TALN 2009, Senlis, 24–26 juin 2009

Vers des contraintes plus linguistiques en résolution de coréférences

Étienne Ailloud Manfred Klenner
Institute of Computational Linguistics
Zurich University
Zurich, Switzerland
{ailloud,klenner}@cl.uzh.ch

Résumé. Nous proposons un modèle filtrant de résolution de coréférences basé sur les notions de transitivité et d'exclusivité linguistique. À partir de l'hypothèse générale que les chaînes de coréférence demeurent cohérentes tout au long d'un texte, notre modèle assure le respect de certaines contraintes linguistiques (via des filtres) quant à la coréférence, ce qui améliore la résolution globale. Le filtrage a lieu à différentes étapes de l'approche standard (c-à-d. par apprentissage automatique), y compris avant l'apprentissage et avant la classification, accélérant et améliorant ce processus.

Abstract. We propose a filter model of coreference resolution that is based on the notions of transitivity and linguistic exclusivity. Starting from the general assumption that coreference sets remain coherent throughout a text, our model enforces the checking of some compatibility criteria (filters) between coreference candidates, thereby improving resolution performance. This filtering is achieved at different stages of the workflow of machine-learning-based coreference resolution, including at the standard learning and testing steps, where it may help reduce the computational load and better distribute the actual occurrences to be learned.

Mots-clés : Résolution de coréférences, apprentissage automatique, linguistique informatique par contraintes.

Keywords: Coreference resolution, Machine learning, Constraint-based NLP.

1 Introduction

La résolution de coréférences est une tâche utile dans de nombreuses applications de traitement du langage naturel, par exemple l'extraction d'informations. Pour identifier une entité parmi un ensemble de documents, il est crucial de pouvoir identifier les diverses expressions pouvant y référer, y compris les expressions *anaphoriques* (p. ex.des pronoms). Trouver un antécédent à chacune de celles-ci se nomme la résolution d'anaphores, trouver toutes les occurrences de la même entité est la résolution de coréférences, plus générale. Mais même pour la résolution d'anaphores comme sous-tâche, les systèmes actuels ont encore des performances insatisfaisantes (entre 55% et 70% pour la F-mesure sur données réelles). L'approche standard consiste à entraîner un modèle d'apprentissage automatique à produire des décisions binaires (des réponses à « L'expression a coréfère-t-elle avec l'expression b ? »).

Le problème majeur dans l'approche informatique de la résolution de coréférences est son manque de contraintes : l'éventail des contraintes qui régissent le phénomène de coréférence est plus large que ce qui peut être correctement modélisé à l'heure actuelle. Par exemple, se restreindre aux seules contraintes morpho-syntaxiques jouant sur différentes expressions linguistiques ne suffit pas, dans beaucoup de langues (même celles morphologiquement marquées comme l'allemand), à discriminer suffisamment entre différent candidats potentiellement référents (les markables). Les contraintes nécessaires ici sont issues d'autres couches du traitement du langage : sémantique, pragmatique. En l'absence de ces contraintes de haut niveau, la résolution de coréférences devient un problème combinatoire : l'énumération de toutes les possibilités (c-à-d. toutes les paires de markables) occupe vite trop d'espace en mémoire. Aussi la réduction de l'espace de recherche constitue-t-elle la pierre d'achoppement de la plupart des approches d'apprentissage automatique : pour rester réalisables, celles-ci sont restreintes par des heuristiques qui font que seulement une partie des possibilités est explorée. Par exemple, on se limite généralement à des *paires* de markables (en se fondant sur le postulat usuel que l'observation binaire est suffisante pour capturer le phénomène de coréférence) ; la résolution de coréférences est généralement restreinte en portée, également : en utilisant p. ex. des fenêtres mobiles (la dernière référence à une entité connue étant susceptible d'apparaître dans un contexte récent).

Considérons l'exemple en allemand suivant, dans lequel différents constituants nominaux masculins font référence à deux entités différentes :

> [Der jugoslawische Präsident]$_i$ ist in diesem Sinne kalt, denn [er]$_j$ kennt [[seinen]$_k$ Westen]$_l$. [Er]$_m$ hat [ihn]$_n$ dort studiert, wo [er]$_p$ am westlichsten ist, als Banker in New York.
>
> [The Yugoslavian president]$_i$ is in this sense cold, for [he]$_j$ knows [[his]$_k$ western society]$_l$. [He]$_m$ studied [it]$_n$ where [it]$_p$ is at its most west-like, as a banker in New York.

La traduction anglaise de ces phrases révèle immédiatement les coréférences, à travers l'opposition masculin/neutre. Cependant, la morphologie allemande ne permet pas de distinguer ces syntagmes les uns des autres, tous de même genre, nombre et personne, alors que la lecture préférée donne $i \equiv j \equiv k \equiv m \not\equiv l \equiv n \equiv p$ (ou, comme chaînes de coréférence : $\{\{i, j, k, m\}, \{l, n, p\}\}$).

On peut aussi formuler des restrictions de portée plus générale (non locale), p. ex. la *transitivité*. Celle-ci permet, couplée avec l'exclusivité, d'exclure des paires de coréférence impossibles (en considérant la partition en classes d'équivalence selon la relation $\equiv$) : Sachant que $j \not\equiv l$[1],

1. « A non-reflexive pronoun must not be coindexed with a c-commanding NP within the minimal NP or S that

chercher à lier $[\text{Er}]_m$ dans la phrase précédente oblige à choisir (de manière exclusive) *soit* $m \equiv j$, *soit* $m \equiv l$. Cette observation simple limite le choix de chaînes de coréférence à deux possibilités (pour les trois markables considérés) : $\{\{j, m\}, \{l\}\}$ et $\{\{j\}, \{l, m\}\}$; Pourtant un classifieur automatique, limité dans sa vision binaire, peut éventuellement produire *à la fois* $m \equiv j$ et $m \equiv l$, induisant par transitivité la chaîne de coréférence $\{\{m, j, l\}\}$. Celle-ci est inconsistante en ce qu'elle implique que $j \equiv l$, bien que la paire (j, l) n'ait jamais été considérée elle-même (car elle contredit l'exclusivité du sujet j et de l'objet non-réflexif l). Nous y reviendrons en partie 3.

La transitivité et l'exclusivité sont deux restrictions dures, la première étant même indépendante d'une quelconque théorie syntaxique : c'est plutôt au manque de consistance *logique* qu'elle s'attaque. Ceci n'a pas été réellement abordé dans la littérature (voir cependant (Finkel & Manning, 2008) ou (Denis & Baldridge, 2008a)), alors que nous montrons qu'elle peut améliorer les performances globales de la résolution de coréférences. L'élimination des paires de markables qui contredisent la transitivité, couplée à une notion d'exclusivité adaptée, constitue une sorte de filtre de consistance.

Dans la prochaine partie nous décrirons le contexte général de l'apprentissage automatique où se situe notre modèle. Puis la partie 3 ira plus en détails dans les concepts-clés de transitivité et d'exclusivité, conduisant à la description du modèle lui-même en partie 4. Nous présenterons des résultats expérimentaux à l'appui en partie 5, et des approches comparables en partie 6.

2 Apprentissage automatique de paires de coréférence

En partant d'une liste de *markables*—les objets linguistiques se référant potentiellement aux entités du discours—et des informations s'y rapportant (leurs traits linguistiques, p. ex.), l'objectif est de prédire lesquels sont coréférents parmi des markables non vus. La plupart des approches sont basées sur l'apprentissage automatique de paires de markables. Une seconde phase agrège les markables coréférents en classes d'équivalence (les chaînes de coréférence).

Traits Le classifieur extrait l'information qui lui est utile d'un ensemble de traits, déterminés à partir des markables ou de paires d'iceux en utilisant l'information linguistique accessible ; ils sont relatifs à : la structure superficielle (p. ex. la distance entre deux markables), la morpho-syntaxe (p. ex. leur fonction grammaticale), de la sémantique basique (p. ex. la compatibilité ontologique des markables). D'un point de vue linguistique, certains traits encodent en fait peu d'information, comme la distance entre deux markables ; certains sont plus ou moins une transcription directe de l'information fournie (p. ex. quand les catégories lexicales elles-mêmes sont utilisées comme traits).

Phase d'agrégation Une conséquence importante de la production de *paires* de markables est qu'un post-traitement est nécessaire pour accéder aux chaînes de coréférence. Celles-ci sont obtenues en fusionnant les paires de markables coréférents en classes d'équivalence. Leur importance, par opposition aux seules paires en sortie, se reflète aussi dans le schéma d'évaluation : la mesure standard MUC (Vilain *et al.*, 1995), de même que son extension B^3 (Bagga & Baldwin, 1998), mesure l'harmonie des résultats avec le gold standard plutôt sur le plan des chaînes

contains it. » (cf. (Chierchia & McConnell-Ginet, 1990))

que des paires. Ceci est d'autant plus vrai pour la nouvelle mesure ECM (introduite dans (Luo, 2005)), qui cherche d'abord à aligner les chaînes de coréférence avec le standard et ensuite seulement calcule les résultats.

L'exemple de l'introduction illustre le problème majeur dans le traitement binaire de la coréférence : L'agrégation canonique de markables à partir de paires, par simple fusionnement (c-à-d. quand deux markables appartiennent à la même chaîne ss'ils forment une paire coréférente), induit une inconsistance dans les chaînes de coréférence.[2] Cette inconsistance est inévitable, car le classifieur binaire considère les paires de markables comme des événements indépendants : il ne peut pas apprendre à faire le rapprochement entre deux paires, même si elles ont un élément en commun—il ne peut pas apprendre la transitivité.

D'autres procédés d'agrégation plus sophistiqués produisent de l'inconsistance de la même manière. Ils fonctionnent de gauche à droite de manière incrémentielle, en assignant à chaque « anaphore » un « antécédent » idoine parmi les markables la précédant.[3] L'agrégation *closest-first* consiste à prendre pour cela le premier markable satisfaisant dans le contexte gauche de l'anaphore (voir p. ex. (Soon *et al.*, 2001)). L'agrégation *best-first* constitue tout d'abord une liste de candidats à l'antécédence dans le contexte gauche, puis sélectionne parmi ceux-ci celui qui, avec l'anaphore, maximise une certaine probabilité de coréférence (voir p. ex. (Ng & Cardie, 2002)). Typiquement, un classifieur produit non seulement une décision binaire mais aussi un poids associé à cette décision, ce dernier mesurant ladite probabilité. La construction de chaînes d'entités coréférentes commune à ces deux procédés est incrémentielle et gloutonne ; nous appelons ces approches « naïves ». L'agrégation naïve (comptant aussi la canonique *aggressive-merging*) ne garantit pas la consistance. Une prise en compte globale de la transitivité comme nous le proposons est un moyen d'y remédier.

3 Transitivité et exclusivité

Comme requis pour la consistance (logique), nous considérons la relation de coréférence comme une relation mathématique, dont les chaînes de coréférence sont les classes d'équivalence. L'exemple a montré plus haut que l'apprentissage automatique binaire suivi d'une agrégation naïve sont voués à produire des chaînes inconsistantes. Une solution est d'agréger différemment les paires de markables en chaînes, en autorisant le modèle à réviser ces décisions binaires afin de garantir transitivité et exclusivité. Cette approche est plutôt rare dans la littérature, sans doute à cause de la complexité de modélisation : la transitivité est une contrainte globale, à valider sur la totalité d'un texte/paragraphe/unité de discours. La prendre en compte exhaustivement implique une complexité de calcul élevée.[4]

On peut distinguer deux types d'entorse à la consistance :

1. Auto-contradiction : une règle d'exclusivité est enfreinte. Ceci constitue une forme faible

2. Dans l'exemple et son traitement putatif du texte « *[...] er kennt seinen Westen. Er [...]* » par apprentissage automatique, ceci reviendrait à fusionner les deux décisions binaires $\{[\text{seinen Westen}]_l, [\text{Er}]_m\}$ et $\{[\text{er}]_j, [\text{Er}]_m\}$ en la chaîne de coréférence $\{[\text{er}]_j, [\text{seinen Westen}]_l, [\text{Er}]_m\}$—inconsistante parce que $j \neq l$.

3. Ils sont traditionnellement nommés ainsi par souci de généralité, mais une relation cataphorique, p. ex., sera bien sûr orientée dans l'autre direction. Néanmoins, la phase d'agrégation ne connaît qu'une direction.

4. Étant donnés n markables, il y a $\frac{n(n-1)(n-2)}{2}$ possibilités d'exprimer la transitivité entre trois quelconques d'entre eux, comme dans : « si i et j sont coréférents et j et k aussi, alors i et k aussi ». La contrainte de transitivité est donc cubiquement quantifiée en le nombre de markables, ce qui implique que, pour être exhaustif, la complexité de tout algorithme traitant de manière extensionnelle ces contraintes doit être majorée par un exposant trois.

(et linguistique) d'inconsistance, puisqu'elle pourrait éventuellement être détectée et évitée par un classifieur automatique. Un exemple est de poser $j \equiv l$ dans l'exemple plus haut ;

2. Contradiction indirecte : l'exclusivité et la transitivité se contredisent. Dans ce cas la paire fautive n'est même pas générée (par le classifieur), mais apparaît implicitement par transitivité. C'est cela-même que notre modèle vise principalement à améliorer : l'inconsistance ne peut être levée même avec des filtres durs (cf. partie 4), qui eux aussi sont locaux et confinés à leur perspective binaire.

Dans le cas de deux paires transitivement inconsistantes, la question se pose quant à laquelle instantier positivement. Ceci met au jour la non-localité du processus de décision : pour résoudre exactement ce problème, il faudrait effectuer une *optimisation* globale ; les poids de paires isolées ne suffisent pas pour fournir la solution globalement optimale. En fait, il faudrait pondérer des *partitions* entières de chaînes de coréférence pour y accéder, et optimiser sur toutes les manières de partitionner un ensemble de markables. Ceci se trouve être l'espace de recherche de la résolution de coréférences, l'*arbre de Bell*. Malheureusement sa taille est rédhibitoire, limitant son exploration à des heuristiques (comme dans (Luo *et al.*, 2004), où l'arbre est élagué selon un modèle statistique, qui n'a plus alors le bénéfice de l'optimalité).

Notre modèle constitue de même une solution alternative (gloutonne) à l'optimisation. Il consiste à réorganiser l'espace de recherche (comme heuristique) : en permettant aux décisions les plus fiables d'être évaluées d'abord. Il pallie la non-localité par une approche incrémentielle qui aborde la transitivité de manière *intensionnelle* ; nous le présentons dans la prochaine partie.

4 Notre modèle filtrant

Dans cette partie nous décrivons la réduction des possibilités opérée par notre modèle au moyen de filtres appliqués sur les paires apprises et classifiées et pendant l'agrégation. L'exclusivité est réalisée par ce filtrage sur la génération de vecteurs puis sur l'agrégation, la transitivité par l'incrémentialité de l'agrégation.

Filtrage de la génération de vecteurs La première étape consiste à sélectionner les données à passer par l'apprentissage automatique : Nous utilisons des *filtres durs* pour réduire le nombre de ces instances (représentées par des *vecteurs de traits*)—en apprentissage et en classification. Ils éliminent beaucoup d'improbables instances de coréférence en se basant sur des critères immédiats comme l'accord (morpho-syntaxique). Ils font respecter l'exclusivité linguistique et en tant que tels éliminent le premier type d'exclusivité évoqué en partie 3. Ceux utilisés ici assurent l'exclusivité de deux configurations, entre autres : d'abord la *clause-boundness*, vérifiée par deux markables s'ils apparaissent de manière non-appositive dans la même proposition, aucun n'étant un pronom réflexif ou adjectif possessif (p. ex. dans « $[$*He*$]_m$ *studied* $[$*it*$]_n$ », $[$*He*$]_m$ et $[$*it*$]_n$ sont clause-bound, la paire (m, n) n'est donc même pas générée pour l'apprentissage). Ensuite l'exclusivité de la *NP-boundness* assure que deux markables ne soient pas coréférents s'ils apparaissent dans le même syntagme nominal.

Le filtrage est d'autant plus important pour la phase d'apprentissage que la relation de coréférence est un phénomène dispersé : la grande majorité des paires-instances sont en fait négatives (non-coréférentes), de sorte que le modèle apprend plutôt un certain concept de *non*-coréfé-

rence. Le *downsampling*, la réduction du rapport d'instances négatives aux instances positives, peut lui faire mieux cerner les véritables relations entre markables coréférents. Les filtres s'avèrent également utiles en phase de test : un modèle de classification recentré sur les moins nombreuses instances positives pourra aussi classifier plus vite, d'où un gain de temps.

Apprentissage automatique L'apprentissage est ensuite lancé sur ces vecteurs filtrés, comme en partie 2. Le modèle utilise ici des traits classiquement utilisés dans les modèles actuels (cf. (Soon *et al.*, 2001) pour un exemple très suivi) ; Voici par exemple un vecteur généré pour la paire $([\textit{Der jugoslawische Präsident}]_i, [\textit{ihn}]_n)$, dans lequel apparaissent distances entre markables, étiquettes individuelles, comparaisons entre les markables, l'extension du mot pour les pronoms et mesures de salience individuelles (basées sur la fréquence) :

⟨ 0	6	NN	PPER	SUBJ	OBJA	false	false	noun	ihn	3	4	unknown ⟩
distance en phrases		POS		fonction		matching		chaîne pronom		salience		classes sémantiques compatibles
markables												
i	n	i	n	i	n	strict	fuzzy	i	n	i	n	

Nous utilisons le classifieur à mémoire « lazy » TiMBL (Daelemans *et al.*, 2004) pour l'apprentissage ; il produit pour chaque instance une prédiction booléenne, basée sur les nombres d'instances positives et négatives trouvées *similaires* à l'instance testée (la mesure de similarité étant le produit de l'apprentissage).

Pondération On assigne ensuite aux instances prédites positives un *poids* fonction de ces nombres ; il modélise le coût de considérer une instance un bon candidat à la coréférence :

$$w_{ij} = \begin{cases} 0 & \text{si } n_{ij} = 0 \\ \frac{n_{ij}}{n_{ij}+p_{ij}} & \text{sinon} \end{cases} \quad \text{où} \quad \begin{array}{l} n_{ij} \text{ est le nombre d'instances négatives similaires à } (i,j) \\ p_{ij} \text{ est le nombre d'instances positives similaires à } (i,j) \end{array}$$

Une instance qui ne connaît que des instances similaires positives est un candidat sûr, elle reçoit donc un poids nul. Une qui n'a au contraire que des instances similaires négatives reçoit le poids maximum 1. Les instances prédites positives sont celles qui ont au moins autant d'instances similaires positives que négatives (c-à-d. un poids d'au plus $1/2$).

Une fois les instances pondérées, le coeur de notre algorithme consiste à considérer d'abord les meilleurs candidats, induisant par-là un *ordre* sur les paires *positives* produites par TiMBL, inspiré de l'algorithme de Balas (Balas, 1965) : $\mathcal{O}_{\leq 1/2} = \mathcal{C}_{ij}, \mathcal{C}_{kl}, \ldots$ pour $w_{ij} \leq w_{kl} \leq \ldots \leq \frac{1}{2}$.

Agrégation Cette étape met en oeuvre le principe ébauché en partie 3. Elle va plus loin que l'agrégation canonique (cf. partie 2) en ce qu'elle autorise une décision binaire produite par TiMBL à être révisée : À chaque itération, la compatibilité d'une nouvelle paire (i,j) de $\mathcal{O}_{\leq 1/2}$ est testée par rapport à toutes les chaînes construites jusqu'alors qui contiennent i et j. À savoir : si $\mathcal{S}_i$ et $\mathcal{S}_j$ sont des chaînes contenant i et j, respectivement, *chaque* élément de $\mathcal{S}_i$ est successivement comparé à *chaque* élément de $\mathcal{S}_j$. Si un seul de ces tests échoue, (i,j) n'est plus retenue comme bon candidat—c'est une paire inconsistante avec les chaînes en cours de construction. C'est là que l'ordre de Balas a son importance, lorsque les meilleurs candidats sont testés en premier (d'autres ordres se sont effectivement avérés moins performants, comme montré dans (Klenner & Ailloud, 2008)).

Notre définition de la compatibilité implique le même filtrage que pendant la génération des vecteurs. En particulier, ces filtres instantient différentes couches de critères linguistiques :

– *exclusivité* : Les deux prédicats intraphrasaux vus plus haut jouent ici leur rôle : Généralement, deux markables clause-bound sont exclusifs ; de même, deux markables NP-bound sont exclusifs aussi[5].
– *accord morpho-syntaxique* : Ceci dépend en grande partie de la catégorie lexicale de l'élément anaphorique ; en allemand, par exemple, le pronom relatif s'accorde en genre, nombre et personne avec son antécédent ;
– *accord sémantique* : L'antécédent nominal d'une anaphore nominale doit avoir la même (ou compatible) classe sémantique.

Dans les expériences, pour investiguer sur différents comportements linguistiques, nous appliquons ou neutralisons plusieurs de ces filtres à l'agrégation (cf. partie 5).

Notre approche de l'agrégation représente donc un compromis entre l'approche naïve entièrement gloutonne (rejeter sur la base d'une probabilité plus faible une de deux paires-instances qui transgressent la transitivité) et la stratégie d'exploration exhaustive de l'espace de recherche (l'arbre de Bell, cf. partie 3) lourde en calcul (équivalente à un backtracking total) qui trouve la partition optimale. Notre algorithme impose donc *transitivement* toute contrainte d'exclusivité appelée par le test de compatibilité—ces deux principes sont traités intensionnellement.

Extensions du modèle Des expériences actuelles avec le modèle impliquent des contraintes qui ne sont ni transitives ni exclusives ; nous les appelons *contraintes de liage d'entités*, au sens où elles obligent une entité anaphorique donnée à avoir un antécédent. Par exemple, on pourrait édicter que « tout pronom possessif doit être lié », ou bien qu'« un syntagme nominal démonstratif doit être lié au maximum trois phrases en arrière ». Techniquement, de telles contraintes requièrent des outils algorithmiques plus élaborés qu'un simple test d'exclusivité de paire.

On pourrait aussi songer à des filtres d'exclusivité « techniques », excluant des configurations comme celle entre un pronom réflexif (sans indication de genre ou nombre en allemand) et un markable le suivant, cf. j et k dans « $[Er]_i$ *hat* $[sich]_j$ *amüsiert.* $[Sie]_k$ *nicht.* » (« $[Il]_i$ $[s']_j$*est bien amusé.* $[Elle]_k$ *pas.* ») Ce filtre permettrait d'exclure $j \equiv k$, par laquelle l'erreur pourrait se propager : par transitivité, les hypothèses correctes $i \equiv j$ et $j \equiv k$ (c-à-d. les *paires* sont compatibles) amènent $i \equiv k$, qui constitue une paire exclusive. Bien que j et k puissent être prédits coréférents, ils ne forment pas une relation anaphorique ; l'apprentissage gagnerait donc à se passer de cette paire.

Dans une configuration en apparence semblable : « $[Mr.\ Clinton]_i$ *[...]* $[Clinton]_j$ *[...]* $[She]_k$ », les deux paires (i, j) et (j, k) sont bien compatibles, mais on ne peut pas appliquer un filtre tel que celui plus haut, car $[\text{Clinton}]_j$ pourrait réellement être l'antécédent de l'anaphore $[\text{She}]_k$. Nous appelons ceci un *pont d'inconsistance* via j ; l'exclusivité seule n'y remédie pas, cette fois. Dans une telle configuration, il n'y a pas d'autre solution que de faire jouer la transitivité ; celle-ci reste donc nécessaire, aussi précis que soient les filtres d'exclusivité. Ce dernier exemple montre bien qu'il est important de repasser par les filtres durant l'agrégation, pour éliminer les éventuels ponts d'inconsistance.

5. Il y a quelques exceptions, notamment l'usage attributif (adjectival) de syntagmes verbaux imbriqués, où un pronom réflexif peut être co-indexé avec le syntagme nominal entier : cf. $[das\ [sich]_i\ amüsierende\ Mädchen]_i$ ($[la\ fille\ qui\ [s']_i\ amuse]_i$)—le verbe allemand *sich amüsieren* étant aussi réflexif. Mais comme les markables imbriqués ne sont pas annotés pour la coréférence dans notre gold standard, (cf. partie 5), le filtre plus restrictif est quand même utilisé.

5 Résultats

Dans cette partie nous montrons que la transitivité et l'exclusivité peuvent améliorer la phase d'agrégation. Notre gold standard, le Corpus arboré d'allemand écrit de Tübingen (Telljohann *et al.*, 2005, TüBa), se compose d'articles de journaux en allemand ; annotés en structure syntaxique et en coréférence. Celle-ci couvre plus de 1 100 textes pour plus de 23 000 phrases.

Étalon Nous avons utilisé 80% des textes du corpus pour la phase d'apprentissage et le reste (soit 217 textes) pour la classification. Notre système-étalon consiste à lancer TiMBL dans ces conditions, les vecteurs soumis au classifieur ainsi que le gold standard ayant passé au préalable nos filtres durs. Ceci implique que l'évaluation doit être interprétée modulo les filtres de vecteurs : seules les configurations qu'ils acceptent explicitement se retrouvent dans les paires à classifier et celles du gold standard. Par exemple, les configurations impliquant *es* (pronom personnel neutre 3^{e} pers.), majoritairement explétif, sont ignorées. Les données après filtrage représentent 60 414 (resp. 15 659) paires d'instance pour l'entraînement (resp. la classification). Nos résultats reposent sur les seules *mentions réelles* (les markables réellement coréférents avec un autre) ; ceci permet de concentrer l'étude sur un phénomène particulier ((Luo *et al.*, 2004) p. ex. utilise également cette restriction), comme ici le gain apporté par l'agrégation filtrante.

Paramètre	ÉTALON	MORPH	MORPH+EXCL	**MORPH+EXCL+SÉM**
Précision	73.61	78.13	79.23	82.02
Rappel	63.36	65.76	66.09	68.27
F-mesure	68.10	71.41	72.07	74.52

TAB. 1 – Expérience : contribution des différentes couches filtrantes à l'agrégation consistante.

Paramètres expérimentaux du filtrage Le tableau 1 présente les progrès atteints par rapport à l'étalon : différentes couches filtrantes y sont ajoutées (morpho-syntaxe, exclusivité intraphrasale, sémantique) dans le but de montrer l'effet de la transitivité combinée aux filtres d'exclusivité. L'effet de l'exclusivité sur une base morphologique (règles d'accords) apparaît en troisième colonne, avec déjà un gain de 3,31% en F-mesure (+4,52% en précision et +2,4% en rappel). La colonne suivante présente l'effet combiné avec l'exclusivité « pure » (c-à-d. intraphrasale) en sus, soit une nouvelle augmentation, de 0,72%, de la F-mesure. Globalement, l'agrégation transitive qui propage l'exclusivité apporte pour nos données une amélioration de 6,42%. On peut voir que l'exclusivité et les critères morpho-syntaxiques contribuent grandement à rendre consistantes les chaînes de coréférence implicitement produites par TiMBL. L'efficacité du critère sémantique peut être interprétée par la suppression des ponts d'inconsistance.

Notre évaluation fait usage de la mesure de coréférence ECM (présentée dans (Luo, 2005)), qui met l'accent sur les chaînes de coréférence plutôt que de comparer simplement le nombre de paires en accord avec le gold standard. Ceci peut expliquer l'augmentation du rappel en colonnes 3, 4 et 5, alors que notre algorithme ne peut que *rejeter* une paire incompatible en la révisant. C'est que la mesure ECM cherche à aligner les chaînes de coréférence avec le gold standard : la suppression à raison d'un maillon peut engendrer une chaîne supplémentaire alignée avec succès, ce qui augmente le rappel. La mesure ECM offre plusieurs avantages par rapport au traditionnel MUC, comme décrit p. ex. dans (Luo, 2005).

6 Travaux antérieurs

Il y a eu d'importants efforts de recherche sur l'apprentissage automatique de la résolution de coréférences ; on a récemment commencé à aborder les limites de portée inhérentes aux modèles binaires. Par exemple, le modèle proposé dans (Yang *et al.*, 2004) incorpore des traits étendus, qui prennent non seulement en compte les paires de markables mais aussi les chaînes de coréférence partielles construites durant l'agrégation. Cependant le modèle ne fait pas état de contraintes dures sur ces chaînes et reste dans la lignée standard. Comme mentionné en partie 2, l'arbre de Bell ne peut pas être exploré dans son intégralité. Dans (Luo *et al.*, 2004), un modèle statistique y apporte une solution en élaguant l'arbre selon les scores obtenus sur ses chemins, perdant la garantie d'optimalité. Cette approche reste également purement statistique—elle n'apporte pas de contraintes supplémentaires en dehors du classifieur. Une autre perspective est abordée dans (Denis & Baldridge, 2008b), où des modèles différenciés sont entraînés selon les différentes catégories de markables (susceptibles d'exhiber différents motifs de coréférence).

On a plus récemment insisté sur l'importance de la transitivité en résolution de coréférences. Les modèles dans (Denis & Baldridge, 2008a) et (Finkel & Manning, 2008) font tous deux respecter cette contrainte globale par programmation linéaire entière (ILP), ce faisant assurant une certaine notion d'optimalité aux chaînes produites. Aucun toutefois n'y combine des contraintes d'exclusivité qui pourraient l'utiliser pour se propager (bien que la combinaison avec la reconnaissance d'entités nommées dans le modèle similaire de Denis & Baldridge (2008a) engendre une propagation mutuelle des deux tâches). De plus, l'ILP impose l'extentionnalisation des contraintes, ce qui affecte la clarté, voire la faisabilité dans le cas de longs textes, pour la transitivité.

7 Conclusion et travaux futurs

Nous avons proposé un modèle de résolution de coréférences, bâti autour d'un modèle binaire d'apprentissage automatique et se basant sur des filtres pour imposer transitivité et exclusivité. Nous avons montré que le filtrage transitif pendant la phase d'agrégation fournit de meilleurs résultats de par la consistance des chaînes de coréférence produites, ce qui constitue une amélioration par rapport aux approches binaires classiques.

Dans des travaux ultérieurs nous chercherons à trouver de nouveaux filtres d'exclusivité, afin de coller—à terme—au plus près des données empiriques: ceux-ci devront être d'inspiration linguistique et nécessiteront probablement peu d'adaptation à d'autres langues (de plus, la transitivité est une contrainte universelle). En outre nous voulons continuer à expérimenter sur des contraintes, en particulier celles qui imposent le liage d'entités, bien que celles-ci modifient quelque peu l'architecture du modèle.

Remerciements

La recherche effectuée dans le cadre de ce travail est financée par le projet n° 105211–118108/1 du Fonds National Suisse. Nous tenons à remercier Anne Goehring ainsi que les deux relecteurs anonymes.

Références

BAGGA A. & BALDWIN B. (1998). Algorithms for scoring coreference chains. In *Proceedings of the Linguistic Coreference Workshop at The First International Conference on Language Resources and Evaluation (LREC'98)*, p. 563–566.

BALAS E. (1965). An additive algorithm for solving linear programs with zero-one variables. *Operations Research*, **13**(4), 517–546.

CHIERCHIA G. & MCCONNELL-GINET S. (1990). *Meaning and Grammar, An Introduction to Semantics*. MIT Press, Cambridge.

DAELEMANS W., ZAVREL J., VAN DER SLOOT K. & VAN DEN BOSCH A. (2004). TiMBL: Tilburg Memory-Based Learner.

DENIS P. & BALDRIDGE J. (2008a). Coreference with named entity classification and transitivity constraints and evaluation with MUC, B-CUBED, and CEAF. In *Proceedings of Corpus-Based Approaches to Coreference Resolution in Romance Languages (CBA 2008)*, Barcelona, Spain. À paraître.

DENIS P. & BALDRIDGE J. (2008b). Specialized models and ranking for coreference resolution. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2008)*, Hawaii, USA. À paraître.

FINKEL J. & MANNING C. (2008). Enforcing transitivity in coreference resolution. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics - Human Language Technology Conference*, p. 45–48: Association for Computational Linguistics.

KLENNER M. & AILLOUD E. (2008). Enhancing coreference clustering. In C. JOHANSSON, Ed., *Proc. of the Second Workshop on Anaphora Resolution (WAR II)*, volume 2 of *NEALT Proceedings Series*, p. 31–40, Bergen, Norway.

LUO X. (2005). On coreference resolution performance metrics. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, p. 25–32: Association for Computational Linguistics Morristown, NJ, USA.

LUO X., ITTYCHERIAH A., JING H., KAMBHATLA N. & ROUKOS S. (2004). A mention-synchronous coreference resolution algorithm based on the Bell tree. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*.

NG V. & CARDIE C. (2002). Improving machine learning approaches to coreference resolution. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 104–111.

SOON W., NG H. & LIM D. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, **27**(4), 521–544.

TELLJOHANN H., HINRICHS E., KÜBLER S. & ZINSMEISTER H. (2005). Stylebook for the Tübingen treebank of written German (TüBa-D/Z). *Seminar fur Sprachwissenschaft, Universität Tübingen, Tübingen, Germany*.

VILAIN M., BURGER J., ABERDEEN J., CONNOLLY D. & HIRSCHMAN L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, p. 45–52: Association for Computational Linguistics Morristown, NJ, USA.

YANG X., SU J., ZHOU G. & TAN C. (2004). An NP-cluster based approach to coreference resolution. *Proceedings of COLING*.