Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Modèle neuronal pour la résolution de la coréférence dans les dossiers médicaux électroniques

Julien Tourille[1] Olivier Ferret[1] Aurélie Névéol[2] Xavier Tannier[3]
(1) CEA, LIST, Laboratoire Analyse Sémantique Texte et Image, Gif-sur-Yvette, F-91191, France
(2) LIMSI, CNRS, Université Paris-Saclay
(3) Sorbonne Université, Inserm, LIMICS
{julien.tourille,olivier.ferret}@cea.fr, aurelie.neveol@limsi.fr,
xavier.tannier@sorbonne-universite.fr

RÉSUMÉ

La résolution de la coréférence est un élément essentiel pour la constitution automatique de chronologies médicales à partir des dossiers médicaux électroniques. Dans ce travail, nous présentons une approche neuronale pour la résolution de la coréférence dans des textes médicaux écrits en anglais pour les entités générales et cliniques en nous évaluant dans le cadre de référence pour cette tâche que constitue la tâche 1C de la campagne i2b2 2011.

ABSTRACT

Neural approach for coreference resolution in electronic health records

Coreference resolution is an essential step for clinical timeline extraction from electronic health records. Herein, we present a neural approach for coreference resolution in clinical documents written in English for both general and clinical entities and we evaluate it in the reference evaluation framework of the task 1C of the i2b2 2011 campaign.

MOTS-CLÉS : TAL clinique, réseaux de neurones, résolution de la coréférence.

KEYWORDS: clinical NLP, deep learning, coreference resolution.

1 Introduction

La résolution de la coréférence consiste à identifier toutes les mentions d'entités ou d'événements et à les regrouper en classes d'équivalence (Pradhan *et al.*, 2011). Cette tâche n'implique pas de déterminer à quels entités ou événements ces mentions font référence. Il s'agit de déterminer si plusieurs mentions font référence à la même entité ou événement. La résolution de la coréférence a principalement été explorée pour des textes de nature journalistique, avec de premières campagnes d'évaluation dans les années 1990 (Sundheim, 1995; Hirschman & Chinchor, 1998). Ce n'est qu'au cours des dix dernières années que des travaux se sont intéressés à la résolution de la coréférence dans le domaine clinique. La première campagne d'évaluation sur le sujet a été proposée par la fondation i2b2 (Uzuner *et al.*, 2012) et a permis le développement de plusieurs modèles (Jindal & Roth, 2013; Hinote *et al.*, 2011; Grouin *et al.*, 2011; Chowdhury & Zweigenbaum, 2013). L'intérêt pour cette tâche répond au besoin grandissant d'explorer et d'utiliser les données contenues dans les rapports et autres documents textuels qui composent les dossiers médicaux électroniques. Parmi ces données, la

chronologie médicale, c'est-à-dire la suite d'événements médicaux qui ont lieu au cours de la vie d'un patient, est une information importante. Être en mesure d'extraire automatiquement ces chronologies permettrait de mieux comprendre certains phénomènes médicaux tels que l'évolution des maladies et les effets longitudinaux des médicaments (Lin *et al.*, 2016; Sun *et al.*, 2013). Or, les événements médicaux sont mentionnés plusieurs fois dans les dossiers, rendant difficile la construction de ces chronologies sans une étape de résolution de la coréférence.

Dans ce travail, nous nous intéressons à la résolution de la coréférence dans des dossiers médicaux électroniques écrits en anglais. Nous proposons une approche neuronale fondée sur les travaux récents appliqués aux textes journalistique (Wiseman *et al.*, 2016; Clark & Manning, 2016).

2 Données

Dans ce travail, nous utilisons le corpus i2b2 tâche 1C (Uzuner *et al.*, 2012) et plus précisément, la partie i2b2/VA ne contenant pas de documents de University of Pittsburgh Medical Center (UPMC). Nous reproduisons ainsi un des contextes expérimentaux proposés lors de la campagne d'évaluation. Le corpus contient 194 documents cliniques du Beth Israel Deaconess Medical Center (BETH) et 230 documents cliniques de Partners Healthcare (PARTNERS) (cf. tableau 1). Le tableau 2 présente le nombre de chaînes de coréférence ainsi que leur longueur moyenne et maximale.

Institution	Train	Test	Total
BETH	115	79	194
PARTNERS	136	94	230
Combinés	251	173	424

TABLE 1: Statistiques concernant le corpus i2b2/VA tâche 1C

Institution	Nombre chaînes	Long. moy.	Long. max.
BETH	1 816	4,2	122
PARTNERS	1 395	4,4	105

TABLE 2: Statistiques concernant le nombre et la longueur des chaînes de coréférence du corpus i2b2/VA tâche 1C

Cinq types d'éléments sont annotés dans le corpus en y incluant les singletons, *i.e.* les éléments non coréférents : personnes, pronoms, tests, traitements et problèmes. Les chaînes de coréférence peuvent être divisées en deux groupes : les chaînes relatives aux *événements* (tests, traitements et problèmes) et les chaînes relatives aux *personnes*. Les deux groupes présentent des caractéristiques différentes : les chaînes *personnes* sont plus longues (moyenne de 12,44 contre environ 2,5 pour les événements) tandis que les mentions composant les chaînes *personnes* prennent généralement la forme de pronoms personnels, même si des pronoms peuvent également faire partie des chaînes *événements*. Enfin, les événements médicaux ont une structure argumentale implicite que les mentions de personnes n'ont pas (Styler IV *et al.*, 2014). Ces trois différences nous ont amené à considérer la résolution de la coréférence pour ces deux ensembles d'entités comme des sous-tâches distinctes. En conséquence, nous apprenons deux modèles distincts, un pour chaque sous-ensemble. Nous faisons l'hypothèse que le modèle sera capable de distinguer quels pronoms appartiennent aux deux sous-ensembles et nous les incluons dans les modèles caractérisant chaque sous-ensemble.

3 Description du modèle

Notre modèle s'inspire des approches neuronales récemment développées (Lee *et al.*, 2017; Wiseman *et al.*, 2016). Le composant principal de notre approche est un modèle de type *mention-ranking* (Denis & Baldridge, 2008) dont l'objectif est d'ordonnancer l'ensemble des antécédents possibles pour une mention donnée et de choisir le premier. Le cas non-anaphorique, c'est-à-dire l'absence d'antécédent pour une mention donnée, est géré par l'utilisation d'un *dummy antecedent* (Durrett & Klein, 2013; Wiseman *et al.*, 2016).

Notre approche diffère des modèles locaux pour la résolution de la coréférence dans lesquels les paires de mentions sont considérées séparément. Nous proposons d'utiliser des traits globaux extraits des chaînes de coréférence en cours de construction. Ainsi, notre approche s'inspire et s'inscrit dans une lignée de travaux récents examinant l'utilisation de ce type de traits dans des approches neuronales (Clark & Manning, 2015, 2016; Wiseman *et al.*, 2016, 2015).

Plus spécifiquement, nous incorporons de l'information concernant les chaînes de coréférence en cours de construction dans notre modèle *mention-ranking*. Cette information est construite en utilisant un LSTM (Hochreiter & Schmidhuber, 1997) qui examine les différentes mentions des chaînes par ordre d'apparition dans le texte. Le principal avantage de cette approche est qu'elle facilite l'inférence en ne requérant qu'une seule passe sur les mentions (de gauche à droite).

Plongements en entrée Les plongements utilisés en entrée de notre modèle sont construits en concaténant une représentation dense des caractères et un vecteur de mot pré-calculé sur un grand corpus. La représentation dense des caractères est construite suivant la méthode proposée par Lample *et al.* (2016). Un plongement aléatoire est d'abord généré pour chaque caractère présent dans le corpus. Ensuite, les caractères des différents tokens passent à travers un Bi-LSTM. Les deux représentations denses résultantes sont enfin concaténées pour former la représentation finale.

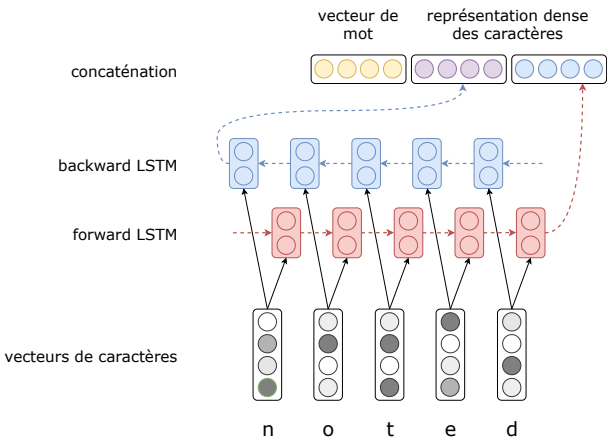

FIGURE 1: Construction des plongements en entrée de notre modèle

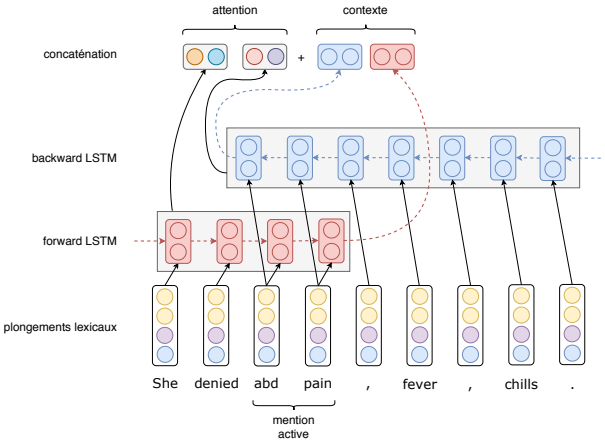

FIGURE 2: Construction des représentations des mentions

Représentation des mentions Le processus de construction des représentations des mentions est présenté à la figure 2. Dans notre exemple, la mention considérée est *abd pain* dans le contexte de la

phrase *She denied abd pain, fever, chills*. Tout d'abord, notre modèle calcule deux représentations contextuelles de la mention considérée grâce à un Bi-LSTM. Le *forward* LSTM prend en entrée le segment allant du premier token de la phrase jusqu'au dernier token de la mention. Le *backward* LSTM prend en entrée le segment allant du dernier token de la phrase jusqu'au premier token de la mention. Les deux représentations denses forment la première partie de la représentation finale.

Nous ajoutons ensuite une représentation dense issue d'un mécanisme d'attention. Ce mécanisme, qui permet d'accorder une importante différenciée aux deux éléments constituant la représentation contextuelle d'une mention, consiste en une somme pondérée des états cachés intermédiaires des deux LSTMs. Les poids utilisés dans la somme pondérée sont calculés en utilisant un réseau de neurones *feed-forward*.

Représentation des chaînes de coréférence Pour le calcul des représentations denses des chaînes de coréférence, nous utilisons, à l'instar de Wiseman *et al.* (2016), un LSTM prenant en entrée les différentes mentions composant les chaînes de coréférence, dans l'ordre de leur apparition dans le document. Un aperçu du processus complet est présenté à la figure 3. Bien entendu, nous maintenons au cours de l'analyse d'un document autant de ces représentations que de chaînes de coréférence, représentations construites en utilisant le même LSTM (*i.e.* les mêmes poids).

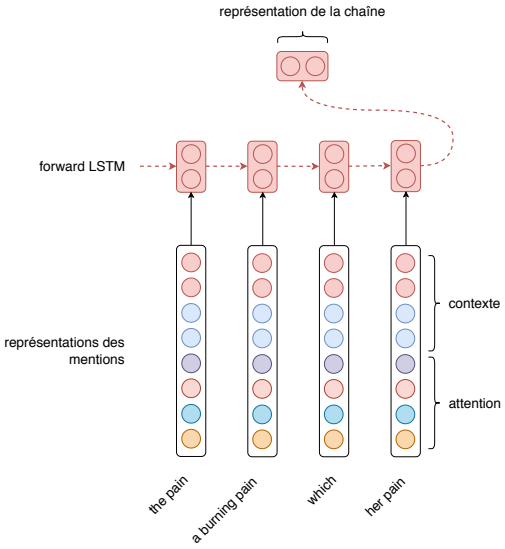

FIGURE 3: Calcul des représentations des chaînes de coréférence

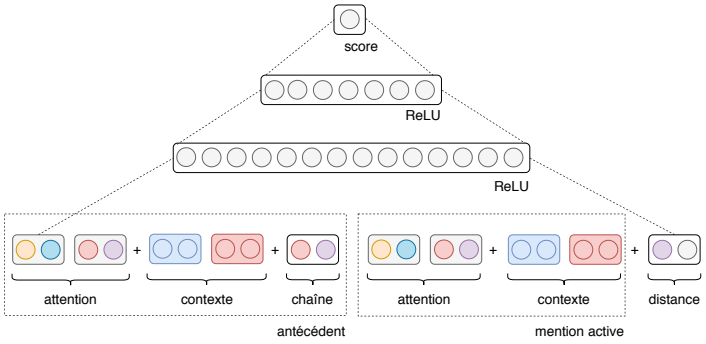

FIGURE 4: Classifieur de paires de mentions

Classifieur de paires de mentions Le rôle du classifieur est d'assigner un score à chaque paire (mention, antécédent). Pour cela, nous utilisons un réseau *feed-forward* qui prend en entrée la concaténation des représentations des mentions, de la chaîne de coréférence dont l'antécédent fait partie et d'une représentation dense de la distance qui sépare la mention active de l'antécédent considéré. La distance est discrétisée en un nombre fixe de catégories : $[1, 2, 3, 4, 5-7, 8-15, 16-31, 32-63, 64+]$ (Clark & Manning, 2016). Le réseau feed-forward est composé de deux couches cachées, chaque couche ayant pour taille la moitié de la taille de la couche précédente. Les scores obtenus pour chaque paire (mention, antécédent) sont concaténés dans un vecteur et nous appliquons une fonction *softmax*. Nous optimisons la vraisemblance de tous les antécédents inclus dans la chaîne de coréférence (Lee *et al.*, 2017).

4 Contexte expérimental

Dans ce travail, nous nous intéressons à la situation dans laquelle nous disposons d'une détection parfaite des mentions dans les documents en utilisant les annotations *gold* du corpus pour ces mentions. Nous testons notre modèle dans plusieurs configurations. Tout d'abord, nous implémentons un modèle de base en enlevant certaines parties de notre architecture. Les plongements présentés en entrée du modèle sont composés uniquement d'un vecteur de mot pré-entraîné. Le mécanisme d'attention ainsi que les représentations denses des chaînes de coréférence et des caractères ne sont pas utilisés.

En partant de cette configuration, nous activons chaque composant pour mesurer son effet sur la performance du système. Par ailleurs, nous implémentons une stratégie de pré-entraînement dans laquelle nous entraînons notre modèle uniquement sur les mentions coréférentes. La tâche se résume alors à trouver l'antécédent correct parmi les mentions précédentes. Cette stratégie permet généralement d'améliorer les performances des systèmes (Clark & Manning, 2015, 2016; Wiseman *et al.*, 2015, 2016). Ensuite, nous implémentons une spécialisation du classifieur de paires de mentions. Le réseau feed-forward qui le compose est unique pour chaque type d'entité. Notre hypothèse est que les éléments permettant d'identifier des mentions coréférentes peuvent différer selon le type d'entité considéré. Enfin, nous implémentons un filtrage des antécédents selon le type de la mention considérée par le modèle. Pour le cas des pronoms, nous considérons tous les antécédents.

Nous implémentons notre modèle avec PyTorch (Paszke *et al.*, 2017). La taille des LSTMs utilisés pour calculer les représentations des mentions et des chaînes de coréférence est fixée à 100. Celle du Bi-LSTM utilisé pour la représentation des caractères est fixée à 25. Les plongements de caractères ont une taille de 25 et sont initialisés aléatoirement. Notre modèle est entraîné sur des mini-lots de 1 document. Nous utilisons l'algorithme d'optimisation Adam et fixons le taux d'apprentissage à 0,001. Nous implémentons une décroissance du taux d'apprentissage de 1 % à chaque itération. Les plongements lexicaux pré-entraînés sont appris sur le corpus MIMIC-III (Johnson *et al.*, 2016) avec une taille de 100. Nous appliquons un *dropout* sur les couches cachées de notre classifieur de mentions avec un taux de 0,2. Un *dropout* est aussi appliqué sur les plongements en entrée avec un taux de 0,5. Nous implémentons l'apprentissage des états initiaux des LSTMs (Gers *et al.*, 2002). Finalement, nous répétons chaque expérience 10 fois pour prendre en considération l'aspect non-déterministe de notre modèle (Reimers & Gurevych, 2017).

5 Résultats

Le tableau 3 présente les résultats de nos expériences, obtenus grâce à la mesure CoNLL calculée avec l'outil de référence de Pradhan *et al.* (2014) [1]. Nous considérons uniquement les clusters regroupant au moins deux mentions et excluons ainsi, à l'instar d'une grande partie des travaux en domaine général, les singletons lors du calcul de la performance. L'utilisation d'un mécanisme d'attention

1. Nous avons préféré cet outil à celui de la campagne i2b2 dans la mesure où suite à divers travaux méthodologiques réalisés sur l'évaluation de la coréférence en domaine général (Pradhan *et al.*, 2011, 2014), il s'y est imposé comme un standard. En outre, Pradhan *et al.* (2014) soulignent que l'outil i2b2 s'appuie sur l'approche de Cai & Strube (2010), dont l'outil associé présente une erreur de mise en œuvre, peut-être également présente dans l'outil i2b2.

permet d'améliorer la performance des modèles *personnes* et *événements*, mais avec une contribution minime comme le suggère la faible différence par rapport au réseau initial. Aucun des deux modèles ne semble bénéficier de l'information issue des représentations denses des caractères, malgré leur contribution dans d'autres tâches (*e.g.* reconnaissance d'entités nommées). La prise en compte des chaînes de coréférence en cours de construction n'améliore pas la performance de notre modèle, avec une baisse supérieure à un point pour les événements. Ces résultats sont en contradiction avec ceux présentés par Wiseman *et al.* (2016). Cependant, les contextes expérimentaux diffèrent. Wiseman *et al.* (2016) travaillent sur des documents journalistiques et rapportent des résultats issus d'un seul run. Or Reimers & Gurevych (2017) suggèrent que les systèmes non déterministes tels que les modèles neuronaux doivent être évalués sur plusieurs runs afin de prendre en compte cette variabilité inhérente.

condition	P	R	F1		P	R	F1		P	R	F1	
	Personnes				Événements				Combinés			
baseline	87,77 (± 0,71)	82,82 (± 0,99)	85,17 (± 0,46)		65,15 (± 1,81)	54,62 (± 1,35)	59,30 (± 0,45)		76,89 (± 1,31)	67,94 (± 0,89)	72,02 (± 0,38)	
attention	88,08 (± 0,82)	83,42 (± 0,55)	85,64 (± 0,38)	↑	66,39 (± 1,45)	54,18 (± 2,08)	59,56 (± 0,86)	↑	77,77 (± 1,01)	67,93 (± 1,27)	72,41 (± 0,40)	↑
caractères	87,87 (± 0,98)	82,55 (± 0,69)	85,08 (± 0,21)	↓	66,39 (± 1,57)	53,26 (± 1,76)	59,02 (± 0,92)	↓	77,78 (± 1,14)	67,13 (± 1,01)	71,96 (± 0,46)	↓
chaîne	87,48 (± 0,77)	82,36 (± 0,72)	84,71 (± 0,49)	↓	65,86 (± 1,03)	52,12 (± 1,10)	58,11 (± 0,58)	↓	77,29 (± 0,69)	66,55 (± 0,64)	71,37 (± 0,37)	↓
filtrage	87,60 (± 1,48)	82,82 (± 0,89)	85,08 (± 0,39)	↓	65,42 (± 1,31)	54,37 (± 1,37)	59,30 (± 0,56)	=	76,98 (± 1,00)	67,85 (± 0,90)	72,03 (± 0,29)	↑
spécialisation	88,52 (± 0,76)	82,34 (± 0,86)	85,25 (± 0,34)	↑	63,66 (± 1,12)	49,62 (± 1,60)	55,68 (± 0,72)	↓	76,64 (± 0,84)	64,92 (± 1,00)	70,17 (± 0,38)	↓
pré-entraînement	88,93 (± 0,30)	82,60 (± 0,79)	85,60 (± 0,41)	↑	68,63 (± 2,09)	56,70 (± 2,77)	61,99 (± 1,11)	↑	79,09 (± 1,43)	69,27 (± 1,53)	73,79 (± 0,53)	↑
optimal	88,65 (± 1,22)	82,91 (± 0,69)	85,62 (± 0,61)		67,72 (± 1,34)	57,51 (± 1,61)	62,16 (± 0,78)		78,36 (± 1,04)	69,86 (± 1,10)	73,82 (± 0,47)	

TABLE 3: Résultat des expériences sur les mentions *gold* du corpus de test

La stratégie de filtrage n'apporte pas d'amélioration nette : nous observons une légère baisse de performance pour le modèle *personnes* tandis que la performance du modèle *événements* reste stable. La spécialisation du classifieur de paires de mentions permet d'améliorer la performance du modèle *personnes* mais diminue fortement la performance du modèle *événements*. Ce résultat négatif pourrait s'expliquer par le volume de données d'entraînement disponible. Le corpus de référence dans le domaine général comprend plus de 2 000 documents (Pradhan *et al.*, 2011) alors que dans notre cas, nous en avons seulement 200. Enfin, le pré-entraînement permet d'améliorer fortement la performance globale de notre modèle, corroborant ainsi les résultats obtenus dans d'autres travaux (Clark & Manning, 2016; Wiseman *et al.*, 2016). L'amélioration est modeste pour le modèle *personnes*. La prévalence des singletons dans des mentions relatives aux personnes (10 %) pourrait limiter l'effet bénéfique du pré-entraînement.

La dernière ligne du tableau 3 donne la performance optimale obtenue en sélectionnant la meilleure configuration pour chaque type d'entités : attention, pré-entraînement dans les deux cas ; spécialisation pour les personnes et filtrage pour les événements. La performance globale de 73,82 en f1-mesure CoNLL est à comparer aux performances des systèmes ayant participé à la tâche 1C de la campagne d'évaluation i2b2 de 2011 pour lesquels nous rapportons les performances dans le tableau 4 (recalculées dans les conditions où nous nous plaçons). Notre système se placerait ainsi entre celui de Cai *et al.* (2011) (classé $5^{ème}$ lors de la campagne d'évaluation) et celui de (Jindal & Roth, 2013) (classé $9^{ème}$ lors de la campagne d'évaluation). Il faut noter que le calcul des performances *via* le script CoNLL a un effet sur le classement initial des systèmes comme on peut le voir dans le tableau 4.

	# i2b2	P	R	F1
Xu *et al.* (2011)	1	82,38	78,13	80,20
Cai *et al.* (2011)	5	75,27	73,96	74,60
notre modèle		78,36	69,86	73,82
Jindal & Roth (2013)	9	65,53	83,48	73,41
Dai *et al.* (2011)	8	76,08	65,65	70,48
Anick *et al.* (2011)	7	79,61	61,67	69,41

TABLE 4: Comparaison de notre système à ceux de la campagne i2b2. Les scores sont obtenus *via* le script officiel CoNLL et calculés en excluant les singletons. Nous rapportons les scores des systèmes pour lesquels une conversion entièrement automatique du format i2b2 vers CoNLL a été possible.

Les modèles développés par les participants lors de la campagne d'évaluation sont fondés sur des traits linguistiques choisis manuellement et intégrés dans des approches à base de règles, d'apprentissage automatique ou une combinaison des deux. Contrairement à ces approches, notre modèle est entièrement neuronal et ne repose pas sur des traits choisis manuellement. Parmi les pistes d'amélioration envisagées figure l'utilisation de ressources externes. Cette possibilité est explorée dans le travail de Zhang *et al.* (2019). Leurs résultats sont néanmoins difficilement comparables aux nôtres dans la mesure où Zhang *et al.* (2019) se focalisent sur les pronoms de façon exclusive et ne traitent donc qu'une partie de la tâche. Par ailleurs, l'utilisation de modèles de langue fondés sur la notion de transformer (Vaswani *et al.*, 2017), tels que le modèle BERT (Devlin *et al.*, 2019), pourrait améliorer les performances de notre modèle, à l'instar du domaine général (Joshi *et al.*, 2019).

6 Conclusion

Nous présentons un modèle neuronal pour la résolution de la coréférence dans le domaine médical, appliqué sur un corpus clinique en anglais. Nous montrons que ce type d'approche permet d'obtenir des performances intéressantes, mais qui restent toutefois inférieures à l'état de l'art à l'aide de modèles fondés sur des traits catégoriels. Dans d'autres expériences, nous avons exploré une situation réelle incluant la détection de mentions en amont de la coréférence (Tourille, 2018). La suite de ce travail pourra explorer plusieurs pistes. Nos expériences font l'hypothèse de la détection parfaite des mentions. Dans une situation d'application réelle, il est nécessaire de procéder à l'extraction des mentions. De plus, les aspects temporels des événements ne sont actuellement pas pris en compte dans notre approche. L'utilisation d'informations temporelles pertinentes permettrait d'apporter des éléments utiles à notre modèle en filtrant les mentions temporellement incompatibles. Le code développé pour convertir le corpus i2b2 du format original vers le format CoNLL est disponible à cette adresse : `https://github.com/jtourille/i2b2-coref-task1c-converter`.

Remerciements

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche sous la référence CABeRneT ANR-13-JS02-0009-01 et d'une aide du labex DigiCosme sous la référence CÔT.

Références

ANICK P., HONG P., XUE N. & AL. (2011). Coreference resolution for electronic medical records. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data* : i2b2.

CAI J., MUJDRICZA E., HOU Y. & AL. (2011). Weakly supervised graph-based coreferencere solution for clinical texts. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data* : i2b2.

CAI J. & STRUBE M. (2010). Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the SIGDIAL 2010 Conference*, p. 28–36, Tokyo, Japan : Association for Computational Linguistics.

CHOWDHURY M. F. M. & ZWEIGENBAUM P. (2013). A controlled greedy supervised approach for co-reference resolution on clinical text. *Journal of Biomedical Informatics*, **46**, 506–515.

CLARK K. & MANNING C. D. (2015). Entity-Centric Coreference Resolution with Model Stacking. In *Proceedings of the Joint Conference of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, p. 1405–1415 : Association for Computational Linguistics.

CLARK K. & MANNING C. D. (2016). Improving Coreference Resolution by Learning Entity-Level Distributed Representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, p. 643–653 : Association for Computational Linguistics.

DAI H. J., WU C. Y., CHEN C. Y. & AL. (2011). Co-reference resolution of the medical concepts inthe patient discharge summaries. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data* : i2b2.

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. 660–669 : Association for Computational Linguistics.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (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*, p. 4171–4186 : Association for Computational Linguistics.

DURRETT G. & KLEIN D. (2013). Easy Victories and Uphill Battles in Coreference Resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 1971–1982 : Association for Computational Linguistics.

GERS F. A., SCHRAUDOLPH N. N. & SCHMIDHUBER J. (2002). Learning Precise Timing with LSTM Recurrent Networks. *Journal of Machine Learning Research*, **3**, 115–143.

GROUIN C., DINARELLI M., ROSSET S., WISNIEWSKI G. & ZWEIGENBAUM P. (2011). Coreference Resolution in Clinical Reports. The LIMSI Participation in the i2b2/VA 2011 Challenge. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data*.

HINOTE D., RAMIREZ C. & CHEN P. (2011). A Comparative Study of Coreference Resolution in Clinical Text. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data*.

HIRSCHMAN L. & CHINCHOR N. A. (1998). MUC-7 Coreference Task Definition. In *Proceedings of the 7th Message Understanding Conference* : Morgan Kaufmann.

HOCHREITER S. & SCHMIDHUBER J. (1997). Long Short-Term Memory. *Neural Computation*, **9**(8), 1735–1780.

JINDAL P. & ROTH D. (2013). Using domain knowledge and domain-inspireddiscourse model for coreference resolution forclinical narratives. *Journal of the American Medical Informatics Association*, **20**, 356–362.

JOHNSON A. E., POLLARD T. J., SHEN L., LI-WEI H. L., FENG M., GHASSEMI M., MOODY B., SZOLOVITS P., CELI L. A. & MARK R. G. (2016). Mimic-iii, a freely accessible critical care database. *Scientific data*, **3**, 160035.

JOSHI M., LEVY O., WELD D. S. & ZETTLEMOYER L. (2019). Bert for coreference resolution : Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, p. 5803–5808 : Association for Computational Linguistics.

LAMPLE G., BALLESTEROS M., SUBRAMANIAN S., KAWAKAMI K. & DYER C. (2016). Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 260–270 : Association for Computational Linguistics.

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 : Association for Computational Linguistics.

LIN C., DLIGACH D., MILLER T. A., BETHARD S. & SAVOVA G. K. (2016). Multilayered Temporal Modeling for the Clinical Domain. *Journal of the American Medical Informatics Association*, **23**(2), 387–395.

PASZKE A., GROSS S., CHINTALA S., CHANAN G., YANG E., DEVITO Z., LIN Z., DESMAISON A., ANTIGA L. & LERER A. (2017). Automatic Differentiation in PyTorch. In *Proceedings of the NIPS 2017 Autodiff Workshop*.

PRADHAN S., LUO X., RECASENS 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, Baltimore, Maryland : Association for Computational Linguistics.

PRADHAN S., RAMSHAW L., MARCUS M., PALMER M., WEISCHEDEL R. & XUE N. (2011). CoNLL-2011 Shared Task : Modeling Unrestricted Coreference in OntoNotes. In *Proceedings of the 15th Conference on Computational Natural Language Learning*, p. 1–27 : Association for Computational Linguistics.

REIMERS N. & GUREVYCH I. (2017). Reporting Score Distributions Makes a Difference : Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 338–348 : Association for Computational Linguistics.

STYLER IV W. F., BETHARD S., FINAN S., PALMER M., PRADHAN S., DE GROEN P. C., ERICKSON B., MILLER T., LIN C., SAVOVA G. & PUSTEJOVSKY J. (2014). Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, **2**, 143–154.

SUN W., RUMSHISKY A. & UZUNER O. (2013). Temporal Reasoning over Clinical Text : The State of the Art. *Journal of the American Medical Informatics Association*, **20**, 814–819.

SUNDHEIM B. M. (1995). Overview of Results of the MUC-6 Evaluation. In *Proceedings of the 6th Message Understanding Conference*, volume 423–442 : Morgan Kaufmann.

TOURILLE J. (2018). *Extracting Clinical Event Timelines : Temporal Information Extraction and Coreference Resolution in Electronic Health Records*. Thèse de doctorat, Université Paris-Saclay.

UZUNER Ö., BODNARI A., SHEN S., FORBUSH T., PESTIAN J. & SOUTH B. R. (2012). Evaluating the State of the Art in Coreference Resolution for Electronic Medical Records. *Journal of the American Medical Informatics Association*, **19**(5), 786–791.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30*, p. 5998–6008. Curran Associates, Inc.

WISEMAN S., RUSH A. M., SHIEBER S. & WESTON J. (2015). Learning Anaphoricity and Antecedent Ranking Features for Coreference Resolution. In *Proceedings of the Joint Conference of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, p. 1416–1426 : Association for Computational Linguistics.

WISEMAN S., RUSH A. M. & SHIEBER S. M. (2016). Learning Global Features for Coreference Resolution. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 994–1004 : Association for Computational Linguistics.

XU Y., LIU J., WU J. & AL. (2011). EHUATUO : a mention-pair coreference system by exploiting document intrinsic latent structures and world knowledge in discharge summaries : 2011 i2b2 challenge. In *Proceedings of the 2011 i2b2/VA/Cincinnati Workshop on Challenges in Natural Language Processing for Clinical Data* : i2b2.

ZHANG H., SONG Y., SONG Y. & YU D. (2019). Knowledge-aware Pronoun Coreference Resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, p. 867–876 : Association for Computational Linguistics.