Université Paris Diderot
UFR de Linguistique

Revisiter la résolution des coréférences

Quentin Gliosca

Sous la direction de Pascal Amsili

Mémoire de M2 Recherche
Cursus de Linguistique Informatique

21 juin 2018

Introduction

En première approximation, la résolution des coréférences consiste à détecter des relations de coréférence entre mentions de référents de discours. Depuis Soon *et al.* (2001), de nombreux modèles statistiques ont été proposés pour résoudre ce problème, mais ce n'est que dans les années 2010 que les méthodes basées sur l'apprentissage ont définitivement dépassé les approches symboliques. L'adoption des réseaux de neurones par Wiseman *et al.* (2015) et les travaux ultérieurs ont récemment encore un peu plus creusé l'écart de performance. C'est dans cette lignée que notre travail se situe.

Au chapitre 1, nous présentons quelques concepts et phénomènes linguistiques qui nous serviront à définir au chapitre 2 plus précisément la résolution des coréférences telle qu'elle est pratiquée aujourd'hui en TAL. À cette fin, nous présenterons brièvement le contexte historique dans lequel la tâche a évolué, puis les schémas d'annotations en coréférences qui ont été proposés et les métriques d'évaluation les plus utilisées.

Au chapitre 3, nous présenterons le modèle mention-mention, l'approche statistique de la résolution des coréférences la plus simple, ainsi que les principaux jalons de son évolution. Le chapitre 4 sera l'occasion de pointer les défauts du modèle mention-mention avant d'introduire le modèle mention-entité, une alternative conceptuellement plus satisfaisante.

Nous terminerons au chapitre 5 notre tour d'horizon des principales approches actuelles de la résolution des coréférences par une présentation du modèle proposé par Lee *et al.* (2017), une extension du modèle mention-mention qui se veut « de bout en bout » du fait de l'abandon de la chaîne de prétraitement utilisée jusqu'alors.

Les chapitres suivants seront plus particulièrement consacrés à notre travail, largement inspiré par les différentes innovations introduites par Lee *et al.* (2017). Au chapitre 6, nous montrons le manque de connaissances syntaxiques dont souffre leur modèle et détaillons les pistes pour l'améliorer que nous avons commencé à explorer.

Comme il nous est ensuite apparu que cette faiblesse du modèle de Lee *et al.* (2017) était en fait extérieure à la résolution des coréférences à proprement parler et provenait d'un imbroglio dans la définition même de la tâche, les deux derniers chapitres se proposent d'explorer une nouvelle approche de la résolution des coréférences.

Au chapitre 7, nous plaidons en faveur d'une nouvelle définition de la tâche et d'une nouvelle méthode d'évaluation qui se focaliseraient davantage sur l'essence même de la résolution des coréférences. Enfin, le dernier chapitre introduit un nouveau modèle qui, en épousant de près notre nouvelle définition de la tâche, apporte un gain de performance accompagné d'une réduction significative des ressources calculatoires nécessaires par rapport au modèle de Lee *et al.* (2017).

Chapitre 1

Références et anaphores

La tâche de résolution des coréférences telle qu'elle est pratiquée aujourd'hui en TAL ne peut être rigoureusement définie qu'en isolant précisément les phénomènes linguistiques couverts parmi tous ceux ayant trait de près ou de loin à la référence ou l'anaphore. Nous tenterons donc dans ce chapitre de définir et illustrer brièvement quelques concepts et phénomènes linguistiques utiles pour cerner la résolution des coréférences.

1.1 Sémantique des syntagmes nominaux

Depuis Frege, la sémantique formelle a pour ambition de construire un modèle d'interprétation des langues aussi compositionnel que possible, c'est-à-dire de tel sorte que le sens d'un énoncé soit autant que possible calculable à partir du sens des différentes expressions linguistiques qui le composent et de la manière dont elles sont agencées.

Dans cette perspective, on peut catégoriser les expressions linguistiques par leur type de dénotation, ou de manière équivalente par leur fonction sémantique dans la phrase. De nombreuses expressions linguistiques sont ambiguës et leur fonction ne peut donc pas être déterminée a priori, mais seulement dans un certain contexte. C'est typiquement le cas des syntagmes nominaux. Une expression linguistique quelconque ne peut cependant pas avoir n'importe quelle fonction, sa catégorie syntaxique restreignant naturellement l'inventaire des possibilités.

1.1.1 Référentialité

Une des fonctions sémantiques que peut avoir une expression linguistique est de faire comprendre à son interlocuteur de quelle « chose » on souhaite parler, ou plus précisément, sur quel « objet » on souhaite prédiquer ; on les appelle expressions référentes [1] et le prototype en est le syntagme nominal.

Les syntagmes nominaux sont cependant, en français notamment, des formes hautement ambiguës. Selon leur contexte d'emploi, ils peuvent principalement avoir une fonction référentielle, prédicative, quantificationnelle ou explétive. Les noms propres sont presque

1. Le terme *expressions référentielles* est plus souvent utilisé, mais il nous semble à la fois inadéquat pour désigner en français quelque chose qui réfère et moins bien correspondre à l'anglais *referring expressions*.

exclusivement référents [2], mais les pronoms sont fréquemment explétifs et les descriptions définies ou indéfinies peuvent être aussi bien référentes que prédicatives.

Ainsi, le syntagme *un homme charmant* permet en (1-a) d'isoler un individu dont on souhaite parler, il est donc bien référent. En revanche, il est prédicatif en (1-b) puisqu'il dénote une propriété que l'on souhaite attribuer à Jean.

(1) a. *Un homme charmant* entra.
b. Jean est *un homme charmant*.

De la même manière, *Il* est référent en (2-a), mais explétif en (2-b) où son rôle est de fournir un sujet syntaxique vide à *faire froid* qui ne sélectionne aucun argument sémantique.

(2) a. *Il* s'appelle Jean.
b. *Il* fait froid aujourd'hui.

Les syntagmes quantificationnels, comme en (3), ont eux en français une forme spécifique qui les distingue systématiquement des autres syntagmes nominaux.

(3) *Tout homme* est mortel.

1.1.2 Cas particulier des génériques

Si les expressions référentes servent à désigner ce dont on veut parler, il arrive parfois qu'il ne s'agisse pas d'un individu particulier, mais d'un type, comme une espèce. L'exemple (4-a) atteste de la différence entre ce type d'expression et une expression quantificationnelle. Il est impossible d'appliquer la prédication *est menacé d'extinction* à chaque tigre du Bengale, elle s'applique plutôt au sens de l'expression *tigre du Bengale*, un concept.

Dans l'exemple (4-b), un autre cas de généricité, *un grand verre d'eau* n'est pas référent puisque sa dénotation varie avec chaque instantiation de l'événement *Jean boit un grand verre d'eau*.

(4) a. Le tigre du Bengale est menacé d'extinction.
b. Jean boit un grand verre d'eau le matin.

1.2 Anaphores et déictiques

L'interprétation des expressions référentes consiste à identifier l'élément de l'univers du discours que le locuteur souhaite désigner ou à en introduire un nouveau en élucidant les éventuels liens qu'il peut avoir avec les éléments déjà présents ou la situation d'énonciation. Si l'interprétation des noms propres va le plus souvent de soi, d'autres expressions référentes nécessitent pour être interprétées la prise en compte de leur contexte situationnel, les déictiques, ou linguistique, les anaphores.

2. *Je m'appelle X.* constitue une exception notable dans laquelle X ne doit pas être interprété comme la personne appellée X, mais comme l'expression linguistique elle-même.

1.2.1 Introduction

Les pronoms de 1re et 2e personnes sont les exemples prototypiques de déictiques. Par exemple, *tu* se rapporte en (5-a) à l'interlocuteur, son référent ne peut donc être déterminé qu'en fonction de la situation d'énonciation. D'autres types d'expressions référentes peuvent cependant également être déictiques, comme *le vélo* en (5-b) lorsque le locuteur désigne un vélo qu'il aperçoit à son interlocuteur.

(5) a. Merci, *tu* es gentil !
b. Regarde *le vélo* !

La dépendance de l'interprétation vis-à-vis de la situation d'énonciation peut également passer par la référence à une portion du discours, comme en (6) dans le cas où le locuteur fait allusion à ce que son interlocuteur vient de dire. On parle dans ce cas-là de déictiques discursifs.

(6) Chut ! Ne prononce pas *ce mot* !

L'interprétation des anaphores dépend elle du contexte linguistique. Les pronoms personnels de 3e personne dans leur usage le plus fréquent, comme en (7-a), constituent le prototype des anaphores. En toute rigueur, on ne parle d'anaphore que lorsque l'expression linguistique dont son interprétation dépend, son antécédent, la précède. Dans le cas contraire, comme en (7-b), on parle plutôt de cataphore. Le terme d'anaphore est cependant couramment utilisé pour parler indifféremment de l'un ou l'autre des deux cas.

(7) a. Tu connais Jean ? *Il* est gentil.
b. Lorsqu'*il* rentra, Jean partit directement se coucher.

1.2.2 Typologie des anaphores

Le lien entre une anaphore et son antécédent n'est pas obligatoirement l'identité de référence. Il peut aussi s'agir d'un lien indirect comme celui entre un adjectif possessif et son antécédent, ou d'une relation d'association comme en (8). Les anaphores reliées à leur antécédent par ces trois types de relations sont de loin les plus fréquentes et nous les appellerons respectivement anaphores coréférentes, possessives et associatives.

(8) a. Je n'achèterai jamais *cette maison*, *le toit* tombe en ruine.
b. Tu vois *l'orchestre* ? Je connais *le violoniste*.

La nature exacte de la relation entre une anaphore associative et son antécédent est assez variée. Il peut par exemple s'agir d'une relation de partie-tout, comme la relation entre *cette maison* et *le toit* en (8-a), ou d'une relation d'appartenance comme entre *l'orchestre* et *le violoniste* en (8-b).

Il existe cependant encore d'autres types d'anaphores, même si elles sont un peu plus rares. Certaines, qualifiées de virtuelles par Milner (1982), reprennent le sens d'une expression linguistique référente, et non son référent. C'est typiquement le cas des pronoms indéfinis associés au clitique *en*, comme en (9-a). L'exemple (9-b) illustre quant à lui le cas des anaphores liées qui dénotent la variable liée par le quantificateur.

Enfin, en (9-c), l'antécédent de *C* est *directeur des ventes* alors que cette expression n'est pas référente mais prédicative. Il ne peut donc pas y avoir identité de référence entre

les deux expressions et le référent de C est plutôt le poste de directeur des ventes occupé par Jean dont l'existence est inférée lors de l'interprétation de la première phrase.

(9) a. Jean a vu une mouette avant que je n'*en* vois *une*.
b. Toute solution est bonne du moment qu'*elle* conduit à la victoire.
c. Jean est devenu directeur des ventes. *C*'est un poste important.

1.2.3 Anaphores abstraites

Une anaphore coréférente se rapporte le plus souvent à une entité, comme en (10-a). Elle peut cependant aussi se rapporter à un événement (10-b), un fait (10-c) ou une proposition (10-d). Ces trois derniers cas ont été étudiés en détail par Asher (1993) sous le nom d'anaphores abstraites.

(10) a. *Jean* est dans la cuisine. *Il* va partir.
b. Je l'ai vu *partir en courant* et ce n'est pas la première fois que *ça* arrive.
c. Elle m'a dit qu'elle était déjà couchée. *Ça* montre qu'elle était fatiguée.
d. Tu m'as dit hier que *s'il venait, tu partais*. Je *le* crois volontiers.

Chapitre 2

Résolution des coréférences

2.1 Repères historiques

2.1.1 Tâche

Depuis le milieu des années 1990, la tâche de TAL liée aux références et anaphores qui a reçu le plus d'attention est ce qui a été appelé la résolution des coréférences. Elle tire son nom du fait qu'elle consiste à partitionner les portions de texte répondant à certains critères, appelées *mentions*, en groupes appelés *chaînes de coréférence*, dont les différents éléments sont censés être coréférents. La relation entre deux éléments d'une même chaîne de coréférence étant très variable, il s'agit en réalité d'un mélange de résolution des coréférences et de résolution des anaphores.

Plusieurs corpus dédiés à cette tâche ont été élaborés aussi bien pour évaluer les différents algorithmes que pour permettre le développement de modèles statistiques par apprentissage supervisé. Les premiers corpus, constitués dans le cadre des 6e (MUC 6, 1995) et 7e (MUC 7, 1997) éditions des *Message Understanding Conference* (MUC), ont par leur schéma d'annotation largement influencé la définition de la tâche, même si elle a fait l'objet de quelques raffinements ultérieurs.

Les recherches réalisées dans le cadre des MUC étaient motivées par l'extraction d'information, domaine dans lequel il est important de détecter les différentes entités d'un texte et d'identifier les portions qui en parlent. Cet objectif a eu un impact décisif sur la définition de la tâche de résolution des coréférences, notamment par l'attention portée à la résolution des pronoms, qu'ils soient référents ou non.

Aux MUC ont succédé le programme de recherche *Automatic Content Extraction* (ACE), émanant lui aussi du gouvernement américain. Également orienté vers l'extraction d'informations, il a lui beaucoup plus clairement assumé son orientation pratique en abandonnant la terminologie linguistique et renommant la tâche *entity detection and tracking*, soit détection et suivi d'entités, ce qui décrit à notre avis beaucoup mieux ce qu'est encore aujourd'hui la résolution des coréférences en TAL.

Pour bien comprendre en quoi celle-ci consiste, il est nécessaire d'étudier au travers des différents schémas d'annotation de corpus propoposés la variété des phénomènes couverts. C'est ce que nous ferons à la section 2.2 en nous appuyant sur les guides d'annotation des trois corpus ayant successivement servi de standard de facto, après les avoir brièvement présentés.

2.1.2 Corpus

Si les premiers corpus annotés en coréférence, MUC 6 et MUC 7, étaient en anglais et reposaient sur un ensemble d'articles du *Wall Street Journal*, ceux publiés entre 2000 et 2008 dans le cadre des campagnes ACE élargissent l'éventail des langues à l'arabe et au chinois et incluent des transcriptions de journaux radiodiffusés en plus de la traditionnelle presse écrite. Leur schéma d'annotation s'éloigne pour sa part peu de celui des corpus MUC.

Suite à la mise au jour du manque de motivations linguistiques des choix faits pour le schéma d'annotation des corpus MUC (Deemter et Kibble, 2000), apparaît également dans les années 2000 une seconde génération de corpus qui se distingue davantage du schéma d'annotation MUC que les corpus ACE. Le schéma d'annotation développé dans le cadre du projet MATE, pour une large part appliqué au corpus GNOME, (Poesio, 2004) est moins taillé pour une application spécifique et linguistiquement plus motivé. Il fait par exemple la différence entre la coréférence à proprement parler et la prédication.

C'est néanmoins le corpus Ontonotes (Hovy *et al.*, 2006) qui, après les campagnes d'évaluation *SemEval 2010*, *CoNLL 2011* et *CoNLL 2012*, a succédé aux corpus ACE comme standard de facto pour l'apprentissage et l'évaluation des systèmes de résolution des coréférences. Constitué de documents aussi variés que des articles de magazines et blogs ou encore des transcriptions de médias audiovisuels et de conversations téléphoniques, il comprend une partie en anglais, une partie en chinois et une plus petite partie en arabe.

2.2 Schémas d'annotation

Dans le cadre d'une tâche aux contours aussi flous que la résolution des coréférences, chaque schéma d'annotation crée une variante de la tâche. Pour tenter de mieux cerner l'objet de la résolution des coréférences, les principaux points de convergence et de divergence des schémas d'annotation des corpus MUC, ACE et Ontonotes sont présentés ci-dessous.

2.2.1 Règles générales

Délimitation des mentions

Comme nous l'avons vu au chapitre 1, les expressions référentes dénotent par définition des référents de discours. Plusieurs expressions imbriquées peuvent cependant dénoter un même référent de discours et il n'est pas aisé de choisir laquelle devra être considérée comme une mention.

Le stratégie généralement adoptée se fonde sur un critère syntaxique non trivial : une mention doit être un syntagme maximal, c'est-à-dire correspondre à la projection maximale d'un des mots de la phrase, et de ce fait inclure tous les modifieurs. En (3) par exemple, c'est *Hsu Tsang-houei, le doyen de la musique taïwanaise* qui doit être retenu comme mention, plutôt que *Hsu Tsang-houei*.

(1) Hsu Tsang-houei, le doyen de la musique taïwanaise, est décédé à la suite d'une chute.

Pour pallier la difficulté de prédire des syntagmes maximaux, les guides d'annotation des corpus MUC et ACE prévoient de fournir une double délimitation des mentions : le syntagme maximal correspondant, mais aussi sa tête syntaxique. [1] Comme nous le verrons à la section 2.3, cette double délimitation a permis d'introduire une certaine souplesse dans la délimitation des mentions lors de l'évaluation. Ontonotes a quant à lui choisi l'approche minimaliste de n'annoter que les syntagmes maximaux.

Restrictions de catégories syntaxiques et types sémantiques

La tâche de résolution des coréférences telle que définie par les MUC était initialement restreinte aux mentions nominales (noms, pronoms et syntagmes nominaux), sans limitation sur le type sémantique du référent qu'elles dénotaient.

En revanche, les corpus ACE restreignent les référents pris en considération à quelques grandes classes sémantiques jugées particulièrement utiles en extraction d'information : personnes, organisations, lieux, véhicules, entités géopolitiques, etc.

Le corpus ACE 2005 se singularise par l'inclusion de certaines classes d'événements parmi les types de référents considérés (naissance, mariage, décès, fusion d'organisations, attaque, etc.) L'éventail des catégories syntaxiques autorisées est en conséquence élargi aux verbes et on parle d'*event detection and tracking*, soit détection et suivi événements.

Comme MUC, Ontonotes ne pose pas de restrictions sur les types sémantiques, mais autorise comme ACE les mentions verbales, à la condition toutefois que celles-ci soient coréférentes avec au moins une mention nominale. Par exception à la règle générale de délimitation, les mentions verbales sont impérativement constituées d'un unique mot, la tête du syntagme verbal.

Relation de coréférence

Le guide d'annotation des MUC, et les suivants à des degrés divers, définissent les chaînes de coréférence en s'appuyant sur une relation d'équivalence entre mentions dont elles constituent la fermeture transitive. Dans les corpus MUC et ACE, l'éventail des liens entre deux mentions considérés comme des liens de coréférence est extrêmement large.

Outre la coréférence réelle entre deux expression linguistiques référentes, leur notion de coréférence comprend les relations de prédication et d'apposition. Comme l'ont montré Deemter et Kibble (2000), cela pose d'importants problèmes de cohérence.

Dans le cas de l'exemple (2), *Jacques Langlois* est considéré comme coréférent aussi bien de *chef de rayon chez Carrefour* que de *directeur des ventes*. Par transitivité, *chef de rayon chez Carrefour* et *directeur des ventes* sont coréférents alors qu'ils ne s'est agi à aucun moment de la même personne.

(2) Jacques Langlois, jusqu'à présent chef de rayon chez Carrefour, est devenu directeur des ventes.

Le problème tire son origine du fait que l'interprétation d'un syntagme nominal dans un emploi prédicatif ou appositif est son sens, et non un référent. À la suite de MATE, Ontonotes revient donc sur les choix des MUC en excluant des liens de coréférence les relations d'apposition ou de prédication entre deux syntagmes.

1. Dans le cas des corpus MUC, il ne s'agit pas systématiquement de la tête syntaxique. Par exemple, dans le cas d'un nom propre, la tête annotée sera constituée de l'ensemble du nom propre, et pas seulement d'un des mots qui le composent.

Ontonotes ne s'est cependant pas complètement affranchi de la libéralité des MUC concernant la notion de coréférence. Par exemple, tout type de lien entre un pronom référent et son antécédent sont considérés comme valant coréférence. Des syntagmes prédicatifs ou quantificationnels sont donc de ce fait de potentiels antécédents.

La confusion entre coréférence et anaphore subsiste également dans le cas des adjectifs possessifs qui sont considérés comme coréférents avec leur antécédent, le possesseur de la chose dénotée par le nom modifié. C'est par exemple le cas de *Jean* et *sa* en (3).

(3) *Jean* est marié à Marie et *sa* sœur s'appelle Jeanne.

Singletons

Dans les corpus MUC et Ontonotes, les mentions qui ne sont considérées comme coréférentes avec aucune autre, dites *mentions singletons*, ne sont pas du tout annotées. Comme nous le verrons à la section 2.3, cela complique l'évaluation de la résolution des coréférences.

Les campagnes ACE étant pour leur part résolument orientées vers une caractérisation précise de chaque entité (ou événement) mentionnée, toutes les mentions ont été annotées dans les corpus, qu'elles soient coréférentes avec d'autres ou non.

2.2.2 Quelques cas particuliers

Du fait de la complexité de la langue, les guides d'annotation ont dû prévoir des règles *ad-hoc*, ou tout au moins des précisions, pour gérer de nombreux cas particuliers susceptibles de faire douter les annotateurs. Nous en présenterons brièvement quelques-uns.

Antécédents discontinus

Les mentions devant être des constituants syntaxiques continus, aussi bien les corpus MUC que les corpus ACE et Ontonotes ignorent les pronoms dont les antécédents sont discontinus comme *ils* en (4).

(4) *Jean* est allé au cinéma avec *Marie*. *Ils* sont rentrés tard.

Génériques

Le schéma d'annotation de MUC ne pose pas de restrictions particulières sur les cas de généricité. Ainsi, deux syntagmes nominaux dénotant le même type, comme *Le tigre du Bengale* et *Il* en (5), seront considérés comme coréférents.

(5) *Le tigre du Bengale* est menacé d'extinction. *Il* risque le même sort que le tigre de Java.

De son côté, Ontonotes autorise également ce type de lien, mais, à la différence de MUC, uniquement parce que *Il* est un pronom. En effet, deux syntagmes non pronominaux dénotant le même type ne seront pas considérés comme coréférents.

Métonymies

Enfin, le traitement des métonymies nécessite quelques précautions. En (6) par exemple, les deux mentions en italique ne semblent pas coréférentes avant coercion [2] de la métonymie, mais après si. Les concepteurs des schémas d'annotation de MUC, ACE et Ontonotes ont choisi que la coréférence serait décidée après coercion des métonymies, ce qui correspond bien ici à l'intuition.

(6) *La Maison blanche* a déposé son projet de loi sur la santé au C ongrès hier. Le sénateur Dole dit que le projet de *l'administration* a peu de chance de passer.

Les entités géopolitiques sont des cas particulièrement problématiques. Pour les traiter, une classe sémantique les regroupant aussi bien dans leur dimension géographique que politique a été postulée. Ainsi les deux occurrences de *la Russie* sont-elles considérées comme coréférentes en (7).

(7) *La Russie* est une fédération. [...] La superficie de *la Russie* est de 17,125 millions de km^2.

2.2.3 Conclusion

Les hésitations des schémas d'annotation reflètent dans une certaine mesure le flou dans la définition de la tâche dont a souffert la résolution des coréférences depuis les MUC. Si Ontonotes a essayé de davantage s'appuyer sur la notion de référence linguistique, il ne s'est pas complètement affranchi du cadre initial posé par les corpus MUC.

Certaines de ses innovations ont en outre introduit de nouveaux problèmes. C'est notamment le cas des verbes dont les règles d'annotation sont conçues pour qu'ils puissent servir d'antécédents, mais ne les reconnaissent pas comme des mentions à part entière, capables de coréférer avec une mention de n'importe quelle catégorie syntaxique. Deux mêmes verbes peuvent ainsi être marqués comme coréférents parce qu'ils sont également coréférents avec un syntagme nominal, et deux autres non annotés parce que ce n'est pas le cas.

2.3 Métriques d'évaluation

Bien que souvent considérée comme un problème de classification [3], la résolution des coréférences est en réalité une combinaison de deux tâches : l'identification des mentions et leur partitionnement en chaînes de coréférence. Seule cette deuxième tâche est bien un problème de classification, la première étant formellement un problème de classement binaire de chacun des empans du texte.

Évaluer ces deux sous-tâches séparément présente certaines difficultés du fait que dans la majorité des corpus disponibles, les mentions singletons ne sont pas annotées. Si l'évaluation des systèmes de résolution des coréférences en utilisant les mentions de référence,

2. On appelle ici coercion d'une métonymie le glissement de sens du contenant vers le contenu. La métonymie fait donc référence avant coercion au contenant, et après coercion au contenu.

3. On emploiera ici la terminologie française de Miclet et Cornuéjols (2010). Dans cette terminologie, le terme anglais *clustering* correspond en français à la *classification*, à savoir l'action de construire un plan de classement pour un ensemble d'éléments, et le terme anglais *classification* correspond en français au *classement*, l'action de ranger un élément dans une des classes selon un plan de classement préétabli.

donc sans les singletons, a parfois été pratiquée, il s'est avéré que c'était en fait un problème bien plus facile à résoudre que le problème original (Stoyanov *et al.*, 2009). En ne considérant qu'un nombre restreint de mentions qui doivent à coup sûr faire partie des chaînes de coréférence, la marge d'erreur des systèmes de résolution est limitée.

D'autre part, l'identification des mentions dans des corpus qui n'annotent pas les singletons ne peut pas être évaluée indépendamment car elle est inextricablement liée à la résolution des coréférences elle-même. En effet, déterminer si un empan est une mention implique de décider s'il a un antécédent ou sert d'antécédent à une autre mention, ce qui ne peut *a priori* être déterminé qu'après résolution des coréférences.

Des métriques spécifiques ont donc été conçues pour évaluer de façon conjointe les deux sous-tâches de la résolution des coréférences. Aucune n'étant parfaite, la littérature sur le sujet est abondante et le choix des métriques fait encore débat. Toutes ont cependant pour point commun de proposer trois scores : un rappel, une précision, et un score F1, la moyenne harmonique [4] des deux précédents.

Les erreurs dans les bornes des mentions prédites ont été traitées avec une sévérité plus ou moins grande selon l'époque et le corpus utilisé. Lors des campagnes d'évaluation MUC, une prédiction était considérée comme correcte dès lors que la mention prédite contenait la tête annotée et qu'elle ne dépassait pas du syntagme maximal. La plus récente campagne SemEval 2010 s'est montré un peu plus sévère : 1 point était accordé si la mention prédite coïncidait parfaitement avec celle annotée, mais seulement 0.5 dans le cas où elle contenait sa tête et ne s'étendait pas au-delà des bornes de référence.

C'est aux campagnes CoNLL 2011 et CoNLL 2012 qu'on doit la méthodologie d'évaluation actuelle qui fait preuve de la sévérité maximale. Dans ce cadre-là, une mention prédite n'est considérée comme correcte que si ses bornes coïncident exactement avec celles de référence. Ce choix est cohérent avec la décision prise par les concepteurs d'Ontonotes de ne fournir dans les annotations qu'une unique délimitation des mentions : les syntagmes maximaux.

Nous présenterons dans ce qui suit les trois métriques les plus utilisées, MUC, B^3 et $CEAF_e$ en nous inspirant largement de la synthèse proposée par Poesio *et al.* (2016). Pour chacune d'entre elles, nous appelerons $\mathcal{K}$ (comme *Keys*) l'ensemble des chaînes de coréférence annotées manuellement et $\mathcal{R}$ (comme *Responses*) celui des chaînes prédites par le système à évaluer.

2.3.1 MUC

Définition

La métrique MUC (Vilain *et al.*, 1995) a été conçue pour servir d'évaluation à la tâche de résolution des coréférences de la 7e édition des MUC. Elle évalue la similarité entre les chaînes de coréférence prédites et celles de référence en considérant le nombre minimum de liens de coréférence qu'il faudrait changer dans les chaînes prédites pour obtenir celles de référence et *vice-versa*.

Plus précisément, le rappel évalue le nombre minimal de liens qu'il faudrait ajouter dans les chaînes de coréférence prédites pour obtenir celles de référence, et la précision le nombre minimal de liens qu'il faudrait ajouter aux chaînes de référence pour obtenir celles prédites.

4. Pour une précision p et un rappel r strictement positifs, le score F1 est $\frac{2rp}{r+p}$.

Vilain *et al.* (1995) fondent leur calcul sur des partitions $\mathcal{P}(K_i, \mathcal{R})$ de chaque entité de référence K_i par intersection avec les entités prédites $R_j \in \mathcal{R}$. Plus précisément,

$$\mathcal{P}(K_i, \mathcal{R}) = \{K_i \cap R_j \mid R_j \in \mathcal{R}\} \cup \bigcup_{m \in K_i - \mathcal{R}} \{\{m\}\}$$

Le nombre minimum de liens entre mentions nécessaire pour former l'ensemble K_i est $|K_i| - 1$ et celui nécessaire pour connecter les différents éléments de $\mathcal{P}(K_i, \mathcal{R})$ est $|\mathcal{P}(K_i, \mathcal{R})| - 1$. C'est la somme sur les entités de référence $K_i \in \mathcal{K}$ de la différence entre ces deux quantités qui forme le numérateur du rappel r, qui est donc :

$$r = \frac{\sum_{K_i \in \mathcal{K}} |K_i| - |\mathcal{P}(K_i, \mathcal{R})|}{\sum_{K_i \in \mathcal{K}} |K_i| - 1}$$

On peut calculer la précision de la même manière en partitionnant cette fois les entités prédites par intersection avec les entités de référence :

$$\mathcal{P}(R_i, \mathcal{K}) = \{R_i \cap K_j \mid K_j \in \mathcal{K}\} \cup \bigcup_{m \in R_i - \mathcal{K}} \{\{m\}\}$$

La précision p est donc :

$$p = \frac{\sum_{R_i \in \mathcal{R}} |R_i| - |\mathcal{P}(R_i, \mathcal{K})|}{\sum_{R_i \in \mathcal{R}} |R_i| - 1}$$

Critiques

Deux critiques principales ont été formulées à l'encontre de la métrique MUC. D'abord, l'agglomération erronée de deux chaînes de coréférence est pénalisée de la même façon quelle que soient leurs tailles puisque, dans tous les cas, un unique lien est responsable de cette fusion. Il est pourtant fondamental pour une métrique d'évaluation de classifications, et *a fortiori* de résolution des coréférences, de prendre en compte le fait que regrouper à tort deux petits ensembles est moins grave que d'en regrouper deux très gros.

La seconde critique, étroitement liée, est que MUC favorise les grosses chaînes de coréférence puisqu'une agglomération erronée entraîne une augmentation du nombre de liens erronés de seulement 1. Ce problème conduit à des résultats complètement contre-intuitifs dans lesquels notamment un partitionnement en une seule chaîne de coréférence aura un score plus élevé que bien d'autres pourtant de meilleure qualité.

Reprenons en guise d'illustration l'exemple de Poesio *et al.* (2016) :

$$K = \{\{1, 2, 3, 4, 5\}, \{6, 7\}, \{8, 9, A, B, C\}\}$$

$$R_1 = \{\{1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C\}\}$$

$$R_2 = \{\{1, 2, 3, 4, 5\}, \{6, D\}, \{8, 9, A, B, C\}\}$$

R_2 est intuitivement bien plus proche de K que R_1 puisque qu'elle contient deux importantes chaînes de coréférence entièrement correctes et une autre qui l'est pour moitié. Pourtant, R_1 obtient un score F1 de $\frac{9}{10}$ et R_2 un score de $\frac{8}{9}$, inférieur à celui de R_1.

2.3.2 B^3

Définition

La métrique B^3 a été proposée par Bagga et Baldwin (1998) pour corriger les défauts de MUC. Elle permet en particulier de pénaliser plus fortement un lien erroné s'il connecte deux chaînes de coréférence de tailles importantes que deux chaînes plus petites.

Le rappel et la précision sont des moyennes de scores calculés respectivement pour chaque mention de référence et chaque mention prédite.

Le rappel r est donné par :

$$r = \frac{\sum_{K_i \in \mathcal{K}} \sum_{R_i \in \mathcal{R}} \frac{|K_i \cap R_i|^2}{|K_i|}}{\sum_{K_i \in \mathcal{K}} |K_i|}$$

Et la précision p par :

$$p = \frac{\sum_{R_i \in \mathcal{R}} \sum_{K_i \in \mathcal{K}} \frac{|K_i \cap R_i|^2}{|R_i|}}{\sum_{R_i \in \mathcal{R}} |R_i|}$$

Critiques

La métrique B^3 a pour défaut un comportement indésirable lorsqu'elle est utilisée pour évaluer une réponse fondée sur des mentions prédites, donc potentiellement inexactes. En effet, elle attribue systématiquement un rappel de 1 à une réponse dans laquelle toutes les mentions prédites sont ensemble, même s'il en manque, et une précision de 1 lorsque chaque mention prédite est dans sa propre chaîne de coréférence, même si toutes ne sont pas correctes.

2.3.3 $CEAF_e$

Définition

Introduite par Luo (2005), $CEAF_e$ aborde le problème de l'évaluation des systèmes de résolution des coréférences encore différemment puisqu'elle nécessite de déterminer au préalable un alignement entre les entités de référence et les entités prédites. Cet alignement est choisi au regard d'une mesure ϕ de similarité entre deux chaînes de coréférence K_i et R_j donnée par

$$\phi(K_i, R_j) = \frac{2|K_i \cap R_j|}{|K_i| + |R_j|}$$

.

L'éventuelle différence entre le nombre d'entités de référence et le nombre d'entités prédites est gérée en considérant les alignements entre m chaînes de coréférence de $\mathcal{K}$ et m chaînes de $\mathcal{R}$ avec $m = min(|\mathcal{K}|, |\mathcal{R}|)$. Appelons G_m l'ensemble de ces alignements et notons $\mathcal{D}_g$ le domaine de définition d'un alignement g. L'alignement optimal g^* pour la mesure de similarité ϕ est défini par

$$g^* = \underset{g \in G_m}{\operatorname{argmax}} \sum_{K_i \in \mathcal{D}_g} \phi(K_i, g(K_i))$$

La complexité du calcul de cet alignement optimal par un algorithme naïf est factorielle en m, mais peut être ramenée en $O(Mm^2log(m))$ par l'algorithme de Kuhn-Munkres de recherche du couplage de poids maximal avec $M = max(|\mathcal{K}|, |\mathcal{R}|)$.

Une fois l'alignement g^* déterminé, le rappel et la précision sont respectivement donnés par :

$$r = \frac{\sum_{K_i \in \mathcal{D}_g^*} \phi(K_i, g^*(K_i))}{\sum_{K_i \in \mathcal{K}} \phi(K_i, K_i)}$$

et

$$p = \frac{\sum_{K_i \in \mathcal{D}_g^*} \phi(K_i, g^*(K_i))}{\sum_{R_i \in \mathcal{R}} \phi(R_i, R_i)}$$

Critiques

Le défaut le plus évident de CEAF_e est d'ignorer complétement les chaînes de coréférence prédites qui ne participent pas à l'alignement optimal, alors qu'elles peuvent être partiellement correctes. Un système qui prédit une multitude de petites entités homogènes sera par exemple fortement pénalisé.

2.3.4 Conclusion

La résolution des coréférences est composée de deux sous-tâches : l'identification des mentions et leur partitionnement en chaînes de coréférence. Du fait que nombre de corpus n'annotent pas les mentions singletons, ces deux tâches doivent être évaluées conjointement.

De nouvelles métriques sont régulièrement proposées, mais les plus utilisées aujourd'hui sont MUC, B^3 et CEAF_e. En particulier, la moyenne de ces trois métriques a été utilisée pour départager les systèmes lors des campagnes d'évaluation CoNLL 2011 et 2012. Elle est depuis largement utilisée sous le nom de score CoNLL car elle permet d'attribuer un score unique aux systèmes de résolution des coréférences.

Chapitre 3

Modèle mention-mention

Nous présenterons dans ce chapitre une approche très largement répandue de la résolution des coréférences connue en anglais sous les noms de *mention-pair model* et *mention-ranking model*, selon le type de modèle statistique utilisé.

C'est dans le cadre de cette approche que plusieurs innovations comme les modèles d'ordonnement, les descripteurs lexicaux et les réseaux de neurones ont été introduites dans les systèmes de résolution des coréférences. Nous profiterons donc également de ce chapitre pour retracer une part importante de l'histoire de la résolution des coréférences statistique.

3.1 Généralités

Même si nous verrons plus tard que les travaux de Lee *et al.* (2017) ont remis cette segmentation en question, l'approche classique de la résolution des coréférences consiste à traiter en cascade ses deux sous-problèmes : l'identification des mentions et leur partitionnement en chaînes de coréférence. Nous décrirons dans un premier temps les processus mis en œuvre par chacune de ces sous-tâches.

3.1.1 Identification des mentions

Dans l'approche classique de la résolution des coréférences, le texte à traiter passe avant la résolution des coréférences à proprement parler par une importante chaîne de traitement qui permet à la fois d'identifier les mentions possibles et de les caractériser par certaines propriétés linguistiques qui serviront d'entrée au modèle de résolution des coréférences.

Cette chaîne de traitement comporte au minimum, outre la segmentation en phrases et en mots, l'étiquetage morphosyntaxique et morphologique, la reconnaissance d'entités nommées ainsi que l'identification des syntagmes nominaux, voire une analyse syntaxique complète.

Le *Berkeley Coreference System* (Durrett et Klein, 2013) sélectionne comme mentions les entités nommées non-numériques, les mots étiquetés comme pronoms ou pronoms possessifs, ainsi que l'ensemble des syntagmes nominaux maximaux. D'autres systèmes appliquent un filtrage pour augmenter la précision, mais c'est généralement le rappel qui est considéré comme important lors de l'identification des mentions.

3.1.2 Partitionnement en chaînes de coréférence

Le partitionnement en chaînes de coréférence des mentions est formellement un pur problème de classification. Beaucoup de stratégies sont envisageables pour l'aborder, mais celle qui a servi de base aux travaux en résolution des coréférences par apprentissage supervisé, le modèle mention-mention, s'inspire des particularités de la tâche à réaliser.

Comme nous l'avons vu, le phénomène de l'anaphore est une des principales motivations de la tâche de résolution des coréférences et il est statistiquement très important. Les pronoms représentent par exemple à eux seuls plus de 40 % des mentions dans le corpus Ontonotes.

Dans le modèle mention-mention, décrit dans ce chapitre, les mentions du document sont traitées les unes à la suite des autres de gauche à droite. Pour chacune d'entre elles, un antécédent est éventuellement choisi parmi les mentions qui la précèdent. Les chaînes de coréférence sont ensuite construites par fermeture transitive de la relation d'antécédence obtenue.

3.2 Le modèle par classement

En 2001, Soon *et al.* ont proposé le premier modèle statistique de résolution des coréférences applicable à l'ensemble des syntagmes nominaux et dont les performances égalaient les systèmes symboliques de l'époque. Leur travail a de ce fait servi de base à nombre de méthodes proposées dans les années 2000.

3.2.1 Modélisation

Dans l'approche proposée par Soon *et al.* (2001), le modèle statistique prend la forme d'une fonction de classement binaire entraînée à prédire si l'association entre une mention et un candidat antécédent est vraisemblable ou non. La fonction de classement prend en l'occurrence la forme d'un arbre de décision, mais ce sont les modèles logistiques (Berger *et al.*, 1996) qui s'imposent quelques années plus tard dans le domaine.

Construction des chaînes de coréférence

Dans le cadre du modèle mention-mention par classement, le processus de résolution des coréférences consiste à parcourir les mentions de gauche à droite et à confronter chacune d'entre elles à toutes celles qui les précèdent pour déterminer leur éventuel antécédent. Le classeur peut répondre négativement pour tous les candidats antécédents, auquel cas la mention introduit une nouvelle chaîne de coréférence et on passe à la suivante. Il peut cependant aussi prédire plusieurs antécédents pour une même mention. Plusieurs stratégies sont alors possibles.

Celle adoptée par Soon *et al.* (2001), dite *closest-first*, est inspirée par la localité des pronoms et fait l'hypothèse que l'antécédent devrait être situé le plus près possible de la mention dès lors que le classeur préconise le lien. Pour chaque mention, les mentions précédentes sont donc parcourues de droite à gauche et la recherche est arrêtée dès que le classeur a autorisé un couple mention-antécédent.

Ng et Cardie (2002b) proposent une stratégie alternative, dite *best-first*, qui ne conserve parmi les liens d'antécédence avec une mention prédits par le classeur que celui qui a obtenu le score le plus élevé. La fermeture transitive peut alors être calculée sans risquer

un trop grand nombre de fusions intempestives de chaînes de coréférence puisque seuls les liens de coréférence les plus vraisemblables sont retenus.

Si l'on souhaite au contraire favoriser le rappel plutôt que la précision, la stratégie *agressive* propose de ne pas s'imposer d'avoir au plus un antécédent pour chaque mention. La fermeture transitive est alors calculée à partir de l'ensemble des liens d'antécédence autorisés par le classeur.

Exemples d'apprentissage

Les algorithmes d'apprentissage de classeurs binaires nécessitent en général d'avoir des exemples positifs et des exemples négatifs en quantités comparables. À défaut, le classeur entraîné tend à prédire systématiquement la classe majoritaire.

Or, utiliser l'intégralité des liens et non-liens d'antécédence [1] inférables à partir des chaînes de coréférence annotées conduit à un important déséquilibre en faveur des exemples négatifs. Une stratégie *ad-hoc* de construction des exemples est donc nécessaire.

Soon *et al.* (2001) retiennent comme exemples positifs les couples associant une mention et son antécédent le plus proche. Pour former les exemples négatifs, les mentions sont appariées avec tous les candidats antécédents situés entre elles et leur antécédent le plus proche.

La méthode de construction des exemples étant étroitement liée à la stratégie de génération des chaînes de coréférence adoptée, Ng et Cardie (2002b) proposent dans le cadre de la stratégie *best-first* de construire les exemples positifs à partir des antécédents « les plus sûrs ». Pour un pronom, il s'agit toujours de l'antécédent le plus proche, mais pour les autres mentions, ils proposent d'utiliser l'antécédent non pronominal le plus proche.

3.2.2 Descripteurs

Pour modéliser le problème par une fonction de classement, les couples associant une mention m et un candidat antécédent c doivent être représentés par un vecteur que l'on notera $\Phi(m, c)$. Chaque composante de ce vecteur doit représenter une caractéristique de la paire pertinente pour inférer s'il s'agit d'une paire mention-antécédent crédible ou non.

Soon *et al.* (2001) proposent les 12 descripteurs suivants :

1. le nombre de phrases entre m et c (0 s'ils sont dans la même phrase) (`DIST`)
2. si m est un pronom (personnel ou possessif) ou un adjectif possessif (`I_PRONOUN`)
3. si c est un pronom ou un adjectif possessif (`J_PRONOUN`)
4. si m et c sont identiques après suppression d'éventuels déterminants (`STR_MATCH`)
5. si m est un syntagme défini, c'est-à-dire commence par *the* (`DEF_NP`)
6. si m est un syntagme démonstratif, c'est-à-dire commence par *this*, *that*, *these* ou *those* (`DEM_NP`)
7. si m et c s'accordent en nombre (`NUMBER`)
8. si les classes sémantiques de m et c sont compatibles (`SEMCLASS`)
9. si m et c s'accordent en genre ou si l'une des mentions a un genre inconnu (`GENDER`)
10. si m et c sont toutes les deux des noms propres (`PROPERNAME`)

1. On appelle ici antécédents d'une mention l'ensemble des mentions de la même chaîne de coréférence qui la précédent.

11. si m et c sont des alias l'une de l'autre (`ALIAS`)
12. si m est apposée à c (`APPOSITIVE`)

Classes sémantiques

Les classes sémantiques utilisées sont *femme*, *homme*, *personne*, *organisation*, *lieu*, *date*, *heure*, *somme d'argent*, *pourcentage* et *objet*, organisées au sein d'une hiérarchie de subsomption.

Chacune de ces classes est associée au sens le plus pertinent au sein du réseau sémantique WordNet (Miller, 1995).

Une mention est associée au sens le plus fréquent de sa tête. La classe sémantique de la mention est la classe la plus proche dans la hiérarchie des concepts parmi les parents du sens sélectionné. Les classes sémantiques de deux mentions sont considérées comme compatibles si l'une est descendante de l'autre ou si ce sont les mêmes.

Alias

Deux mentions sont des alias l'une de l'autre si le fait qu'elles se rapportent à la même entité nommée peut se déduire facilement à l'aide de quelques transformations. S'il s'agit de dates, les différences de format sont neutralisées avant comparaison. S'il s'agit de deux mentions de personne, elles sont comparées par leur dernier mot. Enfin, deux mentions d'organisation sont des alias l'une de l'autre si l'une est un sigle extrait de l'autre.

Remarques

Le jeu de descripteurs de Soon *et al.* (2001) nécessite quelques précautions lorsqu'un modèle logisitique est utilisé plutôt qu'un arbre de décision. D'abord, les descripteurs non-binaires qui prennent des valeurs entre lesquelles il n'existe pas de hiérarchie doivent être binarisés.

Par exemple, l'accord en genre peut prendre les valeurs $\{Oui, Non, Inconnu\}$, il sera donc transformé en 3 descripteurs binaires différents : m et c s'accordent en genre, elles ne s'accordent pas en genre, le genre d'une des deux mentions est inconnu.

Ensuite, les conjonctions de descripteurs pertinentes doivent être explicitées pour recevoir un poids dédié. La distance entre les deux mentions devrait par exemple recevoir un poids différent selon que m est un pronom ou un nom propre.

3.2.3 Résultats

Soon *et al.* (2001) évaluent leur système sur les corpus MUC 6 et MUC 7 en suivant les méthodologies utilisées pour chacune des deux campagnes d'évaluation. Ils obtiennent des scores F1 de 62.6 % sur MUC 6 et de 60.4 % sur MUC 7. Ces scores sont parmi les meilleurs obtenus par les systèmes symboliques proposés lors des campagnes d'évaluation.

De plus, puisque l'algorithme d'apprentissage choisi (C5) construit un arbre de décision, le modèle obtenu est aussi bien interprétable qu'un système de règles définies manuellement. L'arbre de décision obtenu partir du corpus MUC 6 est reporté à la figure 3.1.

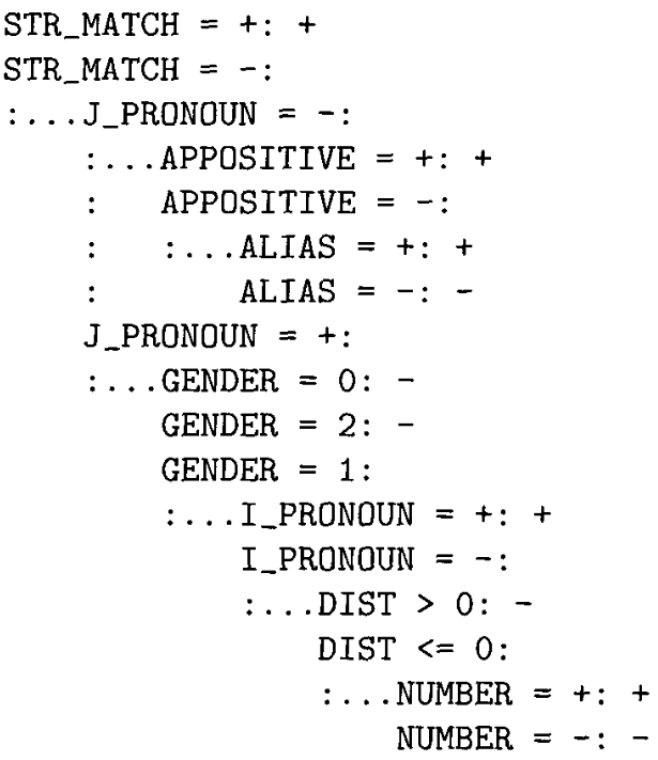

FIGURE 3.1 – Abre de décision obtenu par la méthode de Soon *et al.* (2001) sur MUC-6

3.3 Le modèle par ordonnement

3.3.1 Résolution des pronoms

Inspirés par Ravichandran *et al.* (2003), Denis et Baldridge (2007) proposent une modélisation alternative de la tâche restreinte à la résolution des pronoms anaphoriques. Au lieu d'apprendre une fonction de classement binaire, la fonction apprise est une fonction d'ordonnement [2], paramétrée non pas pour prédire la vraisemblance d'un lien d'antécédence, mais pour prédire l'antécédent le plus probable d'un pronom.

Concrètement, le modèle utilisé est un modèle logistique d'ordonnement qui apprend une distribution de probabilités sur les candidats antécédents sachant le pronom. Soient m le pronom à résoudre et $C = \{c_1, ..., c_n\}$ un ensemble de candidats antécédents. Soient Φ une fonction qui représente un pronom associé à un candidat antécédent par un vecteur dans $\mathbb{R}^p$ et w un vecteur de poids de $\mathbb{R}^p$. La distribution de probabilités sur les candidats antécédents de m est définie par :

$$P(c_i|m) = \frac{exp(w.\Phi(m, c_i))}{\sum_{j=1}^{n} exp(w.\Phi(m, c_j))}$$

Étant donnés $\{m^1, ..., m^l\}$ l'ensemble des pronoms anaphoriques du corpus d'apprentissage et $\{c^1, ..., c^l\}$ leurs antécédents de référence, la fonction objectif utilisée pour l'optimisation du vecteur de poids w est la suivante :

2. En anglais, *ranking*. Il s'agit d'une fonction qui, au sens large, hiérarchise un ensemble d'éléments en leur attribuant chacun un rang dans un classement.

$$J(w) = -\sum_{i=1}^{l} log(P(c^i|m^i))$$

Les corpus annotés en coréférence ne fournissent pas directement d'antécédents pour les pronoms, mais uniquement les chaînes de coréférence auxquelles ils appartiennent. La distance jouant un rôle important dans la résolution des pronoms, Denis et Baldridge (2007) font le choix de considérer comme antécédent de référence d'un pronom la mention coréférente la plus proche qui le précède.

Les mauvais candidats antécédents sont quant à eux l'ensemble des mentions apparaissant dans une fenêtre de quatre phrases autour de l'antécédent de référence. Cette fenêtre inclut la phrase dans laquelle figure l'antécédent de référence, la phrase précédente et les deux suivantes.

Avec ce modèle, la procédure d'inférence est identique à celle du classement *best-first*. Pour chaque pronom, on dispose d'un score par candidat antécédent et on choisit le candidat pour lequel le score le plus élevé a été prédit. Étant donné la localité des pronoms, les candidats antécédents ne comprennent que les mentions situées avant le pronom à résoudre qui figurent dans la même phrase ou l'une des trois phrases précédentes.

L'ordonnement est donc la façon la plus naturelle de modéliser la recherche du meilleur antécédent d'un pronom. Comme l'ont montré Denis et Baldridge (2007), c'est aussi la plus performante puisqu'elle apporte sur les corpus ACE un gain de 7.2 points d'exactitude par rapport à une modélisation par classement binaire proche de celle de Soon *et al.* (2001). Une illustration de la meilleure adéquation à la tâche d'un modèle d'ordonnement est donnée à l'annexe A.

3.3.2 Généralisation à la coréférence

Le modélisation par ordonnement de la résolution des pronoms peut être généralisée à la résolution des coréférences en général moyennant quelques aménagements.

Détection d'anaphoricité

D'abord, le modèle de résolution des pronoms doit impérativement prédire un antécédent pour chaque pronom anaphorique. A contrario, dans le cas de la résolution des coréférences, une mention peut être la première de sa chaîne de coréférence, et donc ne pas avoir d'antécédent.

Les modèles par classement binaire peuvent gérer ce cas facilement : aucun antécédent n'est prédit pour une mention si le classeur n'autorise un lien d'antécédence avec aucun des candidats antécédents. Les modèles par ordonnement nécessitent de déterminer – au préalable ou conjointement –, pour chaque mention, si elle introduit une nouvelle chaîne de coréférence, ou bien si elle doit intégrer une chaîne de coréférence déjà commencée, auquel cas elle sera abusivement dite *anaphorique*.

La tâche consistant à déterminer pour quelles mentions il faut prédire un antécédent, appellée détection d'anaphoricité, a été introduite par Ng et Cardie (2002a). Pour améliorer les performances de la résolution des coréférences par classement binaire, ils proposent de ne rechercher un antécédent que pour les mentions prédites comme anaphoriques par un arbre de décision dédié.

C'est essentiellement cette même méthode que Denis et Baldridge (2008) utilisent pour généraliser leur modèle de résolution des pronoms à la coréférence en général. Rah-

man et Ng (2009) proposent au contraire d'apprendre de manière conjointe à détecter l'anaphoricité d'une mention et à prédire, le cas échéant, son antécédent.

Pour cela, un pseudo-candidat antécédent ϵ est ajouté pour chaque mention à la liste des candidats antécédents. Si le modèle le préconise comme candidat antécédent, c'est que la mention n'est pas anaphorique. À la différence des vrais candidats antécédents, le pseudo-candidat n'est décrit que par les descripteurs propres à la mention considérée puisqu'elle n'est appariée à aucune mention précédente.

Candidats antécédents et antécédents de référence

Dans le modèle de Denis et Baldridge (2008), les candidats antécédents d'une mention sont à l'apprentissage constitués de toutes les mentions situées dans une fenêtre de deux phrases autour de l'antécédent de référence qui ne figurent pas dans la même chaîne de coréférence. Lors de l'évaluation, il s'agit de l'ensemble des mentions qui la précèdent.

L'antécédent de référence d'une mention non pronominale est la mention coréférente non-pronominale la plus proche et celui d'une mention pronominale la mention la plus proche, sans restriction de type. À défaut, c'est le pseudo-antécédent ϵ, marqueur de la non-anaphoricité de la mention.

Encourager le modèle à choisir un antécédent non pronominal permet d'éviter au maximum de se fonder sur des décisions d'antécédence trop peu informées. Une exception est faite pour les pronoms pour lesquels la localité est très importante.

Durrett *et al.* (2013) proposent une approche alternative. Les candidats antécédents d'une mention sont l'ensemble des mentions qui la précèdent, que ce soit pendant l'apprentissage ou pendant l'évaluation. De plus, au lieu d'utiliser une heuristique pour choisir un unique antécédent de référence au sein de la chaîne de coréférence de la mention, le choix de l'antécédent est considéré comme une variable cachée du modèle.

Soit C^i l'ensemble des mentions dans la même chaîne de coréférence que m^i qui la précèdent, ou $\{\epsilon\}$ s'il n'y en a aucune. La fonction objectif est alors définie comme suit :

$$J(w) = -\sum_{i=1}^{l} log(\sum_{c_i \in C^i} P(c^i|m^i))$$

L'approche de Durrett *et al.* (2013) permet d'utiliser au maximum l'information contenue dans le corpus d'apprentissage puisque la mise à jour des poids tient compte de l'ensemble des mentions qui précèdent celle en cours de traitement.

Considérer le choix de l'antécédent comme une variable cachée est alors impératif pour éviter les incohérences. Par exemple, si un même nom propre apparaît dix fois dans un document, chercher à maximiser uniquement la probabilité de la neuvième mention lors de la résolution de la dixième entraînerait une baisse de probabilité pour les huit autres corrects. Cette mise à jour conduirait alors probablement à une sur-valorisation de la distance au détriment de l'identité des chaînes de caractère.

Au contraire, en cherchant à maximiser la somme des probabilités de tous les antécédents corrects, les poids sont modifiés de manière à augmenter la probabilité de tous les candidats qui font partie de la même chaîne de coréférence et diminuer toutes celles des autres candidats. Le modèle favorise ainsi les descripteurs les plus discriminants, c'est-à-dire ceux particulièrement présents sur les bons candidats et peu sur les mauvais.

3.4 L'évolution des descripteurs

3.4.1 La spécialisation des poids

Denis et Baldridge (2008) soulignent également l'importance d'apprendre des poids différents pour bon nombre de descripteurs selon le type de la mention pour laquelle on cherche un antécédent. Ils définissent à cet effet cinq groupes de mentions : les pronoms des 1re et 2e personnes, ceux de la 3e personne, les noms propres, les descriptions définies, et les mentions restantes.

Cette différentiation se fonde sur des motivations linguistiques. Selon plusieurs théories (Ariel, 1988; Grosz *et al.*, 1995), les locuteurs attribueraient à chaque référent de discours une certaine saillance qui poserait des contraintes sur les formes linguistiques utilisables pour y référer. Typiquement, un pronom n'est utilisable pour reprendre un référent que si celui-ci est très saillant alors qu'un nom propre, beaucoup plus informatif, sera utilisé pour parler d'un individu peu saillant.

Apprendre des poids différents pour chaque catégorie de mentions permet donc, par exemple, de privilégier un antécédent proche pour un pronom de 3e et un peu plus éloigné pour un nom propre, ou encore de donner une grande importance à la correspondance des chaînes de caractères pour un nom propre et beaucoup moins pour un pronom.

Si Denis et Baldridge (2008) choisissent d'utiliser des modèles complètement différents selon le type de la mention pour laquelle on cherche un antécédent, ce n'est pas la seule manière d'aborder le problème. La spécialisation des poids peut également être obtenue en utilisant un schéma de conjonction des descripteurs approprié, dans lequel chaque descripteur atomique est conjugué avec le type de la mention traitée.

3.4.2 La lexicalisation des descripteurs

Alors que la plupart des jeux de descripteurs utilisés auparavant reposaient sur des généralisations linguistiques comme la notion de syntagme défini ou l'accord en genre et en nombre, Durrett et Klein (2013) proposent un jeu de descripteurs minimal fondé essentiellement sur les mots de la mention et de son contexte qui leur permet de battre les systèmes à l'état de l'art. Toutefois, pour lutter contre la dispersion des données, les mots sont remplacés par leur étiquette morpho-syntaxique lorsqu'ils apparaissent moins de 20 fois dans le corpus d'apprentissage.

Descripteurs unaires
Tête
Premier mot
Dernier mot
Mot précédent
Mot suivant
Longueur
Descripteurs binaires
Correspondance exacte
Correspondance des têtes
Distance en nombre de phrases
Distance en nombre de mentions

Tableau 3.1 – Jeu de descripteurs minimal proposé par Durrett et Klein (2013)

Ce jeu de descripteurs est reproduit dans la table 3.1. Chaque descripteur unaire est calculé aussi bien sur la mention à traiter que sur le candidat antécédent. C'est également ce même ensemble de descripteurs unaires qui est utilisé pour le pseudo-candidat ϵ, mais avec des poids spécifiques.

Ces descripteurs remplacent implicitement des indicateurs linguistiques plus fins. Par exemple, le premier mot de la mention indique (en anglais) s'il s'agit d'un syntagme défini ou non, et les mots qui précèdent et suivent immédiatement une mention permettent

d'inférer si celle-ci est en position sujet ou objet. Enfin, des descripteurs comme l'accord en genre et en nombre sont également naturellement inférables à partir des mots constituant les mentions.

Durrett et Klein (2013) montrent en plus que ces nouveaux descripteurs apportent un gain de performance par rapport aux descripteurs linguistiques utilisés dans les mêmes conditions. Ce gain de performance s'explique par le fait que les descripteurs lexicaux capturent, en plus des généralisations linguistiques, d'autres généralisations statistiques que l'on ne sait pas modéliser explicitement.

Le schéma de conjonction qu'ils utilisent permet notamment de spécialiser les poids conformément à l'idée de Denis et Baldridge (2008). En plus d'être utilisé individuellement, chaque descripteur est combiné avec le type de la mention. Le cas échéant, les descripteurs obtenus sont à leur tour combinés avec le type du candidat antécédent. Les différents types de mentions comprennent les noms propres, chaque pronom dans sa forme lemmatisée et les mentions restantes.

Durrett et Klein (2013) proposent en outre quelques descripteurs supplémentaires capables d'apporter un gain de performance additionnel. Il s'agit notamment du genre du document et de l'identité de locuteur entre les mentions, qui apportent une source d'information importante pour la résolution des déictiques.

Au final, Durrett et Klein (2013) parviennent à atteindre un score CoNLL de 60.13 sur le corpus de la campagne d'évaluation CoNLL 2011. Ce score est à comparer avec, par exemple, le système symbolique vainqueur de la campagne CoNLL 2011 (Lee *et al.*, 2011) qui obtient 56.65.

3.4.3 L'apprentissage de représentations

Partant du constat qu'il est difficile de déterminer manuellement quelles conjonctions de descripteurs sont pertinentes pour la tâche, Wiseman *et al.* (2015) proposent d'apprendre automatiquement au sein d'un réseau de neurones des représentations des paires mention-candidat antécédent utiles pour la résolution des coréférences.

Leurs descripteurs sont pour l'essentiel ceux de Durrett et Klein (2013), mais sans utiliser de schéma de conjonction. Notons $\Phi_a(m)$ le vecteur descripteur d'une mention utilisé pour la détection d'anaphoricité et $\Phi_p(m, c)$ le vecteur descripteur d'une mention et d'un candidat différent de ϵ.

Donnons-nous deux fonctions h_a et h_p capables de donner des représentations non-linéaires de Φ_p et Φ_a. Ces fonctions, respectivement paramétrées par les matrices et vecteurs de poids W_a et b_a d'une part, et W_p et b_p d'autre part, sont définies comme suit :

$$h_a(m) = tanh(W_a.\Phi_a(m) + b_a)$$
$$h_p(m, c) = tanh(W_p.\Phi_p(m, c) + b_p)$$

La fonction s associant un score à une mention et un candidat antécédent est paramétrée par les matrices et vecteurs U, u_0, V et v_0. Elle est définie de la manière suivante :

$$s(m, c) = \begin{cases} U.\begin{bmatrix} h_a(m) \\ h_p(m, c) \end{bmatrix} + u_0 & \text{if } c \neq \epsilon \\ V.h_a(m) + v_0 & \text{if } c = \epsilon \end{cases}$$

Wiseman *et al.* (2015) parviennent de cette manière à dépasser le système de Durrett et Klein (2013) de presque deux points sur la portion anglaise du corpus de la campagne d'évaluation CoNLL 2012.

Un pas supplémentaire vers l'adoption des méthodes standards d'apprentissage profond pour la résolution des coréférences est ensuite franchi par Clark et Manning (2016b), bien que cela ne constitue pas l'essentiel de leur contribution.

Comme celui de Wiseman *et al.* (2015), leur modèle apprend des représentations non-linéaires d'une paire mention-candidat antécédent. Ces représentations sont cependant construites à partir de représentations distribuées de mots explicites et préentraînées avec des méthodes distributionnelles, ce qui améliore la capacité de généralisation du modèle par partage de la connaissance statistique entre les mots, et grâce à la pertinence de l'initialisation des vecteurs.

3.5 Conclusion

Le modèle mention-mention est une approche simple et efficace de la résolution des coréférences. Dans sa version moderne, une fonction d'ordonnement prédit pour chaque mention si elle introduit une nouvelle chaîne de coréférence, ou son antécédent le plus probable dans le cas contraire.

Grâce à sa simplicité, le modèle mention-mention a été utilisé pour montrer aussi bien la supériorité de l'ordonnement sur le classement, que l'intérêt de la lexicalisation des descripteurs et de l'apprentissage des représentations. Ces innovations lui sont cependant orthogonales et ont été ensuite appliquées avec succès à d'autres approches de la résolution des coréférences, comme le modèle mention-entité que nous décrirons au chapitre suivant.

Chapitre 4

Modèle mention-entité

Si le modèle mention-mention est attrayant de par sa simplicité, il présente plusieurs défauts que nous décrirons à la section 4.1. Nous introduirons ensuite le modèle mention-entité, une approche alternative de la résolution des coréférences qui a émergé dès 2004.

4.1 Motivations

4.1.1 Motivations cognitives

Le premier type de motivations pour aller au-delà du modèle mention-mention est de l'ordre de la plausibilité cognitive. Cette approche présentée au chapitre 3 cherche dans un premier temps un antécédent textuel pour chaque mention d'un document et ne construit les chaînes de coréférence par fermeture transitive qu'en dernier lieu.

Pourtant, déterminer la chaîne de coréférence à laquelle une mention appartient consiste à la lier non pas à un antécédent textuel, mais à un référent de discours. Cette distinction est particulièrement saillante dans le cas des mentions non anaphoriques, pour lesquelles la notion d'antécédent textuel n'existe pas.

Comme ce processus fait partie intégrante de l'interprétation, il est en plus vraisemblable qu'il ait lieu en continu plutôt que d'un bloc à la fin de la réception de l'énoncé. Il mobilise en plus probablement toutes les connaissances accumulées sur les différents référents au travers de leurs différentes mentions, plutôt que celles apportées par un unique antécédent textuel.

4.1.2 Motivations empiriques

Chercher un antécédent textuel aux mentions indépendamment les unes des autres et sans tirer parti des informations sur les référents obtenues à partir des prédictions précédentes n'est pas sans conséquences empiriques.

Pour illustrer les conséquences pratiques de la trop grande simplicité du modèle mention-mention, considérons, à la suite de Luo *et al.* (2004), un document qui contiendrait dans cet ordre les trois mentions suivantes : *M. Clinton*, *Clinton* et *Elle*.

La mention *M. Clinton* introduira dans un premier temps une nouvelle chaîne de coréférence. De par leur grande similarité, un lien d'antécédence sera vraisemblablement prédit entre cette mention et *Clinton*. Enfin, *Elle*, que l'on supposera non loin de *Clinton*, sera naturellement liée à cette dernière mention.

Après fermeture transitive, on obtiendra donc une unique chaîne de coréférence contenant les trois mentions. Cette chaîne est évidemment incohérente et donc hautement improbable puisque *M. Clinton* et *Elle* sont incompatibles en genre.

Le problème est que le modèle a pris chaque fois des décisions locales sans tenir compte des liens d'antécédence déjà prédits. Autrement dit, pour faire le lien avec les motivations cognitives, on a cherché comme antécédent de *Elle* une mention, au lieu de chercher une entité. En effet, lors de la résolution du pronom, on savait déjà grâce au premier lien prédit que la mention *Clinton* se rapportait à un individu de genre masculin et donc que *Elle* ne pouvait pas appartenir à la même chaîne de coréférence.

4.2 Modélisation

4.2.1 Présentation

Pour pallier cette faiblesse intrinsèque du modèle mention-mention, Luo *et al.* (2004) proposent une modélisation alternative de la tâche capable de prendre en compte les liens d'antécédence déjà prédits et de mettre continuellement en commun les informations apportées par les différentes mentions d'un référent de discours.

La modélisation proposée, largement reprise par la suite, considère toujours les mentions une par une de gauche à droite, mais les confronte non pas à celles qui les précèdent, mais aux différents référents de discours déjà introduits. Ceux-ci sont représentés par des chaînes de coréférence partielles contenant les mentions déjà liées à ce stade du traitement du document.

En conséquence, les candidats antécédents auxquels le modèle statistique attribue des scores ne sont plus des mentions, mais des chaînes de coréférence partielles correspondant à des référents de discours, c'est-à-dire le plus souvent des entités. Nous appellerons donc cette approche modèle mention-entité.

Si Luo *et al.* (2004) utilisent un classeur binaire pour décider si une mention appartient à une chaîne de coréférence partielle, Rahman et Ng (2009) transposent naturellement au modèle mention-entité la modélisation par ordonnement apparue pour le modèle mention-mention.

4.2.2 Inférence

Luo *et al.* (2004) modélisent l'inférence dans le cadre d'un modèle mention-entité par un arbre de Bell (voir figure 4.1). Représenter l'inférence sous cette forme permet de la concevoir facilement comme un problème de recherche. Chaque feuille représente une partition de l'ensemble des mentions et reçoit comme score le produit des probabilités des différentes décisions prises sur le chemin qui la relie à la racine. La feuille de plus haut score contient le meilleur partitionnement.

Malheureusement, le nombre de nœuds dans l'arbre de recherche est pour n mentions de $\sum_{k=1}^{n} B(k)$, où $B(k)$ est le k-ième nombre de Bell, c'est-à-dire le nombre de partitions d'un ensemble de k éléments. Aussi, trouver le partitionnement de meilleur score est un problème NP-difficile et des stratégies d'approximation doivent être utilisées.

Luo *et al.* (2004) proposent d'utiliser une recherche en faisceaux en plus de différentes heuristiques d'élagage. En revanche, une simple recherche gloutonne est utilisée avec le réseau de neurones de Wiseman *et al.* (2016). Comme c'est souvent le cas avec ce type de modèles, cela suffit pour obtenir des résultats satisfaisants.

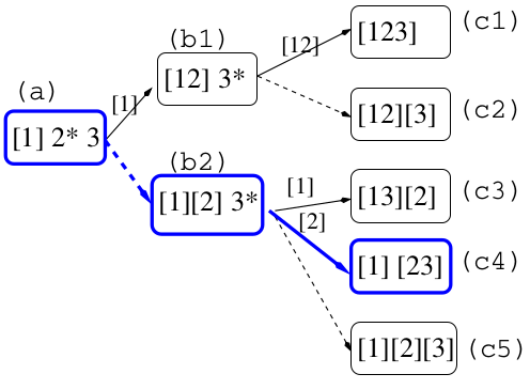

FIGURE 4.1 – Représentation par un arbre de Bell de la procédure d'inférence pour trois mentions. Les nombres entres crochets dénotent des chaînes de coréférence partielles. La mention à résoudre à une étape donnée est suffixée par *. Les flèches en pointillés marquent le choix de commencer une nouvelle chaîne de coréférence, tandis que celles en trait plein sont étiquetées par la chaîne de coréférence partielle choisie pour la mention. (Emprunté à Luo *et al.* (2004))

4.2.3 Oracle

L'utilisation de chaînes de coréférence partielles en entrée du modèle statistique pose également la question de la construction des exemples d'apprentissage. Les candidats antécédents d'une mention sont les différentes chaînes de coréférence partielles constituées à cette étape du processus. Comme dans d'autres tâches de TAL, l'oracle qui guide l'apprentissage en construisant les exemples peut être soit statique, soit dynamique.

Dans le premier cas, les exemples ne dépendent pas des prédictions du modèle et peuvent donc être construits avant le début de l'apprentissage. Les chaînes de coréférence partielles utilisées comme candidats antécédents d'une mention sont construites à partir des chaînes annotées. C'est la méthode employée par Luo *et al.* (2004) et préconisée par Wiseman *et al.* (2016).

Alternativement, on peut souhaiter exposer le modèle à ses propres prédictions lors de l'apprentissage. L'oracle doit alors être dynamique et décider quand utiliser les chaînes de coréférence partielles inférées à partir des annotations, et quand utiliser celles prédites par le modèle. Wiseman *et al.* (2016) rapportent qu'un oracle dynamique ne leur a apporté qu'un gain de performance négligeable qu'ils ont choisi de sacrifier à une réduction du temps d'apprentissage. [1]

4.3 Descripteurs

Enfin, un modèle mention-entité exige une méthode pour combiner les informations fournies par les différentes mentions d'une chaîne de coréférence partielle. Si les premières solutions proposées étaient des heuristiques, celles-ci ont ensuite fait place à l'apprentissage de représentations des chaînes de coréférence partielles.

1. Avec un oracle statique, les représentations des chaînes de coréférence partielles peuvent être précalculées en une seule fois au début du traitement de chaque document.

4.3.1 Approche par heuristiques

Luo *et al.* (2004) proposent de décrire une association entre une mention et une chaîne de coréférence partielle par quelques heuristiques appliquées aux descripteurs habituellement utilisés par les modèles mention-mention. Par exemple, les descripteurs binaires représentant la similarité entre deux mentions (correspondance exacte, alias, etc.) sont évalués à 1 dès lors qu'au moins une des mentions de la chaîne de coréférénce partielle vérifie la propriété de similarité avec la mention à résoudre, et la distance entre les deux mentions est transformée en le minimum des distances entre la mention à résoudre et celles de la chaîne de coréférence partielle.

L'approche de Rahman et Ng (2009) est un peu plus systématique. Ils appliquent les prédicats logiques AUCUN, LA PLUPART OUI, LA PLUPART NON et TOUS à chacun des descripteurs locaux des modèles mention-mention. Ces prédicats permettent de calculer des descripteurs binaires comme *Toutes les mentions de la chaîne de coréférence partielle s'accordent en genre avec la mention à résoudre* ou encore *Au moins une mention de la chaîne de coréférence partielle est un alias de la mention à résoudre*. Les descripteurs locaux non binaires sont binarisés avant l'application des prédicats logiques.

4.3.2 Approche par apprentissage de représentations

Constatant qu'il était difficile de trouver par heuristiques de bons descripteurs globaux, Wiseman *et al.* (2016) proposent de laisser le modèle statistique apprendre de bonnes représentations des chaînes de coréférence partielles. Concrètement, une chaîne de coréférence partielle est représentée par l'état caché d'un réseau de neurones récurrent qui a consommé les représentations des mentions qui la composent.

Wiseman *et al.* (2016) proposent comme ligne de base de moyenner les représentations des différentes mentions de la chaîne de coréférence partielle considérée, et trouvent qu'il est plus avantageux d'utiliser un réseau de neurones pour apprendre automatiquement à déterminer quelles informations retenir des différentes mentions.

4.4 Résultats

Si sa pertinence ne fait aucun doute d'un point de vue théorique, le modèle mention-entité a toujours conduit à des gains de performance très modérés. Par exemple, Wiseman *et al.* (2016) n'obtiennent qu'un gain d'environ 0.8 points CoNLL par rapport à leur modèle mention-mention (Wiseman *et al.*, 2015). Ce maigre gain de performance explique probablement que le modèle mention-entité n'ait jamais réussi à enterrer le modèle mention-mention.

Chapitre 5

Modèle neuronal de bout en bout

Le modèle neuronal de bout en bout proposé par Lee *et al.* (2017), fondamentalement mention-mention, introduit deux innovations principales : l'identification conjointe des mentions et la représentation des mentions par un réseau récurrent additionné d'un mécanisme d'attention.

Ces deux caractéristiques font que le seul prétraitement des documents nécessaire est leur tokenisation [1]. On évite ainsi la longue chaîne de prétraitement de l'approche classique susceptible de créer du bruit pour la résolution des coréférences elle-même.

5.1 Modélisation

5.1.1 Aperçu

Plutôt que d'identifier les empans de texte susceptibles d'être des mentions par étiquetage morphosyntaxique, analyse syntaxique et détection d'entités nommées, Lee *et al.* (2017) choisissent de traiter chaque empan des phrases d'un document comme une mention potentielle et de laisser la tâche de résolution des coréférences filtrer elle-même ce très large ensemble de candidats mentions.

Comme dans l'approche classique, un modèle d'ordonnement confronte chaque candidat mention à ceux qui le précèdent pour déterminer d'éventuels liens d'antécédence. La prédiction de l'antécédent nul ϵ comme antécédent d'un candidat mention signifie soit que celui-ci commence une nouvelle chaîne de coréférence, soit que l'empan n'est pas une mention.

Une telle ambiguïté apparaissait déjà dans l'approche classique. La détection des mentions préalable n'étant pas parfaite, on incitait lors de l'apprentissage le modèle à prédire ϵ comme antécédent des mentions prédites ne faisant pas partie de celles de référence, soit parce qu'erronées, soit parce que singletons.

L'ambiguïté n'est levée que lors de la fermeture transitive des liens d'antécédence. Si un candidat mention à antécédent ϵ est lui-même antécédent d'une autre mention, c'est qu'il commence une nouvelle chaîne de coréférence. Dans le cas contraire, il s'agit soit d'une mention singleton, soit d'une mention erronée.

1. Le modèle tire également profit d'une caractérisation du genre du document et d'informations sur le locuteur dans les dialogues, mais il s'agit de métadonnées naturellement disponibles, et pas d'informations obtenues par prétraitement.

5.1.2 Complexité algorithmique

En faisant l'hypothèse raisonnable que la longueur d'un document est indépendante de la longueur de ses phrases, le nombre d'empans de texte susceptibles d'être des mentions est en $O(T)$, avec T le nombre de mots du document.

Le nombre de couples d'empans que le modèle d'ordonnement doit traiter est donc en principe en $O(T^2)$. Pour borner linéairement la complexité de leur modèle, Lee *et al.* (2017) limitent le nombre de candidats antécédents considérés pour chaque candidat mention aux 250 plus proches, aussi bien à l'apprentissage que lors de l'évaluation.

Derrière cette complexité linéaire en la longueur du document se cachent cependant d'importants coefficients. En particulier, le nombre d'empans dans une phrase augmente quadratiquement avec sa longueur, et le nombre de paires d'empans à traiter est proportionnel à la somme des nombres d'empans dans les phrases du document.

Aussi, pour encore réduire les temps de calcul, Lee *et al.* (2017) limitent la longueur des empans considérés à 10 et introduisent un important élagage des candidats mentions avant comparaison entre eux de façon à ce qu'il en reste exactement $\lambda.T$ avec $\lambda = 0.4$. Dans la pratique, les temps de calcul restent cependant encore conséquents.

5.1.3 Fonctions de score

Scores et élagage des empans

Un score $s_m(m)$ est calculé pour chaque candidat mention m à partir de sa représentation $h_m(m)$ et d'un réseau de neurones à propagation avant FNN_m dédié :

$$s_m(m) = FNN_m(h_m(m))$$

Les $\lambda.T$ candidats mentions qui seront utilisés pour la recherche d'antécédent sont choisis parmi ceux dont les scores sont les plus élevés de manière de manière à éviter les croisements de mentions, proscrits par le schéma d'annotation d'Ontonotes.

Plus précisément, si on note (i, j) un candidat mention commençant au mot d'indice i et se terminant à celui d'indice j, les candidats mentions sont sélectionnés par ordre décroissant de score de telle sorte qu'il n'existe pas de paire de candidats mentions retenus (i, j) et (i', j') telle que $i < i' \leq j < j'$.

Scores d'antécédence

La fonction de score d'antécédence s est définie de la manière suivante :

$$s(m, c) = \begin{cases} 0 & \text{si } c = \epsilon \\ s_m(m) + s_m(c) + s_a(m, c) & \text{sinon} \end{cases}$$

où $s_a(m, c)$ est un score calculé par application d'un réseau de neurones à propagation avant FNN_a à la représentation $h_a(m, c)$ de l'association entre le candidat mention m et le candidat antécédent c.

Fonction objectif

Comme dans Durrett et Klein (2013), la fonction objectif est :

$$J(w) = -\sum_{i=1}^{l} log(\sum_{c_i \in C^i} P(c^i|m^i))$$

avec C^i l'ensemble des mentions dans la même chaîne de coréférence que m^i qui la précèdent, ou $\{\epsilon\}$ s'il n'y en a aucune.

Selon la méthode usuelle, on obtient les probabilités $P(c|m)$ à partir des scores $s(m, c)$ par application de la fonction softmax :

$$P(c|m) = \frac{exp(s(m, c))}{\sum_{c' \in C} exp(s(m, c'))}$$

avec C l'ensemble des candidats mentions non élagués précédant m.

5.2 Représentations

5.2.1 Représentations des mots

À la base des représentations des associations entre une mention et un candidat antécédent se trouvent des représentations distribuées (Pennington *et al.*, 2014; Turian *et al.*, 2010) de chaque mot du document à traiter que l'on notera $(w)_{1 \leq t \leq T}$. Lorsque le mot est inconnu, le vecteur nul est utilisé.

Ces représentations sont complétées par des représentations apprises à partir des différents caractères qui composent les mots, supposées fournir des compléments d'information morphologique et permettre de calculer des similarités même sur la base de mots inconnus.

Concrètement, des représentations distribuées de chaque caractère sont apprises et servent d'entrée à des réseaux convolutifs avec collecte par maximum de taille 50 chacun, avec des fenêtres respectives de 3, 4 et 5 caractères. Pour chaque mot, les sorties des différents réseaux sont concaténées entre elles ainsi qu'avec les vecteurs de mot pour former un vecteur noté x_i qui sera utilisé par la suite pour représenter le mot.

5.2.2 Représentation des candidats mentions

La principale difficulté pour représenter les mentions vient du fait qu'elles sont de taille variable. L'approche historique, proposée par Durrett et Klein (2013) et reprise par la suite dans les premiers modèles neuronaux, consiste à n'utiliser l'information que d'un nombre fixe de mots de la mention et de son contexte réputés être informatifs.

L'approche proposée par Lee *et al.* (2017) est assez différente puisque, grâce à un réseau récurrent bidirectionnel, la représentation d'une mention comprend non seulement tous ses mots, mais également tous ceux de la phrase qui la contient.

À partir des représentations des mots du document, une nouvelle suite de vecteurs notée $(h)_{1 \leq t \leq T}$, est obtenue par application d'un LSTM (Hochreiter et Schmidhuber, 1997) bidirectionnel (Schuster et Paliwal, 1997) sur chacune des phrases du document. Ces nouveaux vecteurs sont des représentations des mots pris dans leur contexte.

À ce stade, nous avons encore un vecteur par mot du document. Pour représenter une mention par un vecteur de dimension fixe, on concatène le vecteur correspondant à son premier mot et celui correspondant à son dernier mot. La base de la représentation d'une mention (i, j) est donc le vecteur $[h_i; h_j]$.

Lee *et al.* (2017) proposent d'utiliser en sus un vecteur d'attention c_i^j calculé de la manière suivante :

$$\alpha_{t,(i,j)} = \frac{exp(FNN_\alpha(h_t))}{\sum_{t'=i}^{j} exp(FNN_\alpha(h_{t'}))}$$

$$c_i^j = \sum_{t=i}^{j} \alpha_{t,(i,j)}.x_t$$

Ce calcul repose sur FNN_α, un réseau à propagation avant dédié, capable de mesurer la pertinence de chacun des mots de la mention pour la tâche à résoudre à partir des vecteurs récurrents encodant les mots dans leur contexte.

Enfin, la taille de la mention, sous la forme d'une représentation distribuée apprise, est ajoutée à la représentation d'une mention par concaténation aux côtés du vecteur d'attention. En notant $\phi(j-i)$ ce vecteur encodant la taille de la mention, on obtient comme représentation de la mention $m = (i, j)$ le vecteur

$$g_m = [h_i; h_j; c_i^j; \phi(j-i)]$$

5.2.3 Représentation des associations mention-candidat antécédent

Soient une mention m et un candidat antécédent a représentés respectivement par les vecteurs g_m et g_a. La représentation de l'association entre la mention et le candidat antécédent est donnée par

$$[g_m; g_a; g_a \circ g_m; \phi(m, a)]$$

où $\circ$ représente le produit composante par composante et $\phi(m, a)$ est un vecteur contenant des représentations distribuées de l'identité de locuteur entre les deux mentions, la distance les séparant en nombre de mots et du genre du document tel que défini par Ontonotes. La distance est empaquetée [2] dans les intervalles 1-1, 2-2, 3-3, 4-4, 5-7, 8-15, 16-31, 32-63, 64+ avant sélection du vecteur la représentant.

5.2.4 Discussion

Mécanisme d'attention

À l'heure actuelle, les réseaux récurrents ne permettent pas de capturer n'importe quelle généralisation mettant en jeu des dépendances à des distances quelconques. Autrement dit, même si l'intégralité des mots de la mention (i, j) est utilisée pour calculer $[h_i; h_j]$, il n'est pas garanti que le modèle soit capable d'encoder la totalité de l'information pertinente dans ce vecteur. C'est pour tenter de pallier ce genre d'inconvénients dans les modèles de traduction neuronale que les mécanismes d'attention ont été conçus (Bahdanau *et al.*, 2014).

Lee *et al.* (2017) proposent pour leur part une interprétation originale de leur mécanisme d'attention. Les coefficients α attribués à chaque mot de la mention représentent l'importance du mot pour la prédiction de la coréférence. En forçant un peu le trait, on peut voir le mot ayant le coefficient le plus important comme étant la tête de la mention.

2. C'est comme ça que nous avons traduit l'anglais *to bin*.

L'analyse de ces coefficients montrent que cette tête de mention obtenue indirectement par le mécanisme d'attention est fortement corrélée avec les têtes syntaxiques, intensivement utilisées dans les systèmes de résolution des coréférence précédents.

Similarité

On remarque la présence du produit de vecteurs composante par composante comme mécanisme de similarité. Ce vecteur produit est donné en entrée à un réseau de neurones à propagation avant qui peut apprendre à distinguer des autres les composantes dont la similarité est caractéristique.

Le modèle peut ainsi en principe différencier la similarité entre les caractères des mentions, celle de leurs têtes ou encore celle de leurs catégories morphosyntaxiques et celle de leur sémantique si ce genre de distinction est appris.

5.3 Résultats

La figure 5.1 reproduit le tableau comparatif des performances des modèles de résolution des coréférences récents proposé par Lee *et al.* (2017). L'évaluation est faite sur le corpus de test de la portion anglaise du corpus CoNLL 2012.

	MUC			B^3			$CEAF_{\phi_4}$			
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Avg. F1
Our model (ensemble)	**81.2**	**73.6**	**77.2**	**72.3**	**61.7**	**66.6**	**65.2**	**60.2**	**62.6**	**68.8**
Our model (single)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
Clark and Manning (2016a)	79.2	70.4	74.6	69.9	58.0	63.4	63.5	55.5	59.2	65.7
Clark and Manning (2016b)	79.9	69.3	74.2	71.0	56.5	63.0	63.8	54.3	58.7	65.3
Wiseman et al. (2016)	77.5	69.8	73.4	66.8	57.0	61.5	62.1	53.9	57.7	64.2
Wiseman et al. (2015)	76.2	69.3	72.6	66.2	55.8	60.5	59.4	54.9	57.1	63.4
Clark and Manning (2015)	76.1	69.4	72.6	65.6	56.0	60.4	59.4	53.0	56.0	63.0
Martschat and Strube (2015)	76.7	68.1	72.2	66.1	54.2	59.6	59.5	52.3	55.7	62.5
Durrett and Klein (2014)	72.6	69.9	71.2	61.2	56.4	58.7	56.2	54.2	55.2	61.7
Björkelund and Kuhn (2014)	74.3	67.5	70.7	62.7	55.0	58.6	59.4	52.3	55.6	61.6
Durrett and Klein (2013)	72.9	65.9	69.2	63.6	52.5	57.5	54.3	54.4	54.3	60.3

FIGURE 5.1 – Tableau comparatif des performances des modèles récents (Emprunté à Lee *et al.* (2017))

Le modèle de Lee *et al.* (2017) surpasse celui de Clark et Manning (2016a) de 1.5 points CoNLL avec un seul modèle et de 3.1 points avec un ensemble de 5 modèles [3]. Notons en outre que le mécanisme d'attention apporte un gain de 1.3 points CoNLL sur le corpus de développement.

3. Combiner les prédictions de plusieurs modèles avec des paramètres initiaux différents permet de mieux généraliser.

Chapitre 6

Coréférence et syntaxe

6.1 Limites des approches actuelles

6.1.1 De l'approche classique au modèle de bout en bout

Comme nous l'avons vu au chapitre 2, la notion de mention textuelle implique de décider quels sont les modifieurs qui doivent être inclus. Rappelons que l'approche classique, mise en œuvre notamment dans Ontonotes, est d'inclure dans les mentions la totalité des modifieurs de leurs têtes syntaxico-sémantiques.

Prédire ces mentions avec leur délimitation exacte est une tâche éminemment syntaxique qui, en toute logique, est confiée dans l'approche traditionnelle de la résolution des coréférences à un analyseur syntaxique. En effet, des candidats mentions sont extraits des arbres syntaxiques des phrases à l'aide de quelques heuristiques en amont de la résolution des coréférences.

Mais comme l'ont fait remarquer Lee *et al.* (2017), les détecteurs de mentions symboliques traditionnels sont loin d'être parfaits. Ils souffrent notamment d'une limitation du rappel qui a des répercussions aussi bien lors des prédictions qu'à l'apprentissage. En effet, dans la mesure où, dans la majorité des corpus, les singletons ne sont pas annotés, des mentions prédites doivent impérativement être utilisées pour l'apprentissage à la fois de la détection d'anaphoricité et de la résolution des coréférences, et *a fortiori* dans les modèles conjoints.

Lee *et al.* (2017) citent un taux de plus de 9 % de mentions annotées ignorées à l'apprentissage faute d'être correctement détectées dans le système de Clark et Manning (2016a). Aussi, ils attribuent la majorité du gain de performance apporté par le modèle de bout en bout au fait de n'ignorer que moins de 2 % des mentions annotées, à savoir celles qui comprennent plus de 10 mots.

Le modèle de bout en bout introduit cependant par un effet de bord encore plus profondément la syntaxe au cœur même de la tâche de résolution des coréférences. En effet, le modèle de résolution des coréférences appris doit à la fois prédire les relations de coréférence à proprement parler et les délimitations des mentions, définies quelque peu arbitrairement par une règle syntaxique. Ce choix introduit malheureusement de nouveaux problèmes.

6.1.2 Problème de la longueur des mentions

D'abord, cette nouvelle approche vient avec une complexité intrinsèque supérieure puisque chaque empan du texte doit être considéré et comparé aux autres. Au contraire, les candidats mentions sont dans l'approche classique largement filtrés en amont ce qui limite la combinatoire.

Dans la pratique, pour que les temps de calcul restent raisonnables, Lee *et al.* (2017) sont même contraints d'ignorer complètement les mentions de longueur supérieure à 10. De cette manière, ils ne laissent donc aucune chance au modèle de faire son travail de résolution des coréférences sur ces mentions-là ce qui, en plus de pénaliser les performances, peut conduire à des résultats surprenants.

Le choix des projections maximales comme délimitation des mentions implique qu'elles peuvent en principe être de longueur arbitraire puisqu'on peut rajouter autant de propositions relatives qu'on souhaite à un syntagme nominal. La solution de Lee *et al.* (2017) n'est donc pas satisfaisante d'un point de vue théorique.

En (1) figure un exemple d'erreur liée à la longueur des mentions. La mention en italique comprend 15 mots et ne peut donc pas être prédite.

(1) One of the two honorable guests in the studio is *Professor Zhou Hanhua from the Institute of Law of the Chinese Academy of Social Sciences.*

6.1.3 Difficulté à prédire les projections maximales

Le problème de l'identification des mentions va cependant encore au-delà dans le modèle de bout en bout. Fréquemment, le modèle s'avère incapable de prédire correctement la mention en tant que syntagme maximal. Même si elle est de longueur inférieure à 10, il échoue à prédire sa délimitation exacte.

En (2), le modèle rattache le syntagme propositionnel *from this case* à *many beneficial experiences*, en conséquence de quoi il prédit la mention *many beneficial experiences from this case.*

(2) I think in fact, we can sum up *many beneficial experiences* from this case and actually extend them to other areas.

L'arbre syntaxique de la phrase nous indique cependant que *from this case* est un modifieur du verbe *sum up*, et pas de *many beneficial experiences.* C'est un très bon exemple du fait qu'on exige d'un bon modèle de résolution des coréférences de faire en prime le travail d'un analyseur syntaxique.

L'exemple (3) illustre quant à lui la difficulté du modèle à prédire des syntagmes maximaux, alors même qu'il n'y a guère d'ambiguïté. En effet, il prédit *Europe* au lieu de *Europe, where consumption of Brazilian juice has fallen.*

(3) The price cut, one analyst said, appeared to be aimed even more at *Europe, where consumption of Brazilian juice has fallen.*

Pour donner un ordre d'idée de l'importance quantitative du problème, l'analyseur d'erreurs de Berkeley (Kummerfeld et Klein, 2013) indique que mieux prédire les bornes des mentions pourrait apporter un gain de performance d'au moins 1.8 points CoNLL.

6.2 Apprendre la syntaxe conjointement

6.2.1 Motivations

Comme nous l'avons vu à la section précédente, l'architecture de Lee *et al.* (2017) exige du modèle de résolution des coréférences d'apprendre à faire en plus office d'analyseur syntaxique. Cet apprentissage ne peut cependant se faire dans leur approche qu'avec une supervision très indirecte et peu de données.

En effet, le modèle ne fait pas de différence entre les nombreux empans qui n'ont aucune chance de faire partie d'une chaîne de coréférence, et une mention singleton qui en commence et termine une simultanément. L'apprentissage de la définition des mentions n'est donc fait qu'implicitement, parce que c'est la seule manière de généraliser sur les liens de coréférence entre empans eux-mêmes.

D'autre part, la longueur des mentions dans Ontonotes suit grossièrement une loi de Zipf. Presque 60 % des mentions sont de longueur 1, presque 20 % de longueur 2, et seulement un peu plus de 7 % des mentions ont une longueur strictement supérieure à 5. La quantité de mentions à la syntaxe complexe, qui pourraient permettre d'apprendre au modèle à faire un minimum d'analyse syntaxique, est donc extrêmement limitée.

Nous avons donc tenté d'ajouter une supervision syntaxique dans le modèle de bout en bout par le biais de l'apprentissage multitâche. Dans les sections suivantes, nous présenterons d'abord une première approche, que nous avons commencé à mettre en œuvre avant de nous tourner vers une toute autre direction, puis une approche alternative plus systématique que nous avons également envisagée.

6.2.2 Première approche

Notre première approche a été de forcer notre modèle de résolution des coréférences à apprendre en plus à prédire si un empan est susceptible d'être une mention, singleton ou non. Conformément aux règles d'annotation, les exemples positifs sont pour cette tâche les empans correspondant aux projections maximales de chacun des mots des phrases, additionnés des verbes. Les empans restants sont des exemples négatifs, soit parce qu'ils ne correspondent pas à des constituants, soit parce qu'ils n'incluent pas tous les modifieurs de leur tête non verbale.

Le corpus Ontonotes fournissant en plus des chaînes de coréférence une couche d'analyse syntaxique en constituants, les projections maximales peuvent être déterminées facilement à partir des arbres syntaxiques des phrases. Il faut néanmoins pour cela identifier préalablement les têtes de chacun des constituants, ce que nous avons fait à l'aide d'une variante de l'algorithme de recherche des têtes implémenté dans *Stanford CoreNLP* (Manning *et al.*, 2014), lui-même basé sur la proposition de Collins (1999).

La seule modification que nous avons effectuée est de considérer les conjonctions de coordination comme les têtes des syntagmes coordonnés plutôt que le premier conjoint. De cette manière, aussi bien chacun des conjoints que le syntagme coordonné correspondent à des projections maximales et peuvent être des mentions.

La nouvelle tâche peut être introduite dans le modèle par ajout d'un classeur binaire sous la forme d'un réseau à propagation avant, avec une couche cachée ou non. Ce réseau prend en entrée les représentations des mentions et partage donc avec la tâche de résolution des coréférences les paramètres du BLSTM et les représentations distribuées des longueurs de mentions.

6.2.3 Un analyseur syntaxique conjoint

Si l'on pousse le raisonnement jusqu'au bout, il pourrait également être intéressant d'expérimenter un modèle conjoint d'analyse syntaxique et de résolution des coréférences qui partagerait entre les deux tâches les paramètres du réseau récurrent chargé d'encoder les phrases. Un analyseur en constituant comme celui de Stern *et al.* (2017) conviendrait par exemple très bien à une telle architecture puisque, comme le modèle de résolution des coréfrences de Lee *et al.* (2017), il s'appuie sur des représentations récurrentes d'empans de phrases.

Dans ce modèle, un empan (i, j) est représenté par le vecteur $[f_j - f_i; b_i - b_j]$ où f_i et b_i sont les sorties pour le mot i de deux LSTM qui consomment la phrase respectivement du début à la fin et de la fin au début. Ces représentations sont un peu différentes de celles utilisées par Lee *et al.* (2017), mais Cross et Huang (2016) suggèrent que ce choix n'est pas crucial. Dans leur cas, la concaténation de vecteurs BLSTM correspondant aux bornes de l'empan se comporte aussi bien.

Stern *et al.* (2017) représentent un arbre syntaxique T par un ensemble de constituants associés à leur étiquette. Soit, en notant respectivement l_t, i_t et j_t l'étiquette, l'indice de début et l'indice de fin du constituant t :

$$T = \{(l_t, (i_t, j_t))\}$$

Le score d'un tel arbre est donné par

$$s(T) = \sum_{(l_t,(i_t,j_t))\in T} (s_{label}(i, j, l) + s_{span}(i, j))$$

où les scores s_{span} et s_{label} sont calculés par des réseaux à propagation avant prenant en entrée la représentation d'un empan telle que décrite plus haut. s_{span} permet de consacrer un certain nombre de paramètres au calcul d'un score d'un empan sans considération d'étiquette, et s_{label} permet d'attribuer un score à chacune des étiquettes possibles.

Grâce à cette décomposition du score d'un arbre en la somme des scores de ses constituants, c'est-à-dire de ses nœuds, l'arbre de meilleur score peut être aisément déterminé par les méthodes classiques d'analyse syntaxique tabulaire basées sur la programmation dynamique.

6.3 Conclusion

Le modèle neuronal de bout en bout souffre d'un manque de connaissances syntaxiques qui se manifeste notamment dans sa difficulté à prédire les bornes exactes des mentions. Nous avons donc envisagé d'introduire dans le modèle une supervision syntaxique *via* l'apprentissage multitâche.

Cette méthode ne permet cependant pas de relâcher la contrainte sur la longueur des mentions puisqu'elle ne réduit pas la complexité intrinsèque du modèle. Les temps d'apprentissage s'en trouvent même naturellement encore augmentés.

À ce stade, nous nous sommes demandé s'il était vraiment pertinent de lier aussi étroitement l'analyse syntaxique et la résolution des coréférences et si nous ne pourrions pas au contraire les découpler complètement. C'est vers cette dernière option que nous nous sommes tournés ; elle fera donc l'objet des chapitres suivants.

Chapitre 7

Redéfinir la résolution des coréférences

Dans ce chapitre, nous reviendrons sur la notion de mention de référent de discours et les différentes façons possibles de les identifier dans un texte. Après avoir illustré la diversité des besoins des applications de la résolution des coréférences en matière de délimitation des mentions, nous proposerons une nouvelle définition de la tâche qui permet de la libérer de la contrainte d'une délimitation syntaxique choisie arbitrairement et d'évaluer les modèles uniquement sur leur capacité à prédire des relations de coréférence.

7.1 Notion de mention

Comme nous l'avons vu aux chapitres 2 et 4, la résolution des coréférences a pour mission d'établir des liens de coréférence entre mentions de référents de discours [1]. Pour mener à bien cette tâche, un moyen de les identifier dans les documents doit être trouvé pour pouvoir les manipuler.

La solution actuellement utilisée par les travaux en résolution des coréférences est celle adoptée pour le corpus Ontonotes, à savoir de représenter les mentions par des empans de texte correspondant aux projections maximales de leur tête syntaxico-sémantique. La décision d'inclure dans les mentions l'ensemble des modifieurs est cependant plutôt arbitraire et n'a rien à voir avec la résolution des coréférences elle-même. Cela montre que l'utilisation d'empans de texte pour représenter des référents de discours est loin d'aller de soi.

Comme deux mentions différentes ne peuvent pas avoir la même tête syntaxico-sémantique, il y a une bijection évidente entre l'ensemble des mentions et celui de leurs têtes. On pourrait donc envisager de fonder la résolution des coréférences directement sur les têtes des mentions plutôt que sur des syntagmes maximaux.

Notons qu'à notre connaissance, c'est la stratégie adoptée dans tous les travaux de résolution des coréférences événementielles pour les mentions verbales : plutôt que d'essayer de définir ce que pourrait être un syntagme maximal dans ce cas-là, les verbes sont considérés comme mentions à part entière, comme c'est également le cas dans les règles d'annotation d'Ontonotes.

1. Nous adoptons ici par souci de clarté les simplifications courantes en TAL : les mentions sont des expressions référentes et les liens à trouver sont des liens de coréférence, comme leur nom le laisse entendre. Le propos s'applique cependant à l'ensemble des cas de figures rencontrés en réalité.

Pour valider expérimentalement notre intuition, nous avons tenté de retrouver automatiquement les mentions d'Ontonotes, annotées sous forme de syntagmes maximaux, à partir des têtes prédites par l'algorithme mentionné à la section 6.2.2. Pour cela, nous avons utilisé une heuristique très simple qui suit au plus près les conventions d'annotation d'Ontonotes : si la tête est un verbe, la mention se résume à cette tête, sinon c'est le syntagme correspondant à sa projection maximale dans l'arbre syntaxique de la phrase.

Cette méthode permet de retrouver les bornes exactes d'environ 99 % des mentions du corpus en utilisant les arbres syntaxiques annotés. Les cas d'échec sont pour une part non négligeable dus à des erreurs d'annotation dans Ontonotes : certains modifieurs n'ont pas été inclus dans les mentions alors qu'ils auraient dû l'être selon les règles d'annotation.

Rappelons en outre que les corpus MUC et ACE avaient fait le judicieux choix de proposer une double délimitation des mentions : les syntagmes maximaux, avec tous les modifieurs, mais aussi des têtes de syntagme. Comme nous l'avons mentionné au chapitre 2, l'évaluation de la résolution des coréférences utilisait abondamment cette double délimitation à l'époque des campagnes d'évaluation qui ont vu naître ces corpus.

7.2 Applications

Le choix des syntagmes maximaux pour représenter les mentions nous semble d'autant plus arbitraire que les besoins des différentes applications de la résolution des coréférences en termes de délimitation des mentions sont très variés. Nous prendrons en guise d'illustration deux exemples importants, le résumé et la traduction automatiques.

Résumé automatique

Le résumé automatique extractif construit le résumé d'un document par sélection de certaines de ses phrases jugées particulièrement informatives, en général par une fonction de score. Durrett *et al.* (2016) remarquent que, sans précautions particulières, 60 % des phrases choisies contiennent des pronoms orphelins. Pour remédier à ce problème, ils proposent d'inclure des contraintes d'anaphoricité dans les scores attribués aux phrases par deux approches complémentaires toutes deux basées sur un système de résolution des coréférences.

La première méthode s'applique lorsqu'une chaîne de coréférence est prédite pour un pronom avec une forte probabilité par le système. Si une mention de l'entité n'a pas encore été incluse dans le résumé, le pronom est remplacé par la première mention de sa chaîne de coréférence. Dans le cas contraire, le système force l'inclusion dans le résumé de contenu supplémentaire, de façon à garantir que la référence du pronom soit claire.

Dans ce dernier cas, les syntagmes maximaux ne jouent aucun rôle et il suffit amplement d'identifier les différentes mentions par leurs têtes. En ce qui concerne le premier cas, une note de bas de page précise que la première mention de la chaîne de coréférence du pronom n'est en fait pas toujours utilisée telle quelle.

Si la tête de la mention est un nom propre, le pronom n'est remplacé que par la partie de la mention qui correspond au nom propre, plutôt que par le syntagme nominal entier. Mais même dans le cas d'un syntagme nominal standard, le remplacement aveugle d'un pronom par des syntagmes maximaux pourrait conduire à des incohérences. Les syntagmes maximaux ne sont donc pas les délimitations les plus pertinentes dans ce cas non plus.

De plus, du fait de la grande hétérogénéité des liens qui unissent deux mentions d'une même chaîne de coréférence, de nombreuses précautions doivent être prises avant de pro-

céder à un remplacement. Durrett *et al.* (2016) précisent qu'ils veillent à remplacer les adjectifs possessifs par un syntagme possessif.

Cet exemple illustre d'une part que les besoins en matière de résolution des coréférence sont très variés et que des empans de texte bien délimités ne sont pas toujours nécessaires, d'autre part que lorsque des délimitations sont requises, il ne s'agit non seulement pas nécessairement des syntagmes maximaux, mais qu'en plus beaucoup d'informations sur les mentions sont nécessaires pour pouvoir exploiter efficacement les chaînes de coréférence.

Typiquement, l'analyse syntaxique traditionnellement effectuée en amont de la résolution des coréférences permettrait ici d'identifier les noms propres [2] et de supprimer les propositions relatives des mentions. La gestion des possessifs pourrait elle éventuellement se contenter de quelques heuristiques sur les mots étant donnée la grande simplicité de l'anglais en la matière.

Traduction automatique

À la suite de Le Nagard et Koehn (2010), l'utilisation de la résolution des coréférences en traduction automatique a été abondamment décrite. Les principaux modèles de traduction automatique traduisent les phrases du document source une par une indépendamment les unes des autres, ce qui n'est pas sans conséquence sur la cohérence de la traduction.

À titre d'illustration, reprenons l'exemple de Le Nagard et Koehn (2010). *Google Translate* traduisait à l'époque *The window is open. It is black.* par *La fenêtre est ouverte. Il est noir.* Le pronom *it* est ici mal traduit en l'absence d'information sur le genre de son antécédent.

Pour éviter ce genre d'incohérences qui, dans des cas moins triviaux, nuisent grandement à la compréhensibilité du texte traduit, Le Nagard et Koehn (2010) proposent de prétraiter les documents en amont du modèle de traduction. Un système de résolution des coréférences est appliqué sur le document source pour identifier les antécédents des pronoms qui sont remplacés par des pronoms factices porteurs d'informations sur le genre que doit avoir leur traduction.

Dans ce cas d'utilisation de la résolution des coréférences, c'est la tête des mentions qui est utile puisque c'est elle qui est porteuse du genre du référent. Un syntagme maximal sans analyse syntaxique pour identifier la tête de la mention s'avérerait de fait inexploitable.

7.3 Une nouvelle définition de la tâche

Au regard de l'indépendance intrinsèque de la tâche de résolution des coréférences vis-à-vis de la façon d'identifier les mentions et des besoins différents des applications en la matière, nous proposons de redéfinir la notion de mention et par-là la tâche de résolution des coréférences elle-même.

Comme nous l'avons expliqué à la section 7.1, la résolution des coréférences à proprement parler consiste à établir des liens de coréférence entre mentions de référents de discours. Si ces mentions sont actuellement représentées par des empans textuels correspondant aux projections maximales de leurs têtes syntaxico-sémantiques, nous avons montré qu'il était équivalent de les identifier par leurs têtes elles-mêmes puisqu'il existe une bijection de principe entre les deux.

2. Dans le très utilisé jeu d'étiquettes morphosyntaxiques du Penn Treebank (Marcus *et al.*, 1993), les noms propres sont clairement marqués par les étiquettes *NNP* et *NNPS*.

Nous avons montré au chapitre 6 que prédire des syntagmes maximaux impliquait d'être capable de construire l'arbre syntaxique des phrases. Or nous n'avons aucune raison d'imposer une telle surcharge de travail aux modèles de résolution des coréférences, usuellement appris sans supervision syntaxique. Les têtes des mentions de référents de discours sont à notre avis un choix plus judicieux pour les identifier, car elles n'introduisent pas dans la tâche une difficulté supplémentaire qui ne lui est pas directement liée.

Nous proposons également de revisiter l'évaluation des chaînes de coréférence prédites en conséquence. Rappelons qu'une des particularités de la résolution des coréférences est de devoir partitionner un ensemble d'éléments qui n'est pas connu à l'avance. Aussi, l'évaluation prend en compte ces deux aspects de la tâche : trouver les mentions, et les partitionner en chaînes de coréférence.

Notre nouvelle définition de la tâche est sans conséquence sur cette seconde composante de l'évaluation et les métriques usuelles de résolution des coréférences peuvent encore être utilisées. C'est l'évaluation de la première composante de la tâche qui doit être révisée.

Les campagnes d'évaluation MUC et ACE avaient déjà tenté de limiter l'impact de la prédiction des bornes des mentions lors de l'évaluation des chaînes de coréférence. Cependant, au vu du manque de pertinence de ces bornes pour la résolution des coréférences elle-même, nous proposons de faire reposer la validité d'une mention prédite uniquement sur l'identification de sa tête. Concrètement, ce sont des chaînes de coréférence de têtes qui devraient être évaluées, puisque la tête est la partie la plus informative de la mention et qu'elle suffit pour l'identifier de façon univoque.

Notons que les systèmes traditionnels qui utilisent une analyse syntaxique pour détecter les mentions et leurs têtes comme descripteur sont très facilement adaptables puisqu'il leur suffit de mettre ces têtes dans les chaînes de coréférence prédites plutôt que les syntagmes maximaux. Ils peuvent ainsi être réévalués pour juger plus fiablement leur capacité réelle à prédire des relations de coréférence, indépendamment de l'exactitude des bornes des mentions prédites par l'analyseur syntaxique en amont.

Cette nouvelle définition de la tâche ouvre également la voie à de nouvelles approches de la résolution des coréférence qui se concentreront sur l'essence même de la tâche, sans jamais tenter de prédire des délimitations aux mentions. Nous avons vu à la section 7.2 que des systèmes capables de prédire des groupes de têtes pouvaient se révéler d'une grande valeur pour certaines applications et qu'au contraire, les systèmes qui ne prédisaient que des syntagmes maximaux n'étaient que difficilement exploitables.

Chapitre 8

Modèle basé sur les têtes

Exclure la prédiction des bornes des mentions de la résolution des coréférences elle-même ouvre la voie à un nouveau type de modèles qui ne les délimitent pas explicitement. Nous proposons dans ce chapitre une variante du modèle de Lee *et al.* (2017) qui, en épousant nativement la nouvelle définition de la tâche, le surpasse de 1.6 points CoNLL et réduit significativement aussi bien les temps de calcul que l'empreinte mémoire.

8.1 Modèle

Les délimitations des mentions sont intensivement utilisées par la plupart des systèmes de résolution des coréférences. D'une part, par le biais du traditionnel descripteur qui indique si deux mentions correspondent en termes de chaînes de caractère, d'autre part au travers des mots contenus dans la mention. Rappelons que Wiseman *et al.* (2015) utilisent comme descripteurs les premier et dernier mots de la mention, ainsi que sa tête syntaxique, le mot qui la précède et celui qui la suit.

Le modèle de Lee *et al.* (2017) a cependant largement remis en cause l'utilité de la délimitation des mentions pour la prédiction des coréférences. D'abord, la correspondance entre mentions n'est plus calculée explicitement, mais remplacée par une similarité entre représentations apprises. Ensuite, si Lee *et al.* (2017) représentent en substance les mentions par la concaténation des vecteurs BLSTM associés à leurs bornes, chacun de ces vecteurs encodent l'intégralité de la phrase de la mention.

Nous proposons d'aller encore plus loin en ne représentant une mention de tête i que par le vecteur BLSTM h_i correspondant à sa tête. On notera la disparition du descripteur indiquant la longueur du syntagme maximal, puisque celui-ci n'est pas utilisé explicitement, ainsi que du coûteux mécanisme d'attention à la tête.

Sur le plan théorique, cette approche est très satisfaisante puisqu'elle laisse le modèle déterminer l'importance relative pour la prédiction des liens de coréférence de chacun des mots autour de la tête de la mention. Il n'a en plus pas à prédire d'arbitraires et complexes bornes de mentions, ce qui, nous l'avons vu, nécessite par définition de très importantes connaissances syntaxiques.

D'autre part, cette nouvelle approche permet une diminution de la complexité intrinsèque du modèle. En utilisant le fait que les têtes des mentions suffisent pour les identifier de façon univoque dans un texte, l'espace de recherche est largement réduit et plus aucune mention ne doit être arbitrairement ignorée à cause d'un trop grand nombre de modifieurs.

Comme nous l'avons mentionné à la section 7.2, certaines applications peuvent tout à fait se contenter des prédictions de ce nouveau modèle : des groupes de têtes de mentions.

Notre approche est cependant applicable également lorsque d'autres délimitations sont requises. Grâce au choix des têtes syntaxiques pour les identifier de façon univoque, des traitements adéquats permettent de retrouver autant d'information que nécessaire sur les mentions.

Par exemple, les traditionnelles délimitations en termes de syntagmes maximaux peuvent être retrouvées efficacement par projection maximale de la tête de la mention au sein de l'arbre syntaxique de la phrase où elle apparaît.

8.2 Expériences

Pour pouvoir comparer de façon fiable les performances et les temps de calcul de notre nouveau modèle avec celui de Lee *et al.* (2017), nous avons utilisé des hyperparamètres identiques à ceux décrits dans leur article. [1] Nous avons également choisi de mener les expériences avec leur propre code source [2], seulement légèrement modifié pour pouvoir faire tourner le modèle basé sur les têtes. L'ordinateur utilisé pour les calculs est équipé d'un processeur Intel Core i7-7700 CPU doté de 8 cœurs et fonctionnant à 3.6 GHz.

Pour l'évaluation, nous avons utilisé la portion anglaise du corpus CoNLL 2012. Puisque nous avons adopté la nouvelle définition de la tâche de résolution des coréférences, la méthode d'évaluation utilisée est celle décrite à la section 7.3.

Comme le corpus CoNLL 2012 est basé sur Ontonotes, il ne fournit cependant pas explicitement les têtes des mentions. Nous les avons donc déterminées avec l'algorithme mentionné à la section 6.2.2 sur la base de la couche d'annotation syntaxique, aussi bien pour l'apprentissage que l'évaluation du modèle basé sur les têtes. Puisque les modèles de Lee *et al.* (2017) et Clark et Manning (2016a) produisent des empans plutôt que des têtes, la même procédure a été appliquée à leurs sorties avant l'évaluation.

Le tableau 8.1 compare les temps de calcul et l'empreinte mémoire de notre modèle à ceux du modèle de Lee *et al.* (2017). Ses scores de coréférence sont reportés dans le tableau 8.2 aux côtés de ceux de Lee *et al.* (2017) et Clark et Manning (2016a). Notre modèle dépasse celui de Lee *et al.* (2017) de 1.6 points CoNLL et réduit significativement la quantité de ressources matérielles nécessaires.

Modèle	**Temps (s)**	**Max RAM (Mo)**
Notre modèle	101	3181
Lee *et al.* (2017)	240	6972

Tableau 8.1 – Temps de calcul et empreinte mémoire pour l'évaluation sur le corpus de test CoNLL 2012.

8.3 Discussion

Avec l'évaluation basée sur les têtes, le modèle *single* de Lee *et al.* (2017) obtient un score de 68.25 points CoNLL, ce qui est supérieur aux 67.22 points qu'ils reportent

1. Le modèle basé sur les têtes n'utilise évidemment pas l'hyperparamètre L correspondant à la longueur maximale des mentions considérées.

2. `https://github.com/kentonl/e2e-coref`.

Modèle	MUC			B^3			$CEAF_e$			CONLL		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
Notre modèle	78.12	**77.84**	**77.98**	68.37	**67.3**	**67.83**	63.31	**64.32**	**63.81**	69.94	**69.80**	**69.87**
Lee *et al.* (2017)	79.62	74.29	76.86	70.08	62.65	66.15	63.66	59.91	61.73	71.12	65.62	68.25
Clark et Manning (2016a)	**81.49**	72.41	76.68	**72.4**	60.01	65.63	**65.66**	57.27	61.18	**73.18**	63.23	67.83

Tableau 8.2 – Scores de coréférence sur le corpus de test CoNLL 2012 (portion anglaise)

avec une évaluation basée sur les syntagmes maximaux. Ce gain s'explique par le fait que transformer les syntagmes maximaux en têtes réduit les erreurs de délimitations. Notons que nous avons utilisé pour cela la syntaxe annotée, les résultats pourraient donc être un peu moins bons avec une syntaxe prédite, plus réaliste.

L'écart avec Clark et Manning (2016a) de 1.5 points CoNLL avec une évaluation basée sur les syntagmes maximaux est drastiquement réduit à 0.42 points avec une évaluation basée sur les têtes, ce qui suggère qu'un des principaux apports du modèle de Lee *et al.* (2017) était une amélioration dans la prédiction des délimitations des mentions.

Notre modèle basé sur les têtes dépasse de son côté le modèle de Lee *et al.* (2017) de 1.62 points CoNLL sur la nouvelle tâche. Ce gain s'explique pour une large part par une augmentation du rappel dû au fait de ne pas avoir filtré les mentions selon un critère de longueur. La baisse de précision peut quant à elle peut-être s'expliquer par l'abandon du mécanisme d'attention à la tête et du descripteur encodant la longueur du syntagme maximal.

En abandonnant le mécanisme d'attention et raisonnant sur l'ensemble des mots plutôt que des empans de phrase, notre modèle basé sur les têtes permet en outre une accélération d'un facteur de presque 2.5 par rapport au modèle de Lee *et al.* (2017). L'empreinte mémoire maximale du modèle est elle aussi réduite par un facteur supérieur à deux.

Ces réductions de la quantité de ressources nécessaire sont d'autant plus appréciables que le modèle de Lee *et al.* (2017) est particulièrement gourmand en ressources. En effet, l'implémentation proposée par AllenNLP (Gardner *et al.*, 2017) a par exemple réduit le nombre maximum de candidats antécédents de 250 à 100 à cause de contraintes d'empreinte mémoire, ce qui a un effet négatif sur les performances.

Enfin, grâce à une baisse intrinsèque de complexité en la longueur des phrases, notre nouvelle approche n'a pas besoin d'ignorer les mentions avec trop de modifieurs, ce qui évite des résultats surprenants et conceptuellement injustifiables.

8.4 Perspectives

En tant que première tentative d'approche basée uniquement sur les têtes des mentions, notre modèle nous semble prometteur. En cernant au plus près la tâche de résolution des coréférences, il apporte une augmentation du rappel qui amène un équilibre naturel entre rappel et précision absent aussi bien du modèle de Lee *et al.* (2017) que de celui de Clark et Manning (2016a).

Des expériences préliminaires montrent qu'intégrer un mécanisme d'attention à une fenêtre de mots autour des têtes pourrait apporter un gain de performance supplémentaire.

Il serait également parfaitement envisageable de tenter d'optimiser notre modèle pour les métriques de coréférence grâce aux techniques d'apprentissage par renforcement, comme l'ont fait (Clark et Manning, 2016a).

Enfin, comme beaucoup avant nous, nous avons mené nos expériences dans le cadre du modèle mention-mention. Les modèles mention-entité sont théoriquement supérieurs, et il serait donc intéressant de voir comment se comporte l'approche basée sur les têtes au sein d'un tel modèle.

Conclusion

Le modèle de Lee *et al.* (2017) et l'important gain de performance qu'il a apporté ont initié une remise en question de l'approche traditionnelle de la résolution des coréférences. Plutôt que d'utiliser l'analyse syntaxique pour détecter les mentions en amont de la résolution des coréférences, leur modèle considère chaque empan de phrase comme une mention potentielle. Cette manière de faire a introduit une importante augmentation du rappel en évitant la propagation des erreurs de prétraitement.

Nous avons cependant montré que la résolution des coréférences telle qu'elle est pratiquée aujourd'hui est étroitement liée à l'analyse syntaxique. De ce fait, le modèle de Lee *et al.* (2017) peine à prédire les bornes exactes des mentions. Pour limiter les temps de calcul, il ignore en outre complètement les mentions de longueur supérieure à 10. En plus d'avoir des conséquences empiriques, cette simplification n'est guère satisfaisante d'un point de vue théorique puisque les mentions, usuellement définies en termes de syntagmes maximaux, peuvent être de longueur arbitraire.

Pourtant, la résolution des coréférences à proprement parler a pour mission d'établir des liens entre mentions de référents de discours, pas entre syntagmes. Après avoir montré que les mentions des référents de discours pouvaient parfaitement être identifiées dans les textes par leurs seules têtes syntaxico-sémantiques, nous avons proposé d'abandonner la délimitation des mentions en termes de syntagmes maximaux au profit de ces têtes.

Cette nouvelle définition des mentions implique de redéfinir la tâche de résolution des coréférences comme l'identification des mentions de référents du discours par le biais de leurs têtes et leur partitionnement en chaînes de coréférence. Nous proposons que l'évaluation de la tâche soit également révisée en conséquence, de manière à ce qu'une mention soit considérée comme bien identifiée si et seulement si la tête proposée est exacte.

Cette révision de la tâche ouvre la voie à de nouvelles approches qui ne se soucient pas du tout de prédire des délimitations textuelles pour les mentions. Dans cette perspective, nous avons proposé une variante du modèle de Lee *et al.* (2017) qui le surpasse de 1.6 points CoNLL sur la tâche de résolution des coréférences telle que nous l'avons redéfinie.

En raisonnant sur un ensemble de mots, pris comme potentielles têtes de mentions, plutôt qu'un ensemble d'empans considérés comme de potentiels syntagmes maximaux, notre modèle n'a plus aucun besoin d'ignorer les mentions avec beaucoup de modifieurs et poursuit l'amélioration du rappel initiée par Lee *et al.* (2017). Il nécessite en outre plus de deux fois moins de ressources calculatoires.

Nous espérons que nos nouvelles définition et méthode d'évaluation de la résolution des coréférences s'avéreront utiles et que notre modèle servira de ligne de base à des travaux ultérieurs. Parmi les développements possibles, on peut citer la réintroduction d'un mécanisme d'attention, l'utilisation de techniques d'apprentissage par renforcement, et la mise au point d'un modèle mention-entité basé sur les têtes.

Table des matières

Bibliographie

Ariel, M. (1988). Referring and Accessibility. *Journal of Linguistics*, 24(1):65–87.

Asher, N. (1993). *Reference to abstract objects in discourse*. Studies in linguistics and philosophy. Kluwer Academic Publ.

Bagga, A. et Baldwin, B. (1998). Algorithms for Scoring Coreference Chains. *Recall*, 5.

Bahdanau, D., Cho, K. et Bengio, Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. *In Proceedings of ICLR 2015*.

Berger, A. L., Pietra, V. J. D. et Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

Clark, K. et Manning, C. (2016a). Deep Reinforcement Learning for Mention-Ranking Coreference Models. *In Proceedings of EMNLP 2016*, pages 2256–2262.

Clark, K. et Manning, C. D. (2016b). Improving Coreference Resolution by Learning Entity-Level Distributed Representations. *In Proceedings of the 54th Annual Meeting of the ACL*, volume 1, pages 643–653.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. Thèse de doctorat, University of Pennsylvania.

Cross, J. et Huang, L. (2016). Span-based constituency parsing with a structure-label system and provably optimal dynamic oracles. *In Proceedings of EMNLP 2016*.

Deemter, K. v. et Kibble, R. (2000). On coreferring : Coreference in MUC and related annotation schemes. *Computational linguistics*, 26(4):629–637.

Denis, P. (2007). *New learning models for robust reference resolution*. Thèse de doctorat, The University of Texas at Austin.

Denis, P. et Baldridge, J. (2007). A Ranking Approach to Pronoun Resolution. *In Proceedings of IJCAI 2007*.

Denis, P. et Baldridge, J. (2008). Specialized models and ranking for coreference resolution. *In Proceedings of EMNLP 2008*, pages 660–669.

Durrett, G., Berg-Kirkpatrick, T. et Klein, D. (2016). Learning-Based Single-Document Summarization with Compression and Anaphoricity Constraints. *In Proceedings of the 54th Annual Meeting of the ACL*, volume 1, pages 1998–2008.

Durrett, G., Hall, D. L. W. et Klein, D. (2013). Decentralized Entity-Level Modeling for Coreference Resolution. *In Proceedings of the 51st Annual Meeting of the ACL*, pages 114–124.

Durrett, G. et Klein, D. (2013). Easy Victories and Uphill Battles in Coreference Resolution. *In Proceedings of EMNLP 2013*, pages 1971–1982.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N. F., Peters, M., Schmitz, M. et Zettlemoyer, L. S. (2017). AllenNLP : A Deep Semantic Natural Language Processing Platform.

Grosz, B. J., Weinstein, S. et Joshi, A. K. (1995). Centering : A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Hochreiter, S. et Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L. et Weischedel, R. (2006). OntoNotes : The 90% Solution. *In Proceedings of the Human Language Technology Conference of the NAACL*, pages 57–60.

Kummerfeld, J. K. et Klein, D. (2013). Error-driven analysis of challenges in coreference resolution. *In Proceedings of EMNLP 2013*, pages 265–277.

Le Nagard, R. et Koehn, P. (2010). Aiding pronoun translation with co-reference resolution. *In Proceedings of the Joint 5th Workshop on Statistical Machine Translation*, pages 252–261.

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 Proceedings of the Fifteenth Conference on Computational Natural Language Learning*, pages 28–34. Association for Computational Linguistics.

Lee, K., He, L., Lewis, M. et Zettlemoyer, L. (2017). End-to-end Neural Coreference Resolution. *In Proceedings of EMNLP 2017*, pages 188–197.

Luo, X. (2005). On Coreference Resolution Performance Metrics. *In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*.

Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N. et Roukos, S. (2004). A Mention-Synchronous Coreference Resolution Algorithm Based On the Bell Tree. *In Proceedings of the 42nd Annual Meeting of the ACL*, pages 135–142.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. et McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *In Proceedings of 52nd Annual Meeting of the ACL*, volume System Demonstrations, pages 55–60.

Marcus, M. P., Marcinkiewicz, M. A. et Santorini, B. (1993). Building a large annotated corpus of English : The Penn Treebank. *Computational linguistics*, 19(2):313–330.

Miclet, L. et Cornuéjols, A. (2010). *Apprentissage artificiel - Concepts et algorithmes*. Eyrolles, Paris, 2 édition.

Miller, G. A. (1995). WordNet : a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Milner, J.-C. (1982). *Ordres et raisons de langue*. Éditions du Seuil.

Ng, V. et Cardie, C. (2002a). Identifying Anaphoric and Non-anaphoric Noun Phrases to Improve Coreference Resolution. *In Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, pages 1–7.

Ng, V. et Cardie, C. (2002b). Improving machine learning approaches to coreference resolution. *In Proceedings of the 40th Annual Meeting of the ACL*, pages 104–111.

Pennington, J., Socher, R. et Manning, C. (2014). Glove : Global vectors for word representation. *In Proceedings of EMNLP 2014*, pages 1532–1543.

Poesio, M. (2004). The MATE/GNOME proposals for anaphoric annotation, revisited. *In Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*.

Poesio, M., Stuckardt, R. et Versley, Y., éditeurs (2016). *Anaphora Resolution*. Theory and Applications of Natural Language Processing. Springer Berlin Heidelberg, Berlin, Heidelberg.

Rahman, A. et Ng, V. (2009). Supervised models for coreference resolution. *In Proceedings of EMNLP 2009*, pages 968–977.

Ravichandran, D., Hovy, E. et Och, F. J. (2003). Statistical QA-Classifier vs. Re-ranker : What's the difference ? *In Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 69–75.

Schuster, M. et Paliwal, K. (1997). Bidirectional Recurrent Neural Networks. *Trans. Sig. Proc.*, 45(11):2673–2681.

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.

Stern, M., Andreas, J. et Klein, D. (2017). A Minimal Span-Based Neural Constituency Parser. *In Proceedings of the 55th Annual Meeting of the ACL*.

Stoyanov, V., Gilbert, N., Cardie, C. et Riloff, E. (2009). Conundrums in noun phrase coreference resolution : Making sense of the state-of-the-art. *In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 656–664.

Turian, J., Ratinov, L. et Bengio, Y. (2010). Word representations : a simple and general method for semi-supervised learning. *In Proceedings of the 48th Annual Meeting of the ACL*, pages 384–394.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D. et Hirschman, L. (1995). A model-theoretic coreference scoring scheme. *In Proceedings of the 6th Conference on Message Understanding*, pages 45–52.

Wiseman, S., M. Rush, A. et M. Shieber, S. (2016). Learning Global Features for Coreference Resolution. *In Proceedings of the 2016 Conference of the NACACL*, pages 994–1004.

Wiseman, S. J., Rush, A. M., Shieber, S. M. et Weston, J. (2015). Learning anaphoricity and antecedent ranking features for coreference resolution. *In Proceedings of the 53rd Annual Meeting of the ACL and the 7th IJCNLP*.

Annexe A

Ordonnement contre classement

A.1 Gradients

Pour évaluer la différence de comportement à l'apprentissage des modèles logistiques de classement et d'ordonnement, nous avons besoin de calculer les valeurs des mises à jour du modèle, et donc des gradients des fonctions de coût.

Donnons-nous un corpus d'apprentissage $\{m^1, \ldots, l^l\}$ de mentions à résoudre. Notons C^i l'ensemble des candidats antécédents de m^i, et c^i_* son antécédent de référence, supposé par souci de simplicité en faire partie.

Le gradient de la log-vraisemblance L^c du modèle de classement appliqué en lot sur l'ensemble de notre corpus d'apprentissage est donné par

$$\frac{\partial L^c(w)}{\partial w_k} = \sum_{i=1}^{l} \Phi_k(c^i_*) - \sum_{c^i_j \in C^i} P(c^i_j | m^i, C^i; w).\Phi_k(c^i_j)$$

De même, le gradient de la log-vraisemblance L^o du modèle d'ordonnement appliqué en lot sur notre corpus d'apprentissage est :

$$\frac{\partial L^o(w)}{\partial w_k} = \sum_{i=1}^{l} \Phi_k(c^i_*) - \sum_{c^i_j \in C^i} P(oui \mid m^i, c^i_j; w).\Phi_k(c^i_j)$$

Rappelons que dans la variante la plus simple de la descente (ici ascension) de gradient qui conviendra parfaitement ici, les mises à jour de chacun des poids w_k sont proportionnelles à $\frac{\partial L(w)}{\partial w_k}$.

Les deux gradients partagent un premier terme qui donne du poids à l'ensemble des descripteurs des bons candidats. Ce poids est cependant nuancé par un second terme qui diffère selon le type de modèle considéré.

A.2 Exemple

Reprenons pour illustrer l'exemple de Denis (2007). Supposons que notre corpus d'apprentissage soit composé de trois mentions m^1, m^2, et m^3 décrites par deux descripteurs binaires, f_1 et f_2 de la manière présentée dans le tableau A.1.

Supposons en outre que l'apprentissage n'ait pas encore commencé et que le vecteur de poids w soit nul. Que ce soit dans le modèle de classement ou le modèle d'ordonnement, les probabilités seront distribuées uniformément sur les différents résultats possibles. Dans le

Mention	**Candidat**	**Coréférence**	**Descripteurs vrais**
m^1	c_1^1	oui	f_2
	c_2^1	non	f_1
m^2	c_1^2	oui	f_2
	c_2^2	non	f_1
m^3	c_1^3	oui	f_1
	c_2^3	non	f_2
	c_3^3	non	f_2
	c_4^3	non	f_2
	c_5^3	non	f_2
	c_6^3	non	f_2

TABLEAU A.1 – Corpus d'apprentissage jouet

cas du classement, les événements *Oui* et *Non* seront donc chacun dotés d'une probabilité de $\frac{1}{2}$, et dans le cas de l'ordonnement, chaque candidat antécédent d'une mention m^i sera doté de la probabilité $\frac{1}{|C^i|}$.

Le calcul des gradients de la log-vraisemblance donne

$$\frac{\partial L^c(w)}{\partial w} = (-\frac{1}{2}, -\frac{3}{2}) \qquad \frac{\partial L^o(w)}{\partial w} = (-\frac{1}{6}, \frac{1}{6})$$

Il ressort que, sur ce lot d'exemples, le classeur décide que c'est f_1 qui est le descripteur le plus caractéristique des bons couples mention-candidat antécédent, tandis que l'ordonneur décide que c'est f_2.

En réalité, f_1 est présent sur 1 bonne paire sur 3, tandis que f_2 l'est sur 2 bonnes paires sur 3. f_2 discrimine donc plus efficacement les bonnes paires, et c'est l'ordonneur qui est le plus pertinent.