Résolution d'anaphores et identification des chaînes de coréférence : une approche « minimaliste »

Sylvie Boudreau, Richard Kittredge

Université de Montréal

Abstract

The problem of recognizing co-referring expressions in a text (proper nouns, definite noun phrases, pronouns, etc. which refer to the same entity) is fundamental to automatic "understanding" of natural language. This work deals with coreference chains headed by proper nouns, and compares three varieties of French texts with regard to the distribution of multiple chains and the anaphoric links within them. Our corpus analysis reveals important differences among the three varieties regarding the use and distribution of referring expressions, but suggests a special role for domain nouns in correctly assigning referring expressions to competing chains. We have developed a simple algorithm, using limited linguistic knowledge, for partitioning the referring expressions of a text into distinct coreference chains, and tested it on our *XML*-coded corpus.

Résumé

La reconnaissance des expressions coréférentielles (les expressions qui réfèrent à la même entité, c'est-à-dire les noms propres, les syntagmes nominaux définis, les pronoms, etc.) est importante pour la « compréhension » d'un texte en langue naturelle par un ordinateur. Dans ce travail, nous nous intéressons aux chaînes de coréférence chapeautés par un nom propre et nous comparons trois variétés de textes français dans le but d'étudier la distribution des éléments coréférentiels ainsi que les liens anaphoriques qui unissent les syntagmes d'une même chaîne. Notre analyse en corpus démontre des différences entre les trois variétés de textes en ce qui concerne l'utilisation et la distribution des expressions référentielles, mais signale également le rôle important des noms du domaine pour assigner une expression référentielle à la bonne chaîne. Nous avons développé un algorithme simple, utilisant des connaissances linguistiques limitées, qui partitionne les expressions référentielles d'un texte en chaînes de coréférence distinctes et nous avons testé cet algorithme sur notre corpus codé en *XML*.

Mots-clés : résolution d'anaphores, chaîne de coréférence, nom propre, variété de textes, linguistique de corpus, langage de balisage *XML*.

1. Introduction

Dans un texte, il est fréquent de retrouver beaucoup d'éléments référentiels. Nous nous intéressons plus particulièrement aux noms propres et à leurs coréférents. Ces derniers peuvent être de plusieurs natures et avoir des comportements référentiels assez diversifiés.

La façon traditionnelle de regrouper les expressions référentielles d'un texte s'effectue selon la référence à une même entité du monde extra-linguistique. Les éléments ainsi regroupés appartiennent à une chaîne de coréférence. En étudiant les chaînes de coréférence de plus près, nous observons que leur longueur (le nombre d'éléments qui les composent) est plutôt variable, que leur distribution dans le texte est inégale et que les moyens pour repérer l'antécédent sont nombreux. Une question peut se poser : malgré cette grande variabilité, est-il facile d'établir ces chaînes de façon automatique ?

Dans ce travail, nous nous intéresserons plus particulièrement aux ressources linguistiques minimales nécessaires pour établir les chaînes de coréférence et aux ajustements propres à

une variété de textes donnée. Pour ce faire, nous classerons les différents types de syntagmes référentiels et nous ferons une brève revue de la littérature portant sur la résolution automatique des anaphores et des chaînes de coréférence. Par la suite, à partir de l'analyse d'un corpus composé de trois types de textes différents, nous mettrons en pratique ces notions dans le but de construire un algorithme simple et partiellement adapté au type de texte permettant l'établissement automatique des chaînes de coréférence.

2. Les types de syntagmes référentiels

Les syntagmes référentiels se présentent sous plusieurs formes de surface. Nous y retrouvons des noms propres (*Donna*, *Steve*), des syntagmes nominaux (*une femme*, *son bracelet*), des syntagmes sans nom (*ce dernier Ø*) ainsi que des pronoms (*il*, *lui*). Les syntagmes référentiels nominaux sont décomposés, selon le type de déterminant utilisé (le tableau 1 illustre différents types de déterminants dans un contexte restreint). Bien que certains syntagmes soient indépendants du contexte (indéfini, défini générique et défini autonome), d'autres nécessitent un antécédent pour fixer la référence (défini de reprise, défini lexical, défini associatif et démonstratif). Dans les cas qui dépendent d'un autre élément, il est possible de construire, dans le texte, une fenêtre contextuelle susceptible de contenir l'antécédent. La dimension de la fenêtre peut varier en fonction du type de déterminant de l'élément anaphorique. Par exemple, la fenêtre associée à un déterminant démonstratif est beaucoup plus petite que celle d'un déterminant défini (Corblin, 1987).

Indéfini	*Un chat gris miaule à la fenêtre...*
Défini	
Générique	*Le chat est un animal qui miaule.*
Autonome	*Le chat de la voisine miaule à la fenêtre...*
De reprise	*... Le chat veut entrer dans la maison.*
Lexical	*... Le félin veut entrer dans la maison.*
Associatif	*... La chanson est peu mélodieuse.*
Démonstratif	*... Ce soprano me tombe sur les nerfs !*

Tableau 1 : Les types de syntagmes référentiels

D'un autre point de vue, il est aussi possible de classer les syntagmes référentiels selon les relations qu'ils entretiennent avec les autres syntagmes de la chaîne de coréférence : ils peuvent être la première occurrence du texte (tête) ou dans le corps de la chaîne de coréférence. Dans ce dernier cas, nous distinguons les coréférences (indépendance du contexte, par exemple la répétition d'un nom propre) des anaphores (nécessitant la fixation d'une tête lexicale). Les anaphores peuvent à leur tour être divisées selon le procédé utilisé pour fixer la tête nominale : par reprise (on « recopie une partie du syntagme ») ; par des procédés lexicaux (un synonyme par exemple) ; par l'utilisation d'un pronom, ou par des procédés dictés par la syntaxe. (Voir le tableau 2 et l'exemple (1) dans lequel les relations entre les syntagmes sont indiquées. Dans ce texte, nous avons relevé 12 chaînes de coréférence, dont 9 débutent par un nom propre.).

Tête (t)

Coréférence (c)

Anaphore

De reprise (ar)

Lexicale (al)

Pronominale (ap)

Syntaxique (as)

Tableau 2 : Les relations entre les syntagmes référentiels

(1) *Une femme demande$_t$ à Donna$_t$ de lui$_{ap}$ créer une robe spécialement pour une séance de spiritisme. Celle-ci$_{ap}$ voudrait entrer en communication avec son défunt mari$_t$ afin qu'il$_{ap}$ l'$_{ap}$aide à retrouver son bracelet de diamants$_t$. Steve$_t$ demande à Carly$_t$ de l'$_{ap}$accompagner au bal de la moisson et Cooper$_t$ en fait de même avec Val$_t$. La situation financière de David$_t$ ne s'arrange pas. On lui$_{ap}$ refuse même une demande de crédit. Au bal, Brandon$_t$, Kelly$_t$, Steve$_c$ et Carly$_c$ sont surpris de voir Val$_c$ et Cooper$_c$ danser ensemble. Donna$_c$ arrive seule au bal et provoque la jalousie de Val$_c$ en flirtant avec Noah$_t$. Ce dernier$_{al}$ s'approche tout de même de Val$_c$ et provoque maintenant la jalousie de Cooper$_c$ qui$_{as}$ décide de s'en aller. Mais lorsque Noah$_c$ voit le bracelet$_{ar}$ que Cooper$_c$ a offert à Val$_c$, il$_{ap}$ devient furieux*[1].

3. Travaux antérieurs

J. Hobbs (1978) fut parmi les premiers à étudier la résolution automatique des anaphores. Le prétraitement nécessaire à sa méthode de résolution demande une analyse syntaxique complète et l'établissement des chaînes de coréférence s'effectue à l'aide du parcours de l'arbre syntaxique selon un ordre particulier. Les critères de sélection pour un antécédent reposent sur les compatibilités syntaxiques et sémantiques (genre, nombre, animation...) avec l'élément anaphorique. Due à la façon de parcourir l'arbre syntaxique, les facteurs tels la fonction syntaxique *sujet* d'un antécédent une petite distance entre l'antécédent et l'anaphore sont implicitement favorisés lors de la sélection. Ce n'est que beaucoup plus tard que (Lappin et Leass, 1994) ont utilisé explicitement ces facteurs comme critère d'antécédence. Dans cette méthode, plutôt que de parcourir un arbre syntaxique selon un ordre particulier, les candidats antécédents reçoivent un pointage selon un ensemble de caractéristiques (distance, fonction syntaxique, construction existentielle).

Les travaux de la deuxième moitié des années 1990 sont caractérisés par une volonté de simplification et de diminution des ressources linguistiques nécessaires aux algorithmes de résolution des anaphores. Dans le courant des approches *knowledge-poor*, (Mitkov, 1998) propose un algorithme qui nécessite en prétraitement l'identification des syntagmes référentiels, plutôt qu'une analyse syntaxique complète. En outre, ses critères de préférences

[1] Tiré d'une description d'un épisode de l'émission de télévision *Beverly Hills* au réseau de télévision TVA.

pour la sélection de l'antécédent valorisent les syntagmes définis, le thème de la phrase, les réitérations lexicales et l'appartenance de la tête nominale au vocabulaire du domaine ou aux mots du titre. Ensuite, Baldwin (1997) examine la possibilité de ne résoudre que les liens anaphoriques « les plus probables », favorisant ainsi une meilleure précision des résultats. Il remarque que les antécédents se situent habituellement dans la phrase courante ou la phrase précédente. Les travaux de Bergler (1997) s'attardent aux chaînes de coréférence appartenant au sujet du texte (c'est-à-dire les chaînes qui débutent dans les trois premières phrases du texte). Ces chaînes semblent plus faciles à résoudre car, comparativement aux autres chaînes du texte, elles sont plus longues, elles contiennent plus de réitérations lexicales et le vocabulaire utilisé est plus restreint. Enfin, dans le but de réduire encore la complexité du prétraitement, Siddharthan (2003) propose l'utilisation des informations du texte pour déduire les fonctions syntaxiques ainsi que les traits syntaxiques. En particulier, les prépositions et les positions des syntagmes vis-à-vis du verbe sont deux bons indices pour trouver la fonction syntaxique d'un syntagme nominal. De plus, il atteste que certains traits syntaxiques peuvent être tout simplement déduits à l'aide des autres éléments du texte (déterminants, accord) ou des chaînes déjà identifiées.

Notons que la plupart de ces travaux ont été effectués sur des textes anglais. Cependant, ces méthodes sont assez générales pour être adaptées au traitement des anaphores de textes français.

4. Analyse du corpus

Pour analyser les comportements référentiels des syntagmes référentiels dont la tête syntaxique est un nom propre et des autres syntagmes s'y référant, nous avons effectué une analyse en corpus. Nous avons construit un corpus constitué de trois variétés de textes différentes en français (voir tableau 3). Le premier sous-corpus est constitué de textes journalistiques qui traitent des fusions de compagnies ; ils ont été étudiés par les conférences MUC (Hirschmann, 1997). Le deuxième sous-corpus est composé de critiques de films, qui sont aussi des textes journalistiques. Leur contenu est cependant moins informatif et plus expressif que le premier. Enfin, le troisième corpus, les *HOWTO Linux*, décrit des procédures pour l'installation de logiciels sous le système d'exploitation *Linux*.

	textes	ph	sn	mots
Critiques de films	20	414	1012	9730
Fusions de compagnies	20	356	972	9481
HOWTO Linux	20	1516	2468	28717

Tableau 3 : Composition du corpus

Ce corpus a été étiqueté manuellement, en utilisant le langage de balisage *XML* (*Extensible Markup Language*) selon la procédure suivante : dans un premier temps, nous avons relevé tous les noms propres des textes. Dans un deuxième temps, nous avons déterminé les liens coréférentiels entre les noms propres identifiés. Dans un troisième temps, nous avons complété ces chaînes avec les autres syntagmes référentiels du texte (principalement les pronoms, les syntagmes définis et les syntagmes démonstratifs). Ainsi, à partir de cette méthode, appliquée à l'exemple (1), nous aurions obtenu la chaîne {*Noah ... Ce dernier ...*

Noah ... il}. L'analyse des données ainsi étiquetées a par la suite été réalisée à l'aide de requêtes construites à partir de *XPath* (*XML Path Language*).

4.1. Différences référentielles entre les variétés de textes

Bien que la plupart des procédés référentiels soient semblables d'une variété de textes à une autre, notre analyse a relevé un certain nombre de différences quant au vocabulaire et aux types de constructions référentielles employées. Les noms communs utilisés dans les constructions définies anaphoriques (dont un antécédent est un nom propre) peuvent être classés en différentes catégories, selon le domaine étudié :

Critiques de films

Relations familiales (*mère, frère*)

Métiers (*superhéro, acteur*)

Fusions de compagnies

Synonymes de fusion (*mariage, acquisition*)

Synonymes de compagnie (*société, groupe*)

Termes du domaine des affaires (*conseil, actif*)

HOWTO Linux

Termes informatiques (*fichier, disque, serveur*)

Tableau 4 : Le vocabulaire utilisé en fonction de la variété de textes

D'autres déviations entre les variétés de textes étudiées concernent des constructions référentielles particulières qui ne se trouvent que dans une variété de textes. Par exemple, dans les critiques de films, l'identification du titre du film peut parfois être complexe :

(2) *Astérix et Obélix : Mission Cléopâtre, d'Alain Chabat, possède touts [sic] les atouts de ce que les Américains appellent un blockbuster.*

De plus, dans les *HOWTO Linux*, quelques expressions linguistiques ressemblent à des noms propres :

(3) *Le répertoire /usr/src*

(4) *Le bus PCI*

(5) *Lancez gcc -v*

De cette analyse, nous concluons qu'un algorithme de résolution d'anaphores doit, au minimum, identifier les syntagmes référentiels dont les noms communs appartiennent à un vocabulaire prédéterminé par le domaine. Il serait aussi avantageux de faire un traitement particulier pour d'autres types d'expressions, selon la variété de textes étudiée. Ce traitement ne fait pas partie du présent projet, mais nous envisageons une étude ultérieure de ces constructions.

5.Algorithme

Pour faire suite à l'analyse des données du corpus, nous avons construit un algorithme simple de résolution de chaînes de coréférence. Nous mettons l'emphase sur mot *simple* car, dans

notre travail, nous n'utilisons pas de dictionnaire ni d'analyseur syntaxique ou d'autres outils de la linguistique informatique. Ce choix est motivé par deux raisons principales. La première est que, bien souvent, le chercheur travaille avec des ressources limitées (problèmes d'accès aux dictionnaires, aux corpus, temps, argent). La deuxième concerne la neutralité par rapport aux cadres théoriques. Cette dernière raison pose un problème particulièrement pour le traitement des variétés de textes contenant des phrases considérées agrammaticales dans la langue générale. Par exemple, dans les recettes de cuisine, nous retrouvons des constructions du type *mettre Ø au four à 350 pendant une heure* (cette construction serait agrammaticale dans un autre contexte car le verbe *mettre* nécessite un COD).

L'algorithme est en quatre étapes, divisé selon les niveaux organisationnels linguistiques : 1) étape lexicale ; 2) étape syntaxique I (concernant phénomènes syntaxiques opérant à l'intérieur d'un syntagme) ; 3) étape syntaxique II (concernant les phénomènes syntaxiques entre des syntagmes différents) ; 4) étape coréférentielle. Chacune de ces étapes dépend des précédentes. Dans la suite de cette section, nous illustrerons brièvement le déroulement de l'algorithme à l'aide d'exemples. Il est important de noter que ce procédé n'identifie pas toutes les expressions référentielles d'un texte ; cependant, il est probable que les expressions les plus saillantes ou « importantes » soient sélectionnées. De plus, le repérage des bornes des syntagmes référentiels et l'assignation des fonctions syntaxiques sont approximatifs. Les détails de l'algorithme se trouvent dans Boudreau (2004).

5.1. Étape lexicale

La première étape consiste à identifier les mots utiles pour les étapes suivantes. Pour ce faire, nous avons développé, essentiellement à partir du corpus d'analyse, un vocabulaire de quelques centaines d'éléments constitués de mots grammaticaux et lexicaux. Les mots lexicaux sont principalement les mots spécifiques d'un domaine et varient donc en fonction du type de texte traité. À cela, nous ajoutons les noms propres qui se reconnaissent grâce à la majuscule.

Noms propres (majuscule)

Mots grammaticaux (liste fermée)

Pronoms (*il, lui*)

Déterminants (*un, le, ce*)

Prépositions (*mais, à, de*)

Ponctuations (*,, ., (, ;*)

Autres (auxiliaires, conjonctions)

Mots lexicaux

Noms communs associés au domaine (voir section 4.1)

Noms importants pour tous les textes (*dernier, article, jour*)

Verbes impersonnels (*semble, faut, s'agit*)

Tableau 5 : Les syntagmes référentiels identifiés

5.2. Étape syntaxique I

L'étape syntaxique I construit des syntagmes à partir des indices laissés par l'étape précédente. Par exemple, nous détectons les suites de noms propres, les pronoms, les syntagmes nominaux débutants par un déterminant et contenant un nom commun spécifique au domaine ainsi que certains syntagmes complexes[2].

- Noms propres
  (6) *Coline Serreau*
  (7) *la petite Marie*
- Pronoms (troisième)
- Syntagmes nominaux
  (8) *la compagnie*
- Cas complexes
  (9) *la fille de Coline Serreau*

5.3. Étape syntaxique II

L'attribution (approximative) des fonctions syntaxiques repose sur la présence de préposition ou de signes de ponctuation particuliers et sur la position d'un syntagme référentiel par rapport aux autres référentiels de la phrase.

- Complément du nom
  (10) *la fille de* ***Coline Serreau***
  (*la fille* étant déjà identifié comme étant un syntagme référentiel)
- COI
  (11) *appuyez sur* ***la touche F5***
- Apposition
  (12) *Cléopâtre (****la superbe Monica Bellucci****)*
  (*Cléopâtre* étant déjà identifié comme étant un syntagme référentiel)
- Sujet et COD (cas par défaut)[3]
  (13) ***Gatlif*** *persiste et déçoit*

5.4. Étape coréférentielle

À la lumière des travaux antérieurs et de l'analyse manuelle du corpus, nous avons sélectionné quelques facteurs pouvant influencer le choix d'un antécédent dans un algorithme de résolution d'anaphores :

- la nature du déterminant et de la tête du syntagme ;
- la compatibilité des traits syntaxiques et sémantiques ;
- la fonction syntaxique (sujet > COD > COI > autre) ;
- la distance entre l'anaphore et l'antécédent ;
- la présence de réitération lexicale ;
- l'appartenance de la tête nominale au vocabulaire du domaine.

[2] Dans ce dernier cas, notre algorithme a parfois du mal à déterminer la bonne fin du syntagme.
[3] Nous ne différentions pas les sujets des COD.

Nous avons donc utilisé et pondéré ces facteurs pour notre propre algorithme. Notons que notre approche limite le nombre de syntagmes référentiels à traiter. Cette restriction s'est avérée avantageuse pour nous car la liste des expressions référentielles à partitionner est plus petite et ses éléments référentiels sont plus simples (en particulier, elle contient beaucoup de noms propres). Voici quelques exemples qui ont été bien analysés par notre algorithme :

– Coréférence

(14) *Denys Arcand ... Arcand*

– Anaphore de reprise

(15) *La société ABB ... la société*

– Anaphore lexicale (démonstratif)

(16) *Combustion Engineering ... cette dernière*

– Anaphore pronominale

(17) *Alors que cette fusion n'aura aucun effet sur l'exploitation d'affaires de Téléglobe, elle permettra une plus grande flexibilité opérationnelle.*

6. Conclusion

Dans cette étude, nous avons constaté que même avec des ressources linguistiques limitées, il est possible de construire un algorithme simple qui permet d'établir les chaînes chapeautées par un nom propre dans de courts textes de styles variés. Pour ce faire, nous utilisons en grande partie les éléments linguistiques qui se trouvent dans le texte (prépositions, la morphologie de certains mots, la position dans la phrase). À cela, nous ajoutons une courte liste de mots qui peut varier selon la variété de textes à analyser.

Références

Baldwin B. (1997). CogNIAC : high precision coreference with limited knowledge and linguistic resources. *In Proceedings of ACL/EACL workshop on Operational factors in practical, robust anaphora resolution (ACL97)*, Madrid, Espagne : 38-45.

Bergler S. (1997). Towards reliable partial anaphora resolution. *In Proceedings of ACL/EACL workshop on Operational factors in practical, robust anaphora resolution (ACL97)*, Madrid, Espagne. Présenté à la « Human Language Technology Conference ».

Boudreau S. (2004). *Résolution d'anaphores et identification des chaînes de coréférence selon le type de texte*. Mémoire de maîtrise, Université de Montréal, Montréal.

Corblin F. (1987). *Indéfini, défini et démonstratif : constructions linguistiques de la référence*. Librairie Droz, Paris.

Corblin F. (1995). *Les formes de reprise dans le discours*. Presses Universitaires de Rennes, Rennes, France.

Hirschmann L. et N. Chinchor (1997). MUC-7 coreference task definition. In *Message Understanding Conference Proceedings*.

Halliday M.A.K. et R. Hasan (1976). *Cohesion in English*. Longman, London, 7 édition.

Hobbs J. (1978). Resolving Pronoun References. *Lingua*, 44 : 311-338.

Kittredge R. (1982). Homogeneity and variation of sublanguages. *In* R. Kittredge & J. Lehrberger, éditeur : *Sublanguage : Studies of Language in Restricted Semantic Domains*, Berlin, De Gruyter : 107–137

Kleiber G. (1981). *Problèmes de référence : descriptions définies et noms propres*. Centre d'analyse syntaxique, Metz.

Lappin S. et H.J. Leass (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4) : 535-561.

Mitkov R. (1998). Robust pronoun resolution with limited knowledge. *In Proceedings of COLING-ACL*, Montréal : 869-875.

Siddharthan A. (2003). Resolving pronouns robustly : Plumbing the depths of shallowness. In *Proceedings of the Workshop on Computational Treatments of Anaphora, 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2003)*, Budapest, Hungary : 7-14.