se fait par la construction d'arbres dont les premières mentions de chaque chaîne sont les racines, chaque arbre recouvrant une et une seule chaîne de coréférences.

[3] Qui ne sont pas nécessairement traitées indépendamment par les systèmes automatiques.

L'intérêt principal de ce modèle est qu'il ne nécessite pas *a priori* de modéliser ou de traiter des objets de haut niveau comme les chaînes elles-mêmes, mais seulement de prendre des décisions mettant en jeu des paires de mentions[4], ce qui simplifie la conception des algorithmes de traitements et leur implémentation dans des systèmes concrets, mais aussi les exigences en termes de calculs de ces systèmes.

Cependant, cette simplification a un coût : l'absence de modélisation implicite des chaînes de coréférences fait reposer l'intégralité du processus sur la qualité des représentations internes des mentions (là où des modèles plus complexes peuvent déduire certaines des caractéristiques des mentions de leur appartenance à des chaînes déjà identifiées). La coïncidence de la généralisation de ce modèle et de celle de techniques d'apprentissage puissantes, et en particulier ces dernières années des techniques dites d'« apprentissage profond » reposant sur des réseaux de neurones artificiels, n'est probablement pas un hasard.

En ce qui concerne le français, cette approche semble pertinente : c'est en effet celle choisie par Grobol (2020) et par Wilkens *et al.* (2020) (reprenant et adaptant le système de Kantor et Globerson (2019)), avec dans les deux cas de bons résultats d'un point de vue quantitatif. Il semble donc naturel pour la présente étude de faire le même choix. D'autre part, les travaux de Wilkens *et al.* (2020) montrent que le système proposé par Kantor et Globerson (2019) n'est pas directement applicable à Democrat dans les conditions matérielles qui sont les nôtres. Les documents de Democrat sont en effet significativement plus longs que ceux pour lesquels ce système a été conçu, et ses spécificités rendent complexe un tel changement d'échelle. Les expériences de Wilkens *et al.* (2020) se limitent pour cette raison à une version simplifiée de Democrat où les documents sont tronçonnés en parties de 2000 mots et les chaînes de coréférences se retrouvent donc également tronquées.

Le système DeCOFre (Grobol 2020) résulte du choix inverse : utiliser un modèle moins riche, mais capable de traiter des documents de taille arbitraire. Notre but étant ici de traiter le corpus Democrat dans toute sa spécificité, c'est sur dernier système que nous fondons nos expériences.

3. Le système DeCOFre

Présentons[5] à présent le système DeCOFre (Grobol 2020), utilisé dans les expériences rapportées dans ce travail.

3.1 Représentation d'empans de mots

DeCOFre, comme son précurseur E2EC (K. Lee *et al.* 2017) tire l'essentiel de ses performances d'un mécanisme de représentations contextuelles d'empans de mots arbitraires, c'est-à-dire d'un modèle capable de représenter toute suite de mots par un vecteur de dimension fixe encodant son rôle syntaxique, son contenu lexical

[4] Ces décisions ne sont toutefois pas nécessairement indépendantes les unes des autres : des travaux comme ceux de Kantor et Globerson (2019) et de K. Lee, He, et Zettlemoyer (2018) montrent en effet que la prise en compte de décisions prises en amont dans un document améliore — légèrement mais de façon statistiquement significative — la recherche d'antécédents pour une mention donnée.

[5] Par souci de place, on restera à un niveau relativement général, pour une présentation détaillée et plus formelle, se reporter à Grobol (2020).

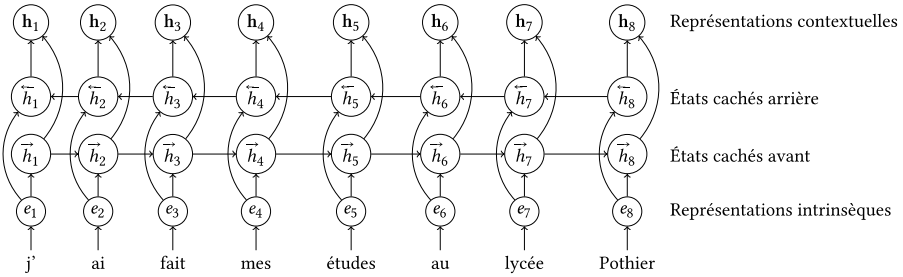

Figure 1: Représentations contextualisées des mots

et une partie de son contenu sémantique[6]. Ce mécanisme repose sur la capacité des réseaux de neurones artificiels récurrents (Graves et Schmidhuber 2005; Hochreiter et Schmidhuber 1997) et attentionnels (Bahdanau, Cho, et Bengio 2015) à contextualiser des représentations vectorielles de mots.

Plus précisément : étant donné une suite de mots $w_1, w_2, \ldots, w_i, \ldots, w_j, \ldots w_n$, pour lesquels on dispose de représentations vectorielles[7], une représentation contextuelle de l'empan $w_i, \ldots, w_j$ s'obtient en appliquant un réseau de neurones récurrent (bi-directionnel dans le cas de DeCOFre, voir Figure 1) à la séquence $w_1, \ldots, w_n$ et en conservant les états de sa couche cachée $h_1, \ldots, h_n$ comme représentations contextualisées des mots. La concaténation de h_i et h_j peut alors suffire à constituer la représentation cherchée. En pratique, on observe qu'il est bénéfique d'y adjoindre également une moyenne pondérée des vecteurs $h_i, \ldots, h_j$ (Figure 2), comme substitut de représentation d'une tête syntaxique de l'empan (voir à ce sujet les résultats rapportés par K. Lee *et al.* (2017)).

3.2 Détection des mentions

Une des innovations de DeCOFre est le procédé utilisé pour la détection des mentions. À la différence d'autres systèmes, comme K. Lee *et al.* (2017), où cette

[6] Ceux-ci pouvant être vides, ou dégénérés. Ainsi des empans comme « de Carthage, dans les », à cheval sur plusieurs constituants, n'ont évidemment pas de rôle syntaxique ni de sens et leur contenu lexical est sans grand intérêt. Encoder cette absence est pourtant crucial, puisqu'elle permet de disqualifier immédiatement les empans de ce type dans la recherche de mentions.

[7] On reste ici agnostique quant à l'origine de ces représentations vectorielles, qui peuvent être des représentations vectorielles de mots classiques (Rumelhart, Hinton, et Williams 1986; Collobert et Weston 2008), potentiellement pré-apprises (Bengio *et al.* 2003; Mikolov *et al.* 2013), des représentations vectorielles de mots contextuelles (Devlin *et al.* 2019; Howard et Ruder 2018; Peters *et al.* 2018) ou de toute autre alternative.

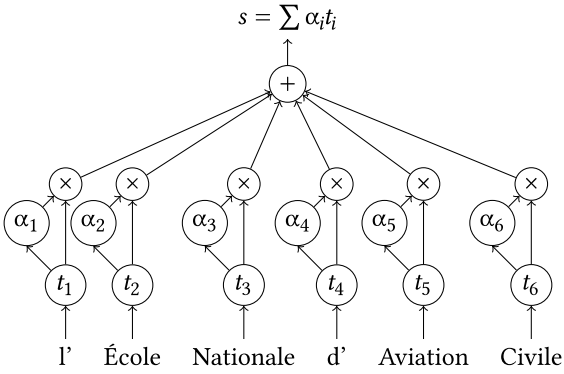

Figure 2: Représentations des pseudo-têtes syntaxiques

étape est traitée implicitement lors de la détection des chaînes de coréférences, la détection des mentions dans DeCOFre repose en effet sur une simple classification des empans de mots. Plus précisément, étant donné une phrase[8] constituée des mots $w_1, \ldots, w_n$, DeCOFre traite la tâche de détection des mentions en attribuant à chacun des empans de mots consécutifs (autrement dit l'ensemble des suites $w_i, \ldots, w_j$ pour $1 \leqslant i \leqslant j \leqslant n$) une classe. On peut pour cela se contenter des deux classes « mention » et « non-mention » ou adopter une typologie plus fine — pourvu que l'on dispose des annotations adéquates. Une fois cette classifications faite, il suffit pour récupérer les mentions d'un document d'en extraire tous les empans auxquels ont été affectés des classes autres que « non-mention ».

En pratique, étant donné le grand nombre d'empans de mots (de l'ordre de n^2), en particulier lorsqu'une segmentation en phrases n'est pas disponible, on limite la longueur des empans considérés à un maximum de l'ordre de quelques dizaines de mots. Concrètement, dans DeCOFre, cette classification est effectuée par un réseau de neurones de taille modeste, l'essentiel du traitement étant ici réalisé au niveau des représentations des empans telles que présentées précédemment.

L'inconvénient majeur de cette formulation est le grand déséquilibre entre les classes considérées : ainsi, les empans non-mentions représentent en général (bien que cela varie suivant les corpus) autour de 99% de tous les empans. Toutefois, ce déséquilibre ne semble pas poser de réel problème en pratique. Au contraire, des stratégies d'échantillonnage visant à le réduire semblent plutôt dégrader les performances finales du système (Grobol 2020).

3.3 Recherche d'antécédent

En ce qui concerne la détection des chaînes de coréférences à proprement parler, DeCOFre adopte l'approche par *classement d'antécédents* introduite par Denis et Baldridge (2008). Cette méthode appartient à la catégorie des méthodes de recherche

[8] Ou tout autre fragment de document dont on sait qu'aucune mention ne chevauche les frontières. Quand un découpage en phrases n'est pas possibles (comme c'est souvent le cas pour l'oral spontané), on peut par exemple considérer des périodes macro-syntaxiques, des tours de parole, ou — en dernier recours — le document dans son intégralité.

d'antécédents définie précédemment et découle d'un constat immédiat lors de l'implémentation concrète d'un système de recherche d'antécédent : que faire quand un système trouve plusieurs antécédents concurrents pour une même mention ? Il est bien évidemment possible de conserver tous les antécédents détectés, mais compte tenu de la faillibilité des systèmes, cette approche conduit bien souvent à fusionner abusivement de nombreuses chaînes, avec des conséquences désastreuses sur la cohérence du résultat. Ainsi, dès les travaux de Soon, Ng, et Lim (2001), il est apparu comme plus prudent de ne conserver qu'un seul antécédent par mention — un choix qui, s'il ne protège évidemment pas des erreurs, permet au moins d'en limiter la portée.

Il reste à déterminer une heuristique permettant de choisir cet unique antécédent. Celle choisie par Soon, Ng, et Lim (2001) est de toujours sélectionner l'antécédent le plus proche, ce qui est particulièrement pratique pour la gestion des anaphores pronominales intraphrastiques, et qui a le mérite d'être complètement indépendant de la méthode de détection des antécédents. Il est cependant clair que cette méthode peut encore pécher par excès de simplicité. Ainsi dans l'exemple suivant :

> « **Docteur Calvin**, votre retraite terminera une ère… ». **Elle** ne m'adressa pas un sourire, ce qui ne me surprit pas, tant il était notoire à l'U.S. Robots que **Susan Calvin** ne souriait jamais.

si les liens de coréférence entre « Docteur Calvin » et « Elle » et entre « Elle » et « Susan Calvin » ne sont pas nécessairement simple à détecter pour un système automatique, repérer le lien entre « Docteur Calvin » et « Susan Calvin » semble beaucoup plus fiable, et se priver de cette victoire facile semble dommage. C'est cette intuition qui sous-tend l'idée, introduite par Ng et Cardie (2002), de plutôt choisir l'antécédent pour lequel la confiance du système est maximale. Ce choix nécessite toutefois d'avoir accès à une mesure de confiance appropriée.

Denis et Baldridge (2008) proposent de résoudre ce problème en développant un modèle qui n'est plus un modèle de classification binaire de paires de mentions en deux classes « coréférent » et « non-coréférent », comme il était alors courant, mais d'apprendre directement un modèle attribuant un « score de coréférence » à chaque paire. Ainsi, sélectionner l'antécédent pour lequel ce score est maximal donne bien un unique antécédent à chaque mention. Mieux, cela permet également de se passer d'une classification explicite des paires de mentions en utilisant un pseudo-antécédent zéro, qui sert d'antécédent aux premières mentions de chaque chaîne : les paires non-coréférentes se voient ainsi simplement attribuer un score plus bas que celui de l'antécédent zéro (certaines formulations proposent alors de lui attribuer le score fixe de zéro et d'attribuer un score négatif aux mentions non-coréférentes, mais ce n'est pas une obligation).

Comme noté précédemment, dans ce type d'approche, la reconstruction des chaînes de coréférences se fait alors immédiatement : l'attribution d'un antécédent à chaque mention donne directement un arbre couvrant pour chaque chaîne de coréférences. Dans le cas de systèmes par classements d'antécédent, il s'agit de ce qu'on appelle couramment un algorithme *glouton* — qui prend à chaque étape du traitement (ici à chaque mention) la décision immédiatement meilleure, sans se

préoccuper des conséquences futures. Cette approche peut paraître simpliste, et des travaux antérieures à l'avènement des systèmes de TAL utilisant des réseaux de neurones artificiels tels que Fernandes, dos Santos, et Milidiú (2014) ou Lassalle et Denis (2015) ont proposé des méthodes exploitant la structure de graphe pondéré que l'attribution de scores aux paires de mentions donne à l'ensemble des mentions. Ces travaux proposent ainsi de considérer l'ensemble des mentions dans sa globalité —plutôt que chaque mention individuellement— en le partitionnant en chaînes de coréférences à l'aide d'algorithmes de recherches de forêts couvrantes optimales. Cependant pour des systèmes capables de déterminer avec suffisamment de précision des scores d'antécédent, les gains obtenus par ces méthodes ne semblent pas suffisant pour justifier la complexité qu'ils induisent, en termes d'ingénierie comme en termes de coût d'apprentissage.

En particulier, dans DeCOFre, ces scores sont déterminés en suivant un modèle dérivé du modèle à deux « grains » proposé par K. Lee, He, et Zettlemoyer (2018). Ce modèle consiste à déterminer pour chaque paire de mentions d'abord un score dit « grossier », dont le calcul est rapide mais la précision limitée, permettant de sélectionner un nombre fixe de candidats antécédents pour lesquels sera calculé un score dit « fin », plus précis mais plus coûteux à calculer. Le calcul pratique de ces scores est effectué par des réseaux de neurones de petites tailles prenant en entrée les représentations d'empans déjà déterminées pour la détection des mentions.

4. Expériences

4.1 Protocole expérimental

Nous évaluons les performances de DeCOFre sur Democrat en utilisant la méthodologie standard de division en trois sous-corpus :

- Un corpus dit d'*entraînement*, auquel le modèle a un accès total durant sa phase d'apprentissage.
- Un corpus dit de *développement* auquel le modèle n'a accès que partiellement durant sa phase d'apprentissage. En particulier, le modèle a accès au score global qu'il obtient sur ce sous-corpus, mais pas au détail de ses erreurs.
- Un corpus dit de *test*, auquel le modèle n'a pas accès lors de sa phase d'apprentissage, et qui sert à évaluer ses performances finales à l'issue de celle-ci.

Ce découpage permet d'entraîner un modèle qui ne soit pas simplement une mémorisation du corpus d'entraînement, puisque le modèle doit également s'assurer d'obtenir des performances satisfaisantes sur le corpus de développement. Le corpus de test permet d'estimer le comportement qu'aurait le modèle sur des données non-annotées inconnues lors de l'apprentissage (mais tout de même suffisamment proches des données d'apprentissage).

La version publiée de Democrat n'inclut pas un tel « *standard split* », nous en avons donc conçu un pour les besoins de cette étude, en nous efforçant de maintenir autant d'homogénéité de genre que possible entre ces sous-corpus. La répartition finale donne en termes de taille 69% pour le corpus d'entraînement, 14% pour le corpus de développement et 14% pour le corpus de test.

Afin d'éviter les écueils liés à la variabilité diachronique, nous ne considérons de plus que la partie de Democrat allant du XIX^e^ au XXI^e^ siècle.

Pour mesurer quantitativement les performances de DeCOFre dans nos expériences, nous rapportons dans ce qui suit les scores F_1 des métriques usuelles pour la détection des chaînes de coréférences : celles (MUC, B^3, CEAF et MELA [CoNLL]) utilisées pour les campagnes d'évaluation CoNLL 2011 et CoNLL 2012 (Pradhan *et al.* 2011, 2012), ainsi que celui de la métrique BLANC (Luo *et al.* 2014; Recasens et Hovy 2011) qui leur est complémentaire.

4.2 Traiter Democrat avec des modèles existants

Le moyen le plus direct de traiter automatiquement Democrat étant donnée l'existence de modèles DeCOFre préentraînés sur ANCOR, est d'appliquer directement ces modèles sur le corpus de test de Democrat. Le modèle que nous utilisons pour cette expérience est le modèle donnant les meilleurs résultats pour ANCOR parmis toutes les configurations décrites par Grobol (2020). Il utilise comme représentation des entrées les représentations vectorielles contextuelles de mots du modèle CamemBERT (Martin *et al.* 2020) et les traits de similarités de chaînes de caractères.

Bien que ce modèle n'ait pas été entraîné sur Democrat, nous ne donnons, pour faciliter la comparaison, qu'une mesure de ses performances sur le corpus de test de Democrat, ainsi que les mêmes mesures réalisées sur le corpus de test d'ANCOR (dans la partition utilisée par Grobol (2020)). Le tableau suivant présente ces résultats.

Tableau 1 : Scores de référence (% F_1) obtenus sur ANCOR et Democrat par un modèle appris sur ANCOR

Corpus	**MUC**	**B^3**	**CEAF**	**CoNLL**	**BLANC**
ANCOR	73.38	84.74	80.67	81.80	75.64
Democrat	45.83	65.46	68.46	59.92	50.86

On observe une grande dégradation de résultats, ce qui n'est pas extrêmement surprenant : outre les différences évidentes entre le français parlé spontané d'ANCOR auquel le modèle est adapté et le français écrit formel de Democrat, les conventions d'annotations des deux corpus ne sont pas exactement identiques. Parmi ces différences, il est une qui semble particulièrement problématique : en effet, dans ANCOR, les pronoms faisant références aux participants d'une interaction ne sont pas annotés comme faisant partie de chaînes de coréférences. Ce choix était motivé par la nature exophorique — plutôt qu'anaphorique — de leur référence, ce qui est significatif pour ANCOR, la focale de ce corpus étant — contrairement à Democrat — les phénomènes *anaphoriques* plutôt que la coréférence stricte.

Outre ces différences de genre linguistique et d'annotations, on observe également que la structure même des mentions et des chaînes de coréférences diffère parfois significativement entre les deux corpus : ainsi les mentions de Democrat ont-elles tendance à être plus longues, plus souvent et plus profondément imbriquées et les chaînes de coréférences tendent à être plus longues, y compris en tenant compte des différences de tailles de documents entre les deux corpus.

Compte tenu de ces différences, une telle baisse de performance en passant à Democrat ne semble pas irréaliste, quelle que soit par ailleurs la différence intrinsèque de difficulté de Democrat pour DeCOFre.

4.3 Apprendre à partir de Democrat

Un autre moyen d'obtenir un modèle de détection des chaînes de coréférences adapté à Democrat est de réentraîner DeCOFre uniquement sur Democrat. On utilise pour cela les mêmes paramètres que dans la configuration précédente[9]. Le tableau suivant rapporte les résultats obtenus sur le corpus de test dans cette configuration.

Tableau 2 : Scores de référence (% F_1) obtenus sur Democrat par un modèle appris sur Democrat

MUC	B^3	CEAF	CoNLL	BLANC
41.52	63.42	50.07	57.25	50.54

Ces résultats sont beaucoup plus étonnants et suggèrent une inadéquation majeure entre l'architecture de DeCOFre et Democrat. En effet, les scores obtenus par un modèle spécifiquement appris sur Democrat sont ici moins bons que ceux présentés précédemment avec un modèle *a priori* étranger à Democrat. Cela suggère que le problème n'est pas simplement que le modèle appris est imprécis, mais que DeCOFre échoue tout bonnement (dans une certaine mesure) à apprendre un tel modèle.

Bien qu'il soit difficile, s'agissant de réseaux de neurones artificiels, de diagnostiquer les causes d'un tel échec, un examen des sorties de ce modèle indique que les coréférences correctement identifiées sont essentiellement des cas d'anaphores fidèles, détectables par heuristique de similarité lexicale de surface, ainsi que des anaphores pronominales intraphrastiques, dont la résolution —contrainte par la syntaxe— ne dépend pas ou presque pas d'une quelconque interprétation sémantique du document. Dans les autres cas, le modèle prédit[10] en général la non-coréférence entre une mention et un candidat antécédent. Autrement dit : ce que DeCOFre a retenu de Democrat, c'est qu'en général, une mention n'a pas d'antécédent, sauf dans certains cas bien précis.

On peut ainsi formuler l'hypothèse suivante : dans l'apprentissage d'un système de recherche d'antécédents, les candidats non-coréférents sont toujours très majoritaires dans l'ensemble des candidats-antécédents d'une mentions donnée. Face à des données difficiles à apprendre, prédire systématiquement la non-coréférence est une heuristique très alléchante pour un système appris automatiquement et prédire la coréférence est un parti risqué. C'est probablement ce qui se produit ici : le manque de régularité dans les paires coréférentes empêche d'apprendre correctement à prédire la coréférence, et le modèle appris fait donc le choix prudent de ne la prédire que dans les cas les plus simples. Ansi, dans « Bientôt, Roger et Paul s'arrêtèrent, ils avaient atteint le bord de la mer, et c'est là qu'ils voulaient passer la nuit. », la coréférence entre les deux « ils » est correctement détectée, mais pas celle entre ces « ils » et « Roger et Paul ».

[9] Pour des raisons d'espace, nous n'incluons pas de résultats de recherches de paramètres optimaux pour Democrat, mais nos expériences en ce sens montrent peu de différences avec les résultats présentés ici, nous nous en tenons donc aux résultats obtenus avec les mêmes paramètres que pour ANCOR.

[10] Bien qu'on se permette dans cette analyse un certain anthropomorphisme pour des raisons de simplicité, rappelons qu'un tel système —quelles que soient ses performances— n'a ni intention, ni agentivité, ni conscience. Ce n'est donc que par abus de langage que nous lui attribuons des qualités comme la prudence ou la témérité.

En contraste, le modèle appris sur ANCOR évalué précédemment a, lui, appris qu'il était possible de prédire la coréférence, même hors de ces cas très simples. En conséquence, bien qu'il commette beaucoup d'erreurs — comme en témoignent les bas scores obtenus — il se risque tout de même à prédire des liens de coréférences dans des cas plus complexes. Ainsi, dans l'exemple précédent, il résout bien la coréférence entre « ils » et « Roger et Paul » et propose également une coréférence entre « là » et « la mer », qui, bien qu'inexacte est plus prometteuse que le repli systématique vers la supposition de non-coréférence.

CONCLUSION ET PERSPECTIVES

Les résultats expérimentaux que nous rapportons ici montrent de façon assez spectaculaire qu'un système de détection automatique des chaînes de coréférences pourtant très performant pour un registre et des conventions d'annotations données peut être très mal adapté pour d'autres configurations.

Partant de ce constat, deux pistes d'amélioration — non-mutuellement exclusives — semblent possibles. L'inadéquation de DeCOFre à Democrat pourrait n'être affaire que de réglages : en effet, la sensibilité des systèmes utilisant des réseaux de neurones artificiels à des détails de conception et d'entraînement est bien connue. Il est ainsi possible que d'autres choix de paramètres, ou un travail d'adaptation de données permettent d'éviter de tomber dans les écueils évoqués précédemment. Mais il se pourrait également que ce soit des choix plus fondamentaux dans l'architecture[11] de DeCOFre qui soient responsables de ses piètres performances sur Democrat. Pour ne donner qu'un exemple, le choix d'utiliser des fenêtres de contexte restreintes pour pallier l'absence de découpage en phrases dans ANCOR soit une erreur pour Democrat, où ce découpage existe bien.

Cependant, en se détachant du cas particulier de DeCOFre, ce constat nous offre également une possibilité de retour à d'autres méthodes plus anciennes et peut-être un peu trop rapidement éclipsée par des systèmes « *end-to-end* ». Il pourrait en effet être intéressant de chercher à compléter un tel système par l'adjonction d'intelligence linguistique. Un des intérêts principaux des systèmes opérant directement sur la surface de documents (en voyant un document comme une simple suite de mots) est qu'ils ne dépendent pas de l'existence de chaînes de prétraitements performantes. En particulier, s'agissant de la coréférence, les systèmes historiques reposaient significativement sur des analyses syntaxiques. L'abandon de ces techniques est en partie lié au manque de précision des analyseurs syntaxiques automatiques, ce qui rendait difficile l'application de ces systèmes à des données réelles non-préalablement annotées. Or, ces dernières années, les gains de performances apportés par les réseaux de neurones artificiels en TAL ont considérablement amélioré les performances des outils utilisés dans de telles chaînes de traitement — en tout cas s'agissant de langues écrites.

Un tel retour à des architectures plus riches ne garantit pas le succès, en particulier à long terme. Les tendances de ces dernières années incitent en effet à la prudence concernant les performances de systèmes en chaînes, souvent sujets aux

[11] Une sorte de contraposée des conclusions de Grobol (2020) sur le manque d'adéquation pour le traitement d'oral transcrit des modèles conçus pour le traitement de l'écrit.

accumulations catastrophiques d'erreurs, et à chercher des solutions dans des approches plus holistiques. Mais, même dans le cas d'un échec, il y a fort à parier que des systèmes plus explicites et plus facilement interprétables, seraient d'une grande aide pour comprendre ce qui dans un corpus comme Democrat peut mettre en échec des systèmes reposants sur l'apprentissage artificiel.

Références

AMSILI P. ET SEMINCK O. (2017): «A Google-Proof Collection of French Winograd Schemas», in *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes*. Association for Computational Linguistics, 24-29. doi: 10.18653/v1/W17-1504.

BAHDANAU D., CHO K. ET BENGIO Y. (2015): «Neural Machine Translation by Jointly Learning to Align and Translate», in Bengio Y. et LeCun Y. (éds.) *Proceedings of the 3rd International Conference on Learning Representations*. San Diego, California, USA.

BENGIO Y. *ET AL.* (2003): «A neural probabilistic language model», *The Journal of Machine Learning Research*, 3, 1137-1155.

COLLOBERT R. ET WESTON J. (2008): «A unified architecture for natural language processing: deep neural networks with multitask learning», in *Proceedings of the 25th international conference on Machine learning*. Association for Computing Machinery (ICML '08), 160-167. doi: 10.1145/1390156.1390177.

DELABORDE M. ET LANDRAGIN F. (2019): «En quoi le pronom « on » a-t-il une valeur anaphorique ? Le cas des successions d'occurrences de « on »», *Les cahiers de praxématique*. (La gestion de l'anaphore en discours : complexités et enjeux), 72, 1-18.

DENIS P. ET BALDRIDGE J. (2008): «Specialized models and ranking for coreference resolution», in *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 660. doi: 10.3115/1613715.1613797.

DEVLIN J. *ET AL.* (2019): «BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding», in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171-4186. doi: 10.18653/v1/N19-1423.

FERNANDES E. R., DOS SANTOS C. N. ET MILIDIÚ R. L. (2014): «Latent Trees for Coreference Resolution», *Computational Linguistics*, 40(4), 801-835. doi: 10.1162/COLI_a_00200.

GRAVES A. ET SCHMIDHUBER J. (2005): «Framewise phoneme classification with bidirectional LSTM and other neural network architectures», *Neural Networks*, 18(5-6), 602-610. doi: 10.1016/j.neunet.2005.06.042.

GROBOL L. (2019): «Neural Coreference Resolution with Limited Lexical Context and Explicit Mention Detection for Oral French», in *Proceedings of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, 8-14.

GROBOL L. (2020): *Coreference resolution for spoken French*. PhD Thesis. Université Sorbonne Nouvelle.

HIRSCHMAN L. ET CHINCHOR N. (1998): «Appendix F: MUC-7 Coreference Task Definition (version 3.0)», in *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference*. Fairfax, Virginia.

HOBBS J. R. (1986): «Resolving Pronoun References», in Grosz B. J., Sparck-Jones K., et Webber B. L. (éds) *Readings in Natural Language Processing*. San Francisco, California, USA: Morgan Kaufmann, 339-352.

HOCHREITER S. ET SCHMIDHUBER J. (1997): «Long Short-Term Memory», *Neural Computation*, 9(8), 1735-1780. doi: 10.1162/neco.1997.9.8.1735.

HOWARD J. ET RUDER S. (2018): «Universal Language Model Fine-tuning for Text Classification», in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 328-339. doi: 10.18653/v1/P18-1031.

JAWAHAR G., SAGOT B. ET SEDDAH D. (2019): «What Does BERT Learn about the Structure of Language?», in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 3651-3657. doi: 10.18653/v1/P19-1356.

KANTOR B. ET GLOBERSON A. (2019): «Coreference Resolution with Entity Equalization», in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 673-677.

KARTTUNEN L. (1976): «Discourse Referents», in McCawley J. D. (éd) *Notes from the Linguistic Underground*. Academic Press (Syntax et Semantics).

LASSALLE E. ET DENIS P. (2015): «Joint Anaphoricity Detection and Coreference Resolution with Constrained Latent Structures», in *Twenty-Ninth AAAI Conference on Artificial Intelligence*. Austin, Texas.

LEE H. *ET AL.* (2011): «Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task», in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics (CONLL Shared Task '11), 28-34.

LEE H. *ET AL.* (2013): «Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules», *Computational Linguistics*, 39(4), 885-916. doi: 10.1162/COLI_a_00152.

LEE K. *ET AL.* (2017): «End-to-end Neural Coreference Resolution», in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 188-197.

LEE K., HE L. ET ZETTLEMOYER L. (2018): «Higher-Order Coreference Resolution with Coarse-to-Fine Inference», in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, Louisiana, USA: Association for Computational Linguistics, 687-692. doi: 10.18653/v1/N18-2108.

LEVESQUE H., DAVIS E. ET MORGENSTERN L. (2012): «The Winograd Schema Challenge», in *Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*. AAAI Press (KR'12), 552-561.

LONGHI J. *ET AL.* (2020): «Le repérage de nominations dans les corpus textuels: de l'exploitation de l'analyse des données textuelles à l'exploration des chaînes de coréférence par le TAL», in *Actes des 15es Journées internationales d'Analyse statistique des Données Textuelles*.

LUO X. *ET AL.* (2014): «An Extension of BLANC to System Mentions», in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 24-29.

MARTIN L. *ET AL.* (2020): «CamemBERT: a Tasty French Language Model», in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 7203-7219.

MIKOLOV T. *ET AL.* (2013): «Efficient Estimation of Word Representations in Vector Space», in Bengio Y. et LeCun Y. (éds) *1st International Conference on Learning Representations*. Scottsdale, Arizona, USA.

MUZERELLE J. *ET AL.* (2014): «ANCOR Centre, a Large Free Spoken French Coreference Corpus: Description of the Resource and Reliability Measures», in *Proceedings of the 9th International Conference on Language Resources and Evaluation*. European Language Resources Association. 843-847.

NG V. ET CARDIE C. (2002): «Improving Machine Learning Approaches to Coreference Resolution», in *Proceedings of the 40th Annual Meeting of the Association for*

Computational Linguistics. Association for Computational Linguistics, 104-111. doi: 10.3115/1073083.1073102.

PETERS M. *ET AL.* (2018): «Deep Contextualized Word Representations», in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. New Orleans, Louisiana, USA: Association for Computational Linguistics, 2227-2237. doi: 10.18653/v1/N18-1202.

POESIO M. ET ARTSTEIN R. (2008): «Anaphoric Annotation in the ARRAU Corpus», in *Proceedings of the 10th International Conference on Language Resources and Evaluation*. European Language Resources Association. 1170-1174.

PRADHAN S. *ET AL.* (2011): «CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes», in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics (CONLL Shared Task '11), 1-27.

PRADHAN S. *ET AL.* (2012): «CoNLL-2012 Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes», in *Proceedings of the Joint EMNLP-CoNLL conference*. Stroudsburg, Pennsylvania, USA: Association for Computational Linguistics, 1-40.

RECASENS M. ET HOVY E. (2011): «BLANC: Implementing the Rand Index for Coreference Evaluation», *Natural Language Engineering*, 17(4), 485-510. doi: 10.1017/S135132491000029X.

RECASENS M., HOVY E. ET MARTÍ M. A. (2011): «Identity, non-identity, and near-identity: Addressing the complexity of coreference», *Lingua*, 121(6), 1138-1152. doi: 10.1016/j.lingua.2011.02.004.

RUMELHART D. E., HINTON G. E. ET WILLIAMS R. J. (1986): «Learning representations by back-propagating errors», *Nature*, 323(6088), 533-536. doi: 10.1038/323533a0.

SOON W. M., NG H. T. ET LIM C. Y. (2001): «A Machine Learning Approach to Coreference Resolution of Noun Phrases», *Computational Linguistics*, 27(4), 521-544. doi: 10.1162/089120101753342653.

WILKENS R. *ET AL.* (2020): «French coreference for spoken and written language», in *Proceedings of the 13th Language Resources and Evaluation Conference*. European Language Resource Association, 80-89.

WINOGRAD T. (1972): «Understanding natural language», *Cognitive Psychology*, 3(1), 1-191. doi: 10.1016/0010-0285(72)90002-3.

ZEMAN D. *ET AL.* (2018): «CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies», in *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Association for Computational Linguistics, 1-21.