Influence des annotations sémantiques sur un système de détection de coréférence à base de perceptron multi-couches

Eric Charton Michel Gagnon Ludovic Jean-Louis
École Polytechnique de Montréal, Montréal, QC, Canada
{eric.charton, michel.gagnon,ludovic.jean-louis}@polymtl.ca

RÉSUMÉ

La série de campagnes d'évaluation CoNLL-2011/2012 a permis de comparer diverses propositions d'architectures de systèmes de détection de co-références. Cet article décrit le système de résolution de coréférence Poly-co développé dans le cadre de la campagne d'évaluation CoNLL-2011 et évalue son potentiel d'amélioration en introduisant des propriétés sémantiques dans son modèle de détection. Notre système s'appuie sur un classifieur perceptron multi-couches. Nous décrivons les heuristiques utilisées pour la sélection des paires de mentions candidates, ainsi que l'approche de sélection des traits caractéristiques que nous avons utilisée lors de la campagne CoNLL-2011. Nous introduisons ensuite un trait sémantique complémentaire et évaluons son influence sur les performances du système.

ABSTRACT

Semantic annotation influence on coreference detection using perceptron approach

The ConLL-2011/2012 evaluation campaign was dedicated to coreference detection systems. This paper presents the coreference resolution system Poly-co submitted to the closed track of the CoNLL-2011 Shared Task and evaluate is potential of evolution when it includes a semantic feature. Our system integrates a multilayer perceptron classifier in a pipeline approach. We describe the heuristic used to select the candidate coreference pairs that are fed to the network for training, and our feature selection method. We introduce a complementary semantic feature and evaluate the performances improvement.

MOTS-CLÉS : Coréférence, Perceptron multi-couches.

KEYWORDS: Coreference, Multilayer perceptron.

1 Introduction

La résolution de coréférence a pour objet de déterminer si deux séquences textuelles (par exemple une entité nommée, un pronom, un syntagme nominal) font référence à une même entité sémantique (par exemple une personne ou un évènement). Le principe de résolution consiste à détecter au sein d'un texte des séquences intitulées *mentions coréférentes* et à les regrouper au sein de *chaînes de coréférences*. Cette tâche du TAL fait l'objet d'un ensemble de propositions algorithmiques récemment revisitées par deux campagnes d'évaluation CoNLL Shared Tasks proposées en 2011 et 2012. Ces campagnes ont démontré la prédominance des systèmes de résolution de co-référence par apprentissage automatique appliqués sur des paires candidates. Le système présenté dans cet article est une évolution de celui que nous avons

présenté dans le cadre de notre participation à l'édition 2011 de cette campagne (Pradhan *et al.*, 2011). Notre approche tente de définir un vecteur de traits d'apprentissage original reposant sur des informations issues d'un processus d'extraction d'information et d'analyse linguistique. Dans cette communication, nous complétons ces travaux antérieurs en intégrant un trait sémantique dans le vecteur d'apprentissage.

Cet article est organisé comme suit. Nous commentons l'état de l'art établi par les campagnes CoNLL en section 2. Puis nous présentons notre système de détection de coréférences en section 3. Nous décrivons comment nous proposons d'enrichir son vecteur en lui adjoignant un trait de nature sémantique, c'est à dire définissant précisément l'identité de certaines des mentions candidates utilisées dans le processus de classification par paires. Cette amélioration induit une progression intéressante du système tel qu'évalué lors de la campagne CoNLL. Nous commentons les résultats de ce système modifié en section 4.1 puis nous concluons.

2 Propositions existantes

De nombreux systèmes fondés sur l'apprentissage automatique ont été proposés pour traiter la résolution de coréférences. Les approches les plus récentes à base de réseaux logiques de Markov (MLNs) (Poon et Domingos, 2008), ou fondées sur une approche de partitionnement de graphe (Sapena *et al.*, 2010) sont prometteuses et demeurent peu explorées. Le modèle de classification proposé par Soon (Soon *et al.*, 2001) est prédominant et très largement implémenté. Dans cette approche, les mentions coréférentes potentielles, contenues dans un document d'entraînement, sont localisées via différents modules dits de *détection de mentions*. Les exemples d'entrainements sont ensuite générés sous la forme de vecteurs de traits qui représentent une paire de mentions potentiellement coréférentes.

En mode applicatif toutes les paires de mentions potentiellement coréférentes d'un document sont soumises sous forme d'un vecteur au classifieur, qui valide ou non leur relation en donnant une réponse binaire ou probabilisée. Un processus d'assemblage, postérieur à la classification, regroupe ensuite au sein de chaînes toutes les mentions coréférentes. L'atout principal de la méthode de Soon est sa grande flexibilité : la réduction du problème de construction de chaînes de coréférences à la reconnaissance préalable de paires coréférentes laisse une grande latitude de conception de système. Cette approche rend aussi la méthode de Soon compatible avec des familles de classifieurs très variées : (Versley *et al.*, 2008) a montré qu'un modèle de type SVM permet d'obtenir un système efficace et lors de la campagne CoNLL 2012, (Fernandes *et al.*, 2012) a montré le potentiel d'un perceptron multicouche pour cette tâche.

Le contenu du vecteur de trait utilisé dans l'architecture de Soon offre également un champ de recherche fertile : on a pu ainsi voir dans la proposition de (Stamborg et Medved, 2012) que des dépendances syntaxiques utilisées en tant que traits pouvaient offrir un bon niveau de performance. Certains travaux soulignent la souplesse de l'approche de Soon en ne retenant que le principe de ses paires et vecteurs de traits qu'ils associent non plus à des classifieurs, mais à des méthodes heuristiques. C'est le cas de la proposition de (Lee *et al.*, 2011) qui a obtenu les meilleures performances lors de la campagne CoNLL 2011. Le principe est de remplacer l'apprentissage automatique et la classification par une approche incrémentale à base de règles pré-établies dites *tamis*. Au cours de 13 étapes successives, ces *tamis* trient les différentes paires de coréférences candidates et les assemblent au sein de chaînes. On notera que (Huang *et al.*,

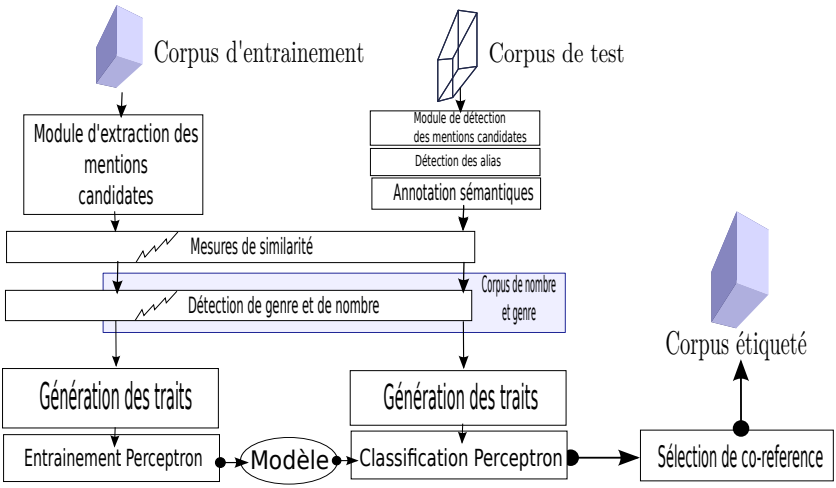

FIGURE 1 – Architecture du système Poly-co.

2009) propose aussi de ne conserver que les paires de vecteurs de traits de l'architecture de Soon, mais utilise un modèle MLN pour assembler les chaînes.

3 Système proposé

La qualité des détecteurs de mentions potentielles jouant un rôle essentiel dans le processus de détection de coréférence (Lee *et al.*, 2011), des efforts d'ingénierie importants sont nécessaires pour élaborer les composants d'un système complet. Notre système n'échappe pas a cette contrainte et une part importante de son implémentation concerne la détection des éléments textuels utilisés pour produire les vecteurs de traits. Nous avons choisi ici de conserver l'architecture de (Soon *et al.*, 2001), alimentée par des vecteurs contenant de nombreux traits de degrés supérieurs. Le corpus Ontonotes (Pradhan *et al.*, 2007) proposé pour entrainer et évaluer les systèmes de détection de coréférences contenant déjà de nombreuses informations telles que la relation syntaxique, la nature syntagmatique, les entités nommées (voir figure 2), nos efforts se sont concentrés sur l'ajout de propriétés évoluées (par exemple les similarités lexicales entre mentions ou les genres des mentions). L'architecture globale présentée dans la figure 1 contient deux parties, la première est dédiée à l'entrainement du système, la seconde à la résolution de coréférence avec un système entrainé.

3.1 Modules de détection et de construction des traits

Les traits des vecteurs de notre système reprennent directement depuis le corpus Ontonotes les catégories morpho-syntaxiques, les syntagmes nominaux et les types d'entités nommées. Nous complétons ces traits en utilisant des modules supplémentaires pour la détection des genres et des nombres, évaluons la détection des alias entre mentions, les similarités entre mentions et introduisons une annotation sémantique. Cinq modules de préparation de vecteurs d'apprentissage sont intégrés à notre système :

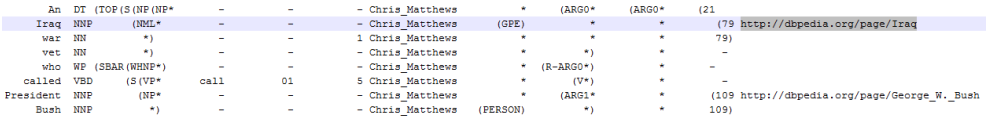

FIGURE 2 – Exemple de corpus Ontonotes avec en dernière colonne l'annotation sémantique.

1. **Module de détection des mentions candidates**, fondé sur des règles d'extraction utilisant les annotations issues de Ontonotes. Il exploite ces annotations pour remplir certains traits (notamment syntaxiques).
2. **Module de détection des alias** entre entités nommées, qui fait intervenir une version précédente du système Poly-co présentée dans (Charton *et al.*, 2010). L'objectif de ce module est d'identifier les différentes variations lexicales d'une même entité en comparant des formes de surface.
3. **Module de calcul de similarité**, qui sert à mesurer la similarité de deux mentions en comparant les chaînes de caractères qui leur sont associées.
4. **Module de détection en genre et en nombre**, détermine le genre et le nombre pour toutes les mentions candidates à l'aide de la ressource fournie par (Bergsma, 2005).
5. **Module de détection sémantique**, détermine par un identifiant unique l'identité de l'objet annoté. Nous évaluons l'influence de ce paramètre dans cette communication.

Lors de la phase d'entrainement, les modules **de détection des mentions candidates** et **de détection des alias** sont remplacés par un seul **module d'extraction des mentions candidates** qui s'appuie directement sur les mentions coréférentes déjà annotées dans le corpus d'entrainement. On obtient ainsi pour entrainer le classifieur un ensemble de paires de mentions candidates positives dont on est certain de la qualité et que l'on complète par un ensemble de paires négatives sélectionnées aléatoirement (cet aspect est détaillé en section 3.3). On se reportera à (Charton et Gagnon, 2011) pour une définition plus précise des modules 1 à 4. Nous décrivons ci-dessous le paramètre sémantique que nous introduisons dans le système Poly-co.

3.1.1 Module de détection sémantique

Nous ajoutons au système Poly-co un trait dit sémantique. Ce trait consiste en une annotation composée d'une URI vers DBPedia. Ce trait vient en complément des annotations fournies sur le corpus Ontonotes [1], tel que présenté dans la figure 2. Le protocole utilisé pour attribuer ces annotations consiste, pour chaque entité nommée candidate, à rechercher son lien correspondant en utilisant un annotateur sémantique [2]. Les corpus d'apprentissage et de test sont traités avec cette méthode. Une correction des erreurs après étiquetage est réalisée visuellement sur le seul corpus d'apprentissage pour limiter l'influence des erreurs d'annotation sur le processus d'entrainement.

Ce lien unique attribué aux entités nommées (GPE, ORG, PERS,LOC, PROD) définit précisément leur identité. Pour l'introduire dans le vecteur de trait sous forme de valeur numérique, nous

1. conll.cemantix.org/2012/data.html
2. Nous utilisons pour cette communication www.wikimeta.org

Nom	Type–valeur	Valeur de trait prise
Propriétés de (A,B)		
IsAlias	vrai/faux	1/0
IsSimilar	réel	0,00 /1,00
Distance	entier	0/d
Sent	entier	0/x
Référence A		
IsNE	vrai/faux	1/0
IsPRP	vrai/faux	1/0
IsNP	vrai/faux	1/0
NE_Semantic type	null / EN	0 / 1-18
PRP_name	null / PRP	0 / 1-30
NP_det	null / DT	0 / 1-15
NP_Type	null / TYPE	0 / 1-3
Gender	M/F/N/U	1/2/3/0
Number	S/P/U	1/2/0
Sémantique	0/URI	0 - 1 à n
Référence B Identique à la référence A		

Table 1 – Paramètres des vecteurs d'apprentissage. Les propriétés communes aux mentions A et B sont détaillées dans la section *Propriétés de (A,B)*. Les traits de la mention A sont détaillés dans la section *Référence A*. Les traits de la mention B sont identiques à ceux de la mention A.

établissons un index de tous les liens sémantiques contenus dans le document dans lequel nous cherchons les chaînes de coréférences et lui attribuons un numéro d'ordre (dans l'exemple de la figure 2, par exemple, le numéro 1 est attribué à *Iraq* et 2 à *Georges Bush*. La valeur 0 est attribuée en l'absence de liens.

3.2 Construction des vecteurs de traits

Le vecteur d'entrainement du système Poly-co (voir tableau 1) est constitué de 24 traits qui décrivent, conformément à l'architecture de Soon, une paire de mentions, (A,B), dans laquelle B est l'antécédent potentiel et A est l'anaphore. Les paramètres sont extraits en utilisant les différents modules de détection. Le rôle du classifieur est ici de fournir une réponse binaire ou probabilisée : A et B co-réfèrent ou non. Quatre paramètres définissent la paire (A,B) (section *Propriétés de (A,B)* du tableau 1) :

- **IsAlias** : il s'agit d'une variable binaire retournée par le **module alias**. La variable prend la valeur *vrai* lorsque A et B sont identifiés comme décrivant la même entité.
- **IsSimilar** : il s'agit du score de similarité calculée par le **module de calcul de similarité**.
- **Distance** : cette valeur représente la distance, c'est-à-dire la différence entre les deux rangs occupées par A et B dans la *liste des mentions candidates*.
- **Sent** : indique le nombre de marqueurs de fin de phrases (ex : « . ! ? ») qui séparent les mentions A et B.

Pour chacun des candidats A et B, un ensemble de neuf traits est ajouté au vecteur. Dans un premier temps, trois variables binaires déterminent si la mention est une entité nommée (**IsNE**), s'il s'agit d'un pronom personnel (**IsPRP**) ou d'un syntagme nominal (**IsNP**). Ensuite, les variables ci-dessous définissent les caractéristiques d'une mention :

– NE_SEMANTIC TYPE est un des 18 types d'entité nommée prédéfini (PERSON, ORG, TIME, etc).
– PRP_NAME s'applique aux pronoms et correspond à une valeur numérique attribuée à chacun des 30 pronoms prédéterminés (ex. : *my, she, it, etc*).
– NP_DET est une valeur qui indique quel déterminant accompagne un syntagme nominal (par exemple, *the*, *this*, *these*, etc).
– NP_TYPE précise si un syntagme nominal est démonstratif, définitif ou quantificateur.
– GENDER et NUMBER indiquent, lorsque les valeurs sont connues, le genre de la mention parmi ***Masculin, Féminin*** *ou* ***Neutre*** et son nombre (***S****ingulier* or ***P****luriel*). Lorsque les valeurs sont inconnues les variables prennent la valeur ***U***.
– SÉMANTIQUE : la valeur du trait est définie selon les modalités présentées en section 3.1.1.

Une valeur *null* (ou 0) est utilisée lorsqu'il n'est pas nécessaire de définir une variable : par exemple, la variable PRP_NAME est positionnée sur 0 lorsque la mention est une entité nommée.

3.3 Entraînement et application du classifieur

Pour entrainer le classifieur, nous utilisons l'algorithme suivant pour préparer les paires. Supposons que la *liste des mentions candidates* contient k mentions $M_1, M_2, \ldots, M_k$, apparaissant dans cet ordre dans le document. L'algorithme commence par la dernière mention du document, c'est-à-dire M_k. Il compare de façon séquentielle M_k avec les mentions précédentes en remontant la liste et s'arrête lorsque (i) une mention en situation de coréférence M_c est trouvée (ii) il a traité un nombre maximum de n mentions (ici n est fixé à 10). Lorsqu'une mention coréférente M_c a été détectée, un vecteur est construit pour toutes les paires de mentions $\langle M_k, M_i \rangle$ où M_i est une mention qui a été traitée. Ces vecteurs sont ajoutés à l'ensemble d'entraînement : M_c est considéré comme exemple positif et tous les autres sont considérés comme négatifs. Le processus est répété avec M_{k-1}, et ainsi de suite, jusqu'à ce que chaque mention soit traitée. Si aucune des n mentions précédentes n'a de lien de coréférence avec M_k, l'ensemble des n paires est écarté et n'est pas utilisé pour les données d'entraînement.

Pour l'application, le processus de détection de coréférence s'appuie sur un algorithme similaire. La mention M_k est comparée aux n mentions précédentes jusqu'à ce que l'on en trouve une pour laquelle le modèle perceptron multi-couches retourne une probabilité supérieure au seuil de 0,5 (ou une valeur binaire dans le cas du classifieur SVM). Si aucun référent n'est trouvé dans la limite des n mentions, M_k est considérée comme une mention non coréférente. Une fois cette procédure appliquée à toutes les mentions d'un document, les coréférences détectées sont utilisées pour construire les chaînes de coréférences.

4 Expériences

Le système complet d'annotation de coréférences Poly-Co [3] est entrainé sur le corpus d'entrainement Ontonotes [4] sur lequel les annotations sémantiques complémentaires ont été apposées. Il est ensuite testé sur le corpus de développement *gold dev-set*. Le tableau 2 présente les résultats obtenus lors de ConLL 2011, sans que le classifieur n'exploite les traits sémantiques, le tableau 3 présente les résultats en intégrant les traits sémantiques. Notre système est entrainé avec

3. Poly-co est téléchargeable sur `https://code.google.com/p/polyco-2/`
4. Le corpus Ontonotes est diffusé par LDC. Un échantillon est téléchargeable sur le site de la conférence ConNLL `http://conll.cemantix.org/2012/data.html`

Scores Poly-co	Mentions			B3			CEAF			MUC		
	R	P	F	R	P	F	R	P	F	R	P	F
Perceptron multi-couches (MLP)	65,91	64,84	65,37	66,61	62,09	64,27	50,18	50,18	50,18	54,47	50,86	**52,60**
SVM	65,06	66,11	**65,58**	65,28	57,68	61,24	46,31	46,31	46,31	53,30	50,00	51,60
Arbres de décision (J48)	66,06	64,57	65,31	66,53	62,27	**64,33**	50,59	50,59	**50,59**	54,24	50,60	52,36

TABLE 2 – Résultats du système, obtenus en appliquant différents classifieurs utilisant les mêmes vecteurs de paramètres sur les données « gold dev-set » du corpus Ontonotes.

Scores Poly-co	Mentions			B3			CEAF			MUC		
	R	P	F	R	P	F	R	P	F	R	P	F
Perceptron multi-couches (MLP)	66,50	65,81	**66,15**	66,70	62,18	64,36	52,31	52,31	**52,31**	54,97	51,86	**53,36**
SVM	65,46	66,60	66,02	65,37	58,79	61,90	48,03	48,03	48,03	54,35	51,00	52,61
Arbres de décision (J48)	66,56	64,97	65,75	67,01	62,5	**64,67**	52,19	52,19	52,19	54,64	51,30	52,91

TABLE 3 – Résultats du système avec les traits sémantiques, obtenus en appliquant différents classifieurs sur les données « gold dev-set » du corpus Ontonotes.

trois types de classifieurs : perceptron multi-couches (MLP), SVM, arbres de décision (J48). Les métriques d'évaluation retenues sont celles adoptées par la campagne ConLL 2011-12, à savoir une mesure de la capacité des systèmes à détecter des mentions d'une part (une simple F-Mesure est retenue), et une moyenne non pondérée des métriques B3, CEAF, et MUC.

4.1 Résultats

Pour la phase d'évaluation de la campagne CoNLL ST 2011, nous avons retenu le modèle MLP qui obtient les meilleures performances sur l'ensemble de données sans annotation sémantique. En raison des faibles différences entre les modèles MLP et J48 il était difficile de définir clairement lequel était le plus adapté avec le modèle de classification retenu. L'introduction de traits sémantiques améliore les performances du modèle Perceptron en regard des deux autres modèles de classification. On observe que l'utilisation d'un identifiant sémantique pour les entités nommées permet d'améliorer d'un point les capacités de détection de mentions du système : ceci s'explique par le fait que l'introduction de cet identifiant améliore la robustesse de classification lorsque les paires sont constituées d'entités nommées. Il en résulte moins de paires mal sélectionnées et donc une augmentation du nombre de mentions correctement détectées. De manière globale, l'introduction de traits sémantiques améliore les performances du classifieur.

5 Conclusions

Cet article présente Poly-co, un système de résolution de coréférence pour l'anglais, facile à adapter à d'autres langues. La version initiale de Poly-co a été construite dans le cadre de la campagne d'évaluation CoNLL ST 2011. Le corpus d'évaluation proposé, Ontonotes, d'un haut niveau de complexité, nous a donné l'opportunité d'évaluer nos algorithmes de détection de mentions dans le cadre d'une tâche complète, regroupant des coréférences entre des entités nommées, des syntagmes nominaux et des pronoms. En introduisant de nouveaux traits sémantiques dans les vecteurs d'apprentissage, nous observons un gain global de performance et soulignons que notre approche à base perceptron multi-couches est une solution intéressante pour la reconnaissance de chaînes de coréférence.

Références

BERGSMA, S. (2005). Automatic acquisition of gender information for anaphora resolution. *Advances in Artificial Intelligence*, pages 342–353.

CHARTON, E. et GAGNON, M. (2011). Poly-co : a multilayer perceptron approach for coreference detection. *In CoNLL : Shared Task*.

CHARTON, E., GAGNON, M. et OZELL, B. (2010). Poly-co : an unsupervised co-reference detection system. *In* BELZ, A. et KOW, E., éditeurs : *INLG 2010-GREC*, Dublin. ACL SIGGEN.

FERNANDES, E., dos SANTOS, C. et MILIDIÚ, R. (2012). Latent structure perceptron with feature induction for unrestricted coreference resolution. *Proceedings of the Joint Conference on EMNLP and CoNLL : Shared Task*, pages 41–48.

HUANG, S., ZHANG, Y., ZHOU, J. et CHEN, J. (2009). Coreference Resolution using Markov Logic Network. *In The 10th International Conference on Intelligent Text Processing and Computational Linguistics*, volume 41, pages 157–168.

LEE, H., PEIRSMAN, Y., CHANG, A., CHAMBERS, N., SURDEANU, M. et JURAFSKY, D. (2011). Stanford's Multi-Pass Sieve Coreference Resolution System at the CoNLL-2011 Shared Task. *In CoNLL Shared Task*, numéro June, page 73.

POON, H. et DOMINGOS, P. (2008). Joint unsupervised coreference resolution with Markov logic. *In Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP '08*, page 650, Morristown, NJ, USA. Association for Computational Linguistics.

PRADHAN, S., RAMSHAW., L., MARCUS, M., PALMER, M., WEISCHEDEL, R. et NIANWEN, X. (2011). CoNLL-2011 Shared Task : Modeling Unrestricted Coreference in OntoNotes. *In Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL 2011)*, Portland, Oregon.

PRADHAN, S., RAMSHAW, L., WEISCHEDEL, R., MACBRIDE, J. et MICCIULLA, L. (2007). Unrestricted coreference : Identifying entities and events in OntoNotes. *In International Conference on Semantic Computing, 2007. ICSC 2007.*, pages 446–453. IEEE.

SAPENA, E., PADRÓ, L. et TURMO, J. (2010). RelaxCor : A global relaxation labeling approach to coreference resolution. *In Proceedings of the 5th International Workshop on Semantic Evaluation*, numéro July, pages 88–91. Association for Computational Linguistics.

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

STAMBORG, M. et MEDVED, D. (2012). Using syntactic dependencies to solve coreferences. *Proceedings of the Joint Conference on EMNLP and CoNLL : Shared Task*, pages 64–70.

VERSLEY, Y., PONZETTO, S., POESIO, M., EIDELMAN, V., JERN, A., SMITH, J., YANG, X. et MOSCHITTI, A. (2008). BART : A modular toolkit for coreference resolution. *In Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, numéro 2006, pages 9–12, Marrakech. European Language Resources Association (ELRA).