Évaluation comparative de systèmes neuronal et statistique pour la résolution de coréférence en langage parlé

Maëlle Brassier[1] Loïc Grobol[2] Théo Azzouza[1] Jean-Yves Antoine[1] Anaïs Halftermeyer[3]

(1) LIFAT, 3 place Jean Jaurès, 41000 Blois, France
(2) MoDyCo, 200 Av. de la République 401B, 92001 Nanterre, France
(3) LIFO, 6 Rue Léonard de Vinci, 45067 Orléans, France
maelle.brassier@etu.univ-tours.fr
{theo.azzouza, jean-yves.antoine}@univ-tours.fr,
loic.grobol@ens.psl.eu,
anais.halftermeyer@univ-orleans.fr

RÉSUMÉ

Nous présentons OFCoRS, un système de résolution de coréférence, basé sur le français parlé et un ensemble de modèles Random Forest. L'objectif de ce papier est de comparer l'approche statistique d'OFCoRS avec l'approche neuronale du système DeCoFre. Nous soulignons particulièrement les similarités et différences entre les deux systèmes. Nous comparons ensuite leurs performances sur le corpus français ANCOR et observons que les performances d'OFCoRS s'approchent de celles de DeCoFre. Une analyse détaillée montre également que les deux systèmes affichent de faibles performances sur les coréférences indirectes, montrant ainsi qu'on ne peut pas considérer le traitement des anaphores complexes comme un problème résolu.

ABSTRACT

Comparative evaluation of neural and statistical coreference resolution on spoken language

We introduce OFCoRS, a spoken French coreference resolution system based on Random Forest models. This paper aims at comparing the statistical approach of OFCoRS with DeCoFre, a neural based system. We put the emphasis on the similarities and differences between the two systems. We then compare both systems on the ANCOR corpus and observe that the performances of OFCoRS come close to the performances of DeCoFre. Additionally, in light of the low results on complex anaphora, we argue that the resolution of indirect anaphora is not completely resolved.

MOTS-CLÉS : résolution de coréférence, modèle neuronal, classification statistique, langage parlé.

KEYWORDS: coreference resolution, neural model, statistical classification, spoken language.

1 Introduction

L'avancée des techniques d'apprentissage profond (*Deep Learning*) a fortement impacté l'ensemble des domaines de l'Intelligence Artificielle et notamment le domaine du TAL, apportant de nombreuses améliorations que ce soit en termes de performances ou via l'affranchissement du travail conséquent de *feature engineering*. Cependant, les bénéfices des réseaux de neurones ne doivent pas éclipser certaines problématiques. L'une des principales préoccupations provient du manque d'explicabilité

des modèles neuronaux. Bien que l'explicabilité soit une source d'inquiétude évidente d'un point de vue éthique, son impact économique semble être largement sous-estimé. En effet, que ce soit dans leurs relations b2b ou b2c, les entreprises ont besoin d'introduire une dimension explicative dans leurs produits et leurs offres. Dans ce travail, nous ne remettons pas fondamentalement en question la contribution du *Deep Learning* au sein d'applications TAL mais préférons étudier si les systèmes neuronaux sont toujours l'approche la plus adaptée, comparé à des méthodes qui ont tendance à offrir une meilleure interprétation.

Dans le cadre d'une tâche de résolution de coréférences, notre contribution repose sur la comparaison des performances de deux systèmes se basant sur (1) une approche neuronale (système DeCoFre) et (2) une approche de classification statistique basée sur des Random Forest (système OFCoRS [1]). Nous détaillons dans un premier temps l'architecture générale des systèmes OFCoRS et DeCoFre, ainsi que les différentes étapes qui les composent. Nous décrivons ensuite conjointement la méthode d'évaluation que nous avons adoptée et le corpus de français oral ANCOR sur lequel ont été évalués les deux systèmes. Enfin, nous mettons en exergue les résultats que nous obtenons et tirons un ensemble de conclusions que nous estimons pouvoir être généralisé à d'autres domaines du TAL.

2 Description des systèmes

2.1 DeCoFre

Le système de Grobol (2019) est un système de résolution de coréférence end-to-end développé pour le français parlé. Inspiré des architectures proposées par Lee *et al.* (2017) et Lee *et al.* (2018), DeCoFre repose sur le paradigme de recherche d'antécédent (Denis & Baldridge, 2008) dont l'objectif est de trouver, pour chaque mention d'un document, l'antécédent le plus vraisemblable parmi un ensemble de candidats. DeCoFre utilise des vecteurs de mots à haute dimension – contextuels ou non-contextuels – qui permettent de disposer d'une représentation de mots, utilisée par la suite par l'architecture du système afin d'obtenir des représentations vectorielles d'empans de mots. Contrairement à Lee *et al.* (2017), le système passe par une étape de détection de mentions explicite via un classifieur d'empans de mots, ce qui permet de détecter facilement les mentions singletons et réduit la complexité de la recherche d'antécédent [2]. Nous utilisons ici l'implémentation de Grobol (2019) ainsi qu'un modèle utilisant des plongements de mots contextuels, développés pour Grobol (2020).

2.2 OFCoRS

Le système de résolution statistique OFCoRS suit un pipeline constitué de plusieurs étapes successives. La première étape consiste à détecter les mentions présentes dans le texte d'entrée. OFCoRS utilise pour cela le détecteur de mentions fourni par DeCoFre. Une fois l'ensemble des mentions détectées, une étape de génération de paires candidates (a, c) consiste à créer l'ensemble de toutes les paires de mentions à partir de l'ensemble M des mentions détectées dans le texte : $\{(a, c) \mid a, c \in M\}$. Nous choisissons de restreindre cette génération via un mécanisme de filtrage par fenêtre où seules

1. https://gitlab.com/Stanoy/ofcors

2. Dans le cadre de cette expérimentation, cette étape de détection n'est pas évaluée puisque nous nous attachons à la résolution de paires de coréférence et non à la détection des mentions. Nous présentons alors le même ensemble de mentions gold aux deux systèmes, comme évoqué en Section 3

les mentions séparées par une distance (en termes de nombre de mentions) inférieure à un seuil W sont considérées : $\{(a, c) \mid a, c \in M \wedge d(a, c) < W\}$. La taille de fenêtre W a été définie de façon empirique, avec comme valeur $W = 30$.

Les mentions et paires de mentions sont ensuite décrites par un ensemble de traits. Cet ensemble de traits s'inspire majoritairement de travaux portés sur des modèles *pair-wise* statistiques tels que (Désoyer *et al.*, 2015) et se compose de deux catégories de traits : les traits non-relationnels, caractérisant une mention de façon individuelle – tels que des traits morphologiques et syntaxiques (genre, nombre, entité nommée... de la mention) –, et les traits relationnels, définissant les relations entre deux mentions d'une paire – comme l'identité entre les traits non-relationnels des mentions ou bien la distance entre ces dernières. Néanmoins, afin de s'affranchir des annotations gold, trop riches d'un point de vue sémantique et pragmatique au vu des résultats que produiraient des systèmes automatiques, nous avons souhaité appauvrir certains des traits annotés. C'est notamment le cas de certaines mentions de type pronom (tels que "y", "on", "en") dont les traits de genre, nombre et type d'entité nommé ne peuvent être inférés de façon automatique et sans résoudre au préalable le lien de coréférence.

Afin d'inclure une information lexico-sémantique au sein de notre étape de classification de paires, nous ajoutons en sus de (Désoyer *et al.*, 2015) des plongements de mots, qui représentent les têtes syntaxiques des mentions. Nous considérons deux types de plongements : contextuels et non-contextuels (calculés respectivement avec les modèles de langage français pré-entraînés `fr_dep_news_trf` et `fr_core_news_lg` de Spacy). Grâce à eux, nous pouvons ajouter un trait de similarité cosinus entre les plongements des têtes lexicales, enrichissant ainsi le traitement de certains types de relation de coréférence d'une dimension sémantique.

À partir de ces traits, un classifieur s'attache ensuite à prédire si une paire candidate de mentions est coréférente ou non. L'étape de classification de paires repose soit sur un classifieur unique traitant indifféremment tous les types de coréférences, soit sur un multi-classifieur. Ce dernier est composé de trois classifieurs, chacun spécialisé sur la détection d'un des trois types de coréférence (avec l'antécédent et la *reprise*) :

- **anaphore pronominale**, lorsque la reprise est un pronom
  Exemple : La voiture est garée dehors. *Elle* est sale.
- **coréférence directe**, lorsque les deux mentions de la paire partagent la même tête lexicale
  Exemple : La voiture est garée dehors. *Cette voiture* est sale.
- **coréférence indirecte**, lorsque les deux mentions de la paire ne partagent pas la même tête et que la reprise n'est pas un pronom
  Exemple : La voiture est garée dehors. *Le véhicule* est sale.

Après avoir évalué plusieurs méthodes de classification différentes, notre choix s'est porté sur un modèle *Random Forest* puisqu'il propose les meilleures performances, tout en bénéficiant d'un caractère interprétable. Enfin, la dernière étape du pipeline consiste à construire une chaîne de coréférences, reliant toutes les mentions faisant référence à une seule et même entité. Ce processus repose sur un algorithme de chaînage qui s'applique sur toutes les paires classifiées comme coréférentes lors de l'étape précédente. Il arrive alors qu'une mention m se retrouve avec k antécédents potentiels. Deux approches sont envisageables pour résoudre de tels conflits d'antécédents : une approche *closest-first* qui choisit l'antécédent le plus proche de la mention considérée (Soon *et al.*, 2001) ou une approche *best-first* qui à l'inverse considérera l'antécédent ayant le plus haut score de probabilité (Aone & William, 1995; Ng & Cardie, 2002). OFCoRS opte pour l'algorithme de *best-first* pour son étape de

création de chaîne. De plus, si une mention m ne possède aucun antécédent candidat, alors m est considéré comme un singleton et n'appartient à aucune chaîne.

3 Méthodologie d'évaluation

Nos expérimentations se sont conduites sur un jeu de données unique : le corpus ANCOR, qui est le plus large corpus de français parlé annoté en coréférence (Muzerelle *et al.*, 2014). Le jeu de données comporte quatre sous-corpus de conversations, chacun présentant des situations d'interaction et degrés d'interactivité différents. Nous divisons ce corpus en trois ensembles d'apprentissage, développement et test, correspondant respectivement à 65 %, 14 % et 21 % du corpus d'origine en termes de mentions. Les ensembles de développement et test sont également re-découpés chacun en trois sous-ensembles, pour l'étude de la significativité statistique des résultats. Les différents modèles que nous avons évalués –et qui varient selon les hyper-paramètres, architecture et utilisation des plongements de mots– sont présentés dans la table 1. Ici, chaque modèle nommé "simple" représente un classifieur unique pour tous les types de relation tandis que les modèles "multi" utilisent un multi-classifieur (trois classifieurs chacun dédié à un type de relation). Les valeurs "contextual" et "non-contextual" correspondent à l'utilisation de plongements de mots et des traits associés, avec respectivement des plongements contextuels et non-contextuels. Ainsi, notre baseline correspond au modèle *OFCoRS-simple* qui fait appel à un classifieur unique, sans utilisation de traits reposant sur des plongements de mots.

L'évaluation de nos modèles de résolution s'est basée sur une implémentation alternative du scorer CoNLL-2012 (Pradhan *et al.*, 2014), à savoir le scorer Scorch (Grobol, 2020) qui ré-implémente directement les définitions princeps des métriques MUC (Vilain *et al.*, 1995), B^3 (Bagga & Baldwin, 1998) and $CEAF_e$ (Luo, 2005) dont CoNLL-2012 est la moyenne arithmétique. La qualité de la détection des paires coréférentes (et non des chaînes de coréférence complètes) sera évaluée en termes de précision, rappel et F-score puisque cette tâche peut s'apparenter à une tâche de classification binaire.

Notre comparaison se concentrant uniquement sur l'évaluation de la classification de paires, nous souhaitons gommer l'impact de l'étape de détection de mentions. C'est pourquoi nous utilisons ici les mentions gold extraites directement du corpus ANCOR, pour les deux système OFCorS et DeCoFre. Nous exploitons également les traits gold non-relationnels des mentions, utilisés par la suite lors du calcul des traits relationnels. Seul le trait de similarité cosinus des plongements des têtes lexicales repose sur un modèle de langage externe, comme décrit précédemment.

4 Résultats

4.1 Comparaison des modèles OFCoRS

La table 1 détaille les meilleures performances obtenues par différentes versions du système statistique OFCoRS sur le *devset* d'ANCOR. Cette comparaison nous permet d'évaluer l'impact de l'ensemble des configurations que nous avons étudiées : à savoir l'intégration de plongements de mots comme trait sémantique ainsi que l'utilisation de plusieurs classifieurs, chacun dédié à un type de coréférence.

Model	devset	subdevset0	subdevset1	subdevset2
OFCoRS-simple (baseline)	75,67	75,35	74,69	76,79
OFCoRS-simple-non-contextual	77,45	77,14	77,11	77,95
OFCoRS-multi	77,65	77,38	77,28	78,13
OFCoRS-multi-non-contextual	**78,25**	**78,17**	**77,78**	**78,68**
OFCoRS-multi-contextual	77,93	78,13	77,17	78,36
DeCoFre	**80,70**	**81,47**	**79,64**	**81,49**

TABLE 1 – Comparaison des performances (score CoNLL-2012 par *scorch*) des différents modèles OFCoRS et DeCoFre sur l'ensemble *devset* et ses sous-ensembles.

Modèle	testset	subtestset0	subtestset1	subtestset2
OFCoRS-multi-non-contextual	78,22	77,15	79,14	78,27
DeCoFre	81,77	81,85	81,12	82,45

TABLE 2 – Comparaison des performances (score CoNLL-2012 par *scorch*) entre le meilleur modèle d'OFCoRS et le modèle DeCoFre sur l'ensemble *testset* et ses sous-ensembles.

Nos résultats mettent en avant le bénéfice des plongements de mots non-contextuels ainsi que de l'approche multi-classifieur : en effet, le modèle *OFCoRS-simple-non-contextual* surpasse la baseline de 1,78 points pour la métrique CoNLL-2012 , tandis qu'une augmentation de performances similaire (1,98 points) est observée avec le modèle *OFCoRS-multi*. Ces deux améliorations sont significatives statistiquement (valeur-p de respectivement 0,0389 et 0,0316) [3]. De plus, nous observons que ce gain est cumulatif : le modèle *OFCoRS-multi-non-contextual* qui combine l'ajout des plongements de mots non-contextuels et l'approche multi-classifieur affiche les meilleurs performances, atteignant un score CoNLL-2012 de 78,25 sur le *devset*, ce qui représente un gain de 2,58 points par rapport à la baseline (différence significative d'une valeur-p de 0,0190). Une étude de l'importance respective des traits pour chaque classifieur confirme l'intérêt de l'approche multi-classifieur et sera détaillée en section 4.3.

4.2 OFCoRS vs. DeCoFre

La table 2 compare la meilleure architecture d'OFCoRS, *OFCoRS-multi-contextual*, avec le modèle de DeCoFre sur l'ensemble *testset* d'ANCOR. Bien que nous constatons que DeCoFre surpasse OFCoRS avec une différence de +3.55 points du score CoNLL-2012 sur le *testset*, leur différence de performances n'écarte pas l'approche classification statistique qui peut être considérée comme une approche efficace et viable pour certains contextes applicatifs. Nous remarquons que les anaphores pronominales affichent les meilleurs résultats de détection, avec une moyenne de F_1-score supérieure à 70 % pour les deux modèles et sur l'ensemble des jeux de données, comme observé dans la table 3. Conjointement, les relations directes obtiennent des résultats satisfaisants, proches de ceux des pronominales bien qu'avec une baisse de rappel relative (d'environ 8 % pour OFCoRS et 9 % pour

3. Ces valeurs sont obtenues via un test-t de Student paramétrique, bien que le manque d'échantillons nous empêche de tirer des conclusions concernant la normalité de la distribution.

Dataset	Modèle	Pronominal			Direct			Indirect		
		P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
testset	OFCoRS	69,06	72,89	70,92	70,91	64,41	67,50	42,05	21,87	28,77
	DeCoFre	81,77	74,79	78,12	83,80	65,11	73,28	48,18	23,61	31,69
subtestset0	OFCoRS	64,26	65,81	65,03	68,52	65,45	66,95	40,07	19,90	26,59
	DeCoFre	82,76	72,73	77,42	83,22	65,31	73,19	48,30	22,83	31,00
subtestset1	OFCoRS	72,92	77,66	75,22	74,15	64,97	69,27	45,62	25,71	32,89
	DeCoFre	80,36	73,30	76,67	84,44	64,53	73,15	44,38	23,05	30,34
subtestset2	OFCoRS	69,50	74,90	72,10	70,01	62,76	66,19	39,60	19,64	26,26
	DeCoFre	84,10	77,29	80,55	85,31	66,28	74,60	54,12	22,55	31,84

TABLE 3 – Comparaison des performances (Précision, Rappel et F_1-score) entre le meilleur modèle d'OFCoRS et le modèle DeCoFre, par type de relation et sur l'ensemble testset et ses sous-ensembles.

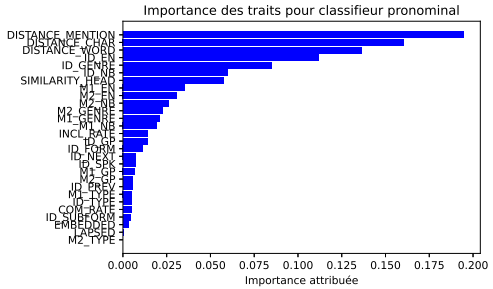
DeCoFre). En revanche, les coréférences indirectes apparaissent comme un problème non-résolu, tout du moins dans le langage français oral : en effet, les deux systèmes peinent à atteindre une précision de 50 % alors que le rappel, lui, stagne dans une fenêtre de 20 % à 25 %. Cette faiblesse des modèles peut être expliquée par la complexité des phénomènes sémantiques (synonymie, hyponymie, hyperonymie...) que mettent en jeu les relations indirectes.

Si nous considérons chaque système séparément, nous observons que DeCoFre dépasse OFCoRS majoritairement sur les scores de précision, et ce pour chaque type de relation. Ce fossé est d'autant plus significatif pour les relations pronominales (12,71 %) et directes (12,89 %) sur l'ensemble *testset*. La différence est moindre et plus variable (avec un écart de 6,13 % sur le *testset* et OFCoRS surpassant DeCoFre de 1,24 % sur le *subtestset1*) pour les indirectes. Néanmoins, les performances des deux systèmes restent basses dans ce cas de figure. Enfin, nous pouvons noter que les résultats de rappel sont relativement similaires pour les deux systèmes et même observer qu'OFCoRS dépasse DeCoFre sur le *subtestset1*.

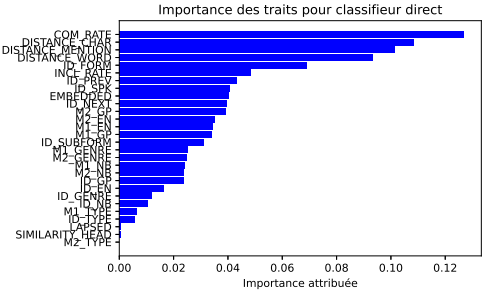
4.3 Analyse qualitative

Après avoir comparé les performances brutes d'OFCoRS et DeCoFre, nous réalisons une analyse qualitative des résultats afin de proposer un faisceau d'explications quant aux performances des systèmes. En dépit des bonnes performances décrites précédemment, il apparaît que la résolution de coréférences pronominales n'est pas aussi facile qu'il n'y paraît. Si les coréférences mettant en jeu des pronoms personnels sont correctement résolues par les deux systèmes, les pronoms démonstratifs et relatifs posent encore de nombreuses difficultés : ils représentent ainsi la majorité des erreurs (False Positive et False Negative) concernant les coréférences pronominales. Par exemple, dans la phrase « *Le musée à côté de la poste, ça ouvre à quelle heure ?* », le pronom "ça" ne porte aucune information morphologique ou sémantique d'un point de vue lexical. Cela rend sa résolution particulièrement difficile . De facto, nos deux systèmes rencontrent de fréquentes difficultés avec ce pronom.

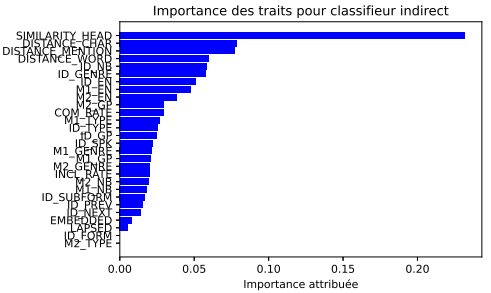
Les sous-figures de la figure 1 présentent l'importance de chaque trait donnée par l'un des classifieurs. L'analyse de l'importance respective de ces traits est une indication du caractère explicatif de notre modèle. Les variations d'importance de ces traits d'un classifieur à l'autre illustrent par ailleurs



(a) Importance des traits attribuée par le classifieur *OFCoRS-multi-non-contextual* traitant les relations pronominales



(b) Importance des traits attribuée par le classifieur *OFCoRS-multi-non-contextual* traitant les relations directes



(c) Importance des traits attribuée par le classifieur *OFCoRS-multi-non-contextual* traitant les relations indirectes

FIGURE 1 – Ensemble des figures représentant l'importance des traits attribuée par chaque classifieur

la capacité du multi-classifieur à s'adapter au mieux à chaque type de relation. La sous-figure 1a montre ainsi que le classifieur pronominal accorde une importance prépondérante aux traits de distance (`DISTANCE_*`)[4], suivi par le trait d'identité de genre (`*_GENRE`) et nombre (`*_NB`). Cette observation est assez intuitive, puisque les pronoms ne portent pas de sémantique lexicale propre, de même que leur forme en terme de séquence de caractères est de peu d'utilité. Les coréférences pronominales sont ainsi formées de paires de mentions partageant le même genre et nombre, tout en ne pouvant pas reposer sur des traits de proximité de formes (`*_RATE`)[5]. À l'inverse, la dimension sémantique est primordiale lors de la résolution de coréférence indirecte, où les têtes lexicales des mentions peuvent partager une relation sémantique complexe telle que la synonymie ou l'hyperonymie. On vérifie effectivement (sous-figure 1c) que le classifieur indirect accorde la plus haute importance au trait de similarité cosinus (`SIMILARITY HEAD`) et beaucoup moins aux autres, particulièrement les traits morpho-syntaxiques (`*_GENRE, *_TYPE`)). Enfin, la sous-figure 1b nous montre que les traits de forme (`*_RATE`) se trouvent parmi les traits les plus discriminants pour le classifieur direct, dû au fait que les mentions des coréférences directes partagent les mêmes têtes lexicales.

Il est intéressant de noter que les traits de distance contribuent de façon significative à l'ensemble

4. ici, * correspond aux différentes types de distance : par mots, par caractères et par mentions. Pour les traits de type *_FEATURE, * correspond à l'une des deux mentions de la paire : M1 ou M2.

5. ici, les traits *_RATE englobent les traits INCL_RATE et COM_RATE qui sont respectivement le taux d'inclusion de la plus petite mention dans la plus grande et le taux de tokens en commun entre les deux mentions.

des classifieurs. Pour le classifieur pronominal, le taux d'importance s'aligne avec la distribution du corpus ANCOR où les coréférences pronominales tendent à être très locales : 85 % d'entre elles sont composées de mentions séparées par au plus 5 mentions ; alors que les distances pour les paires directes et indirectes sont plus variables.

Enfin, l'utilisation des plongements de mots a largement été éprouvée par la communauté pour leur capacité à identifier des relations sémantiques entre les mots, tout en capturant des informations syntaxiques et sémantiques au niveau des mots. Nos résultats confirment le bénéfice d'un trait de similarité cosinus sur les plongements de mots contextuels. De façon surprenante, le modèle *OFCoRS-multi-contextual* qui fait appel à ces plongements contextuels ne parvient pas à surpasser un modèle avec des plongements non-contextuels (77.93 vs. 78.25 en score CoNLL-2012 ; pas de différence significative : valeur-p = 0.1884). Cette absence de gain de performance peut provenir de la fenêtre de contexte que nous considérons (c-à-d un tour de parole complet) ainsi que la nature du corpus ANCOR, qui est principalement composé du sous-corpus ESLO_ANCOR. Dans ce sous-corpus, les documents traitent de longues conversations avec des tours de paroles pouvant être très longs, ce qui peut potentiellement rendre le contexte non pertinent. De plus, le français parlé inclut beaucoup d'artefacts qui ne sont pas présents dans le langage écrit, tels que des répétitions ou d'autres disfluences de la parole. Ces dernières peuvent être ainsi non-présentes ou alors extrêmement rares dans les données utilisées pour l'entraînement des plongements de mots pré-entraînés. Une étude plus poussée sera nécessaire concernant le choix des plongements utilisés.

5 Conclusion

Dans ce travail nous avons introduit OFCoRS, un système de résolution de coréférence statistique et avons comparé ses performances avec un système neuronal, DeCoFre. L'objectif de cette comparaison est double : à la fois comparer les performances d'un modèle statistique reposant sur des traits implémentés manuellement contre une architecture neuronale utilisant des traits profonds afin de mettre en exergue les différences et similarités des deux approches et évaluer si les réseaux de neurones doivent être toujours préférés. En termes de score CoNLL-2012, nous observons que les performances d'OFCoRS se rapprochent de ceux de DeCoFre. Néanmoins, nous observons une différence plus importante au niveau des valeurs de précision, particulièrement sur les relations de coréférence pronominales et directes mais bien plus basse pour les indirectes. Cependant, les résultats des coréférences indirectes démontrent avant tout que la résolution des coréférences indirectes demeure une problématique scientifique encore non résolue.

Au regard des niveaux proches de performances observés avec OFCoRS et DeCoFre, nous estimons qu'il est pertinent de questionner systématiquement l'utilisation d'une approche de résolution plutôt qu'une autre, en considérant parallèlement leur coût de calcul et leur niveau de performances. Cette conclusion rejoint les observations de Poot & van Cranenburgh (2020) qui présentent des résultats de résolution de coréférence sur le néerlandais, où les performances d'un réseau de neurones s'avèrent inférieures à celles d'un système à base de règles. En termes d'interprétabilité, nous proposons une première tentative d'explication du modèle OFCoRS en observant l'importance de chaque trait accordée par chaque classifieur. Bien que cet aspect requiert des études supplémentaires, il permet de souligner les particularités de chaque type de relation de coréférence et l'intérêt de développer des modèles spécifiques à chacun d'entre eux. C'est dans cette direction que nous nous dirigeons, dans l'espoir d'améliorer davantage les performances de notre classifieur statistique et d'attaquer le problème des coréférences indirectes.

Références

AONE C. & WILLIAM S. (1995). Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies. In *33rd Annual Meeting of the Association for Computational Linguistics*, p. 122–129, Cambridge, Massachusetts, USA : Association for Computational Linguistics. DOI : 10.3115/981658.981675.

BAGGA A. & BALDWIN B. (1998). Algorithms for scoring coreference chains. In *Proceedings of the 1st International Conference on Language Resources and Evaluation*, p. 563–566.

DENIS P. & BALDRIDGE J. (2008). Specialized models and ranking for coreference resolution. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, p. 660, Honolulu, Hawai'i, USA : Association for Computational Linguistics. DOI : 10.3115/1613715.1613797.

DÉSOYER A., LANDRAGIN F. & TELLIER I. (2015). Apprentissage automatique d'un modèle de résolution de la coréférence à partir de données orales transcrites du français : le système CROC. In *Actes de la 22ème Conférence sur le Traitement Automatique des Langues Naturelles*, p. 439–445, Caen, France.

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*, p. 8–14, Minneapolis, Minnesota, USA.

GROBOL L. (2020). *Coreference resolution for spoken French*. Thèse de doctorat, Université Sorbonne Nouvelle, Paris, France.

LEE K., HE L., LEWIS M. & ZETTLEMOYER L. (2017). End-to-end Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 188–197, København, Danmark : Association for Computational Linguistics.

LEE K., HE L. & 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*, volume 2, p. 687–692, New Orleans, Louisiana, USA : Association for Computational Linguistics. DOI : 10.18653/v1/N18-2108.

LUO X. (2005). On coreference resolution performance metrics. In *Proceedings of the Human Language Technology Conference and the 2005 Conference on Empirical Methods in Natural Language Processing*, p. 25–32.

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. In ELRA, Éd., *LREC'2014, 9th Language Resources and Evaluation Conference.*, p. 843–847, Reyjavik, Iceland. HAL : hal-01075679.

NG V. & CARDIE C. (2002). Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 104–111, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073102.

POOT C. & VAN CRANENBURGH A. (2020). A Benchmark of Rule-Based and Neural Coreference Resolution in Dutch Novels and News. In *Proceedings of the Third Workshop on Computational*

Models of Reference, Anaphora and Coreference, p. 79–90, Barcelona, España : Association for Computational Linguistics.

PRADHAN S., RECASENS X. L. M., HOVY E., NG V. & STRUBE M. (2014). Scoring coreference partitions of predicted mentions : A reference implementation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 30–35.

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

VILAIN M., BURGER J., ABERDEEN J., CONNOLLY D. & HIRSCHMAN L. (1995). A modeltheoretic coreference scoring scheme. In *Proceedings of the 6th Message Understanding Conference (MUC-6)*, p. 45–52.