Résolution d'anaphores et traitement des pronoms en traduction automatique à base de règles

Sharid Loáiciga
Laboratoire d'Analyse et de Technologie du Langage
CUI - Université de Genève
Battelle - bâtiment A, 7 route de Drize, CH-1227 Carouge
sharid.loaiciga@unige.ch

RÉSUMÉ

La traduction des pronoms est l'un des problèmes actuels majeurs en traduction automatique. Étant donné que les pronoms ne transmettent pas assez de contenu sémantique en eux-mêmes, leur traitement automatique implique la résolution des anaphores. La recherche en résolution des anaphores s'intéresse à établir le lien entre les entités sans contenu lexical (potentiellement des syntagmes nominaux et pronoms) et leurs référents dans le texte. Dans cet article, nous mettons en œuvre un premier prototype d'une méthode inspirée de la théorie du liage chomskyenne pour l'interprétation des pronoms dans le but d'améliorer la traduction des pronoms personnels entre l'espagnol et le français.

ABSTRACT

Anaphora Resolution for Machine Translation

Pronoun translation is one of the current problems within Machine Translation. Since pronouns do not convey enough semantic content by themselves, pronoun processing requires anaphora resolution. Research in anaphora resolution is interested in establishing the link between entities (NPs and pronouns) and their antecedents in the text. In this article, we implement a prototype of a linguistic anaphora resolution method inspired from the Chomskyan Binding Theory in order to improve the translation of personal pronouns between Spanish and French.

MOTS-CLÉS : Résolution d'anaphores, traduction automatique à base de règles, sujets nuls.

KEYWORDS: Anaphora Resolution, Rule-based Machine Translation, nul subjects.

1 Introduction

Bien que "une utilisation inappropriée ou l'échec dans l'utilisation des pronoms rend la communication moins fluide" (Brennan *et al.*, 1987) [1] en compromettant la cohérence de la traduction d'un texte, la résolution d'anaphores en traduction automatique (TA) n'a été que peu utilisé jusqu'à ces dernières années. Ainsi, dans l'exemple (1), même si la traduction

1. Texte original en anglais.

transmet l'essentiel du texte original, sa lecture peut être trompeuse. [2] En effet, dans un contexte multilingue comme la TA, le problème est exacerbé puisque les caractéristiques des deux langues concernées doivent être considérées en même temps et que les langues n'ont pas toujours une correspondance un-à-un dans leur utilisation du système pronominal.

(1) a. *Source* La compañía, una de las más antiguas de Oriente Próximo, tiene numerosos críticos en su propio país y son muchas e insistentes las voces que reclaman su privatización.

b. *Référence* La société, **une** des plus anciennes du Moyen-Orient, a été très critiquée dans **son** propre pays et nombreuses et insistantes sont les voix qui appellent à **sa** privatisation.

c. *Traduction automatique* La société, l'**un** des plus anciens de Oriente Próximo, a beaucoup de critiques dans **leur** propre pays et **ils** sont nombreux et insistants voix appelant à la privatisation.

1.1 La résolution des anaphores et la traduction automatique

L'intérêt pour la tâche de RA a émergé dans la littérature dès les années 1970. La plupart des travaux pionniers ont été développés en utilisant une approche à base de règles, par exemple les algorithmes proposés par Hobbs (1986) et Lappin et Leass (1994) ; en revanche, les propositions les plus récentes utilisent des systèmes statistiques, principalement à la suite de Soon *et al.* (2001), et sont plutôt intéressés par la résolution de la coréférence.

En effet, il est nécessaire de clarifier la distinction entre deux tâches étroitement liés, mais avec des intérêts différents, à savoir la résolution des anaphores (RA) et la résolution de la coréférence. La première s'intéresse à trouver l'antécédent des expressions sans contenu référentiel (en général les pronoms et les expressions de quantification), alors que la deuxième cherche à créer des liens entre toutes les expressions qui pointent sur une même entité discursive (incluant les pronoms mais aussi des syntagmes nominaux (SN) référentiels, p. ex. *Mr. Obama* et *le président des États-Unis*).

Au fil du temps, ces travaux se sont concrétisés dans des systèmes de RA ou bien de résolution de la coréférence. Parmi beaucoup d'autres, on peut citer RAP (Lappin et Leass, 1994), un algorithme qui privilégie les arguments en fonction de la saillance ; MARS (Mitkov *et al.*, 2002), fondé uniquement sur des heuristiques ; BART (Broscheit *et al.*, 2010), fondé sur des contraintes éliminatoires et des préférences de sélection.

La RA pour la TA a été particulièrement dynamisée à partir des études de Le Nagard et Koehn (2010) et Hardmeier et Federico (2010). Ces travaux proposent d'annoter les pronoms dans la langue source avec des informations sur leurs antécédents dans la langue cible. La recherche de l'antécédent a été effectuée à l'aide des algorithmes de Hobbs et de Lappin et Leass dans le premier étude, et à l'aide du système BART dans le second. Dans les deux cas, les améliorations obtenues ne sont que modestes, en raison principalement de la performance peu satisfaisante des systèmes de RA choisis.

2. Traduit de l'espagnol en utilisant *Google Translate* (`http://translate.google.com/\#es/fr/`).

1.2 Résolution d'anaphores inspirée de la théorie du liage

La *théorie du liage* (Chomsky, 1981) est le principal instrument linguistique utilisé pour la résolution d'anaphores (tant pour des systèmes à base de règles que pour des systèmes statistiques). Ce n'est pas une méthode de RA en soi, mais elle comprend un ensemble de contraintes hiérarchiques qui permettent d'exclure des antécédents potentiels au sein de la phrase. Pour ce qui est des pronoms, ces contraintes suivent deux principes : le **Principe A** stipule que les pronoms réfléchis et réciproques trouvent leurs antécédents à l'intérieur de leur catégorie gouvernante (la proposition la plus petite qui les inclut) ; le **Principe B** établit que les pronoms personnels de 3ème personne trouvent leurs antécédents à l'extérieur de la proposition qui les inclut. [3] [4]

Dans cet article nous évaluons l'impact de la résolution d'anaphores dans la traduction automatique des pronoms de notre système de traduction à base de règles ITS-2 (Wehrli *et al.*, 2009). Celui-ci est un traducteur automatique avec une architecture de transfert fondé sur l'analyseur syntaxique Fips (Wehrli, 2007). Le processus de traduction se fait en trois étapes principales : l'analyse syntaxique du texte source, le transfert, et la génération dans la langue cible. Nôtre prototype de composant de RA intervient au cours de l'analyse syntaxique, avant le processus de traduction (Nerima et Wehrli, 2013).

Nous avons choisi la paire des langues espagnol-français pour évaluer la performance du composant. À cet égard, une des difficultés pour le traitement de l'espagnol est l'omission du pronom sujet (*pro-drop*). En d'autres termes, on peut ne pas mentionner les pronoms personnels de sujet et s'appuyer presque entièrement sur une morphologie verbale assez distinctive pour différencier les personnes grammaticales. Ainsi, les verbes fléchis en espagnol, ne comportent pas des pronoms, tel qu'il est montré ci-dessous dans l'exemple (2) en utilisant le symbole ∅.

(2) a. *Espagnol* ∅ Ha prometido un mejor servicio.
b. *Français* Il a promis un meilleur service.

2 Corpus

Pour l'évaluation du composant nous avons utilisé le corpus Ancora (Taulé *et al.*, 2008) et nous avons exploité ses annotations de la coréférence, Ancora-Co (Recasens et Martí, 2010). Il s'agit d'un corpus disponible tant pour l'espagnol que pour le catalan et dont chaque partie est composée d'articles journalistiques. Nous nous sommes servis de la partie espagnole uniquement.

Travailler avec un corpus annoté nous a permis d'avoir une mesure de référence et de comparaison. Ainsi, nous avons fait une sélection de 18 articles, correspondant à un total de

3. Le 3ème principe, le **Principe C**, stipule que les expressions référentielles (syntagmes nominaux pleins) ne peuvent pas être liés (Reinhart, 1983; Büring, 2005). Ce principe n'est pas pertinent dans ce travail.

4. Un autre instrument linguistique pour la RA est la *Théorie des Représentations Discursives Segmentées* (SDRT) de Lascarides et Asher (2007). Celle-ci est un cadre théorique pour l'analyse du discours dans son propre droit, mais il faut noter qu'il trouve ses origines dans la *Théorie de Représentation du Discours* (DRT), initialement proposée par Kamp et Reyle (1993), et dans la *Théorie de la Structure Rhétorique* (RST) formulée par Mann et Thompson (1988). Cette théorie fournit un cadre pour l'interprétation dynamique des pronoms, de l'anaphore temporelle et des présuppositions. Nous reviendrons sur cette théorie ultérieurement dans la section 4.

250 phrases et nous avons gardé la structure de chaque article. [5] Sept catégories de pronoms ont été trouvées sur la base des annotations de Ancora-Co (des pronoms interrogatifs n'ont pas été trouvés).

Nous avons évalué les pronoms qui ont été annotés comme partie d'une chaîne coréférentielle. Pour nos 18 articles, le Tableau 1 indique les chiffres correspondants aux catégories des pronoms trouvés – 229 en total – et leurs distributions.

Type de pronom	**Personnel**			**Relatif**	**Possessif**	**Démons-tratif**	**Indéfini**
	Nul	**OD**	**OI**				
Total	78	9	5	83	43	5	6
%	34.1	3.9	2.2	35.8	18.8	2.9	2.6

TABLE 1 – Distribution des pronoms anaphoriques dans 18 articles extraits du corpus Ancora.

3 Résolution des pronoms anaphoriques

Notre stratégie de résolution rappelle celle utilisée par Hobbs, mais contrairement à son implémentation, nous n'assumons pas d'arbres parfaitement analysés, et en plus, nous avons limité la recherche de l'antécédent à une seule phrase précédente. L'approche utilisée peut aussi se comparer à celle de Lappin & Leass ; néanmoins, nous avons effectué des analyses approfondies afin d'exploiter les ressources linguistiques de notre analyseur syntaxique. Contrairement à ces travaux, notre composant s'applique au cours de l'analyse syntaxique, dès que l'analyseur rencontre un pronom.

L'algorithme de RA est décrit dans les lignes suivantes :

Pour chaque pronom trouvé :

1. **Vérifier la nature du pronom :**
   (a) **Pronom impersonnel.** À l'aide d'informations lexicales, ainsi que de constructions adjectivales tel que *il est évident que ...*, les pronoms impersonnels sont écartés de toute considération ultérieure dans l'algorithme.
   (b) **Pronom réfléchi ou réciproque.** Nous assumons une interprétation simplifié du **Principe A** dans laquelle ce type de pronom renvoie toujours au sujet de la phrase qui le contient pour son interprétation. Dans les cas des phrases infinitives enchâssées, nous assumons un pronom sujet PRO (non-réalisé lexicalement) dont l'antécedent est detérminé par la *théorie de contrôle*. Par exemple, dans la phrase $Paul_i$ *promised Mary* e_i*to take care of himself*$_i$, *himself* renvoie au pronom sujet PRO (e) qui à son tour indique le SN *Paul* comme antécédent.
   (c) **Pronom personnel de 3ème personne.**
      i. Regarder les SN de la phrase précédente qui constituent des arguments. Nous partons de l'hypothèse que tous les antécédents sont des arguments.

5. Vu que les pronoms touchent à la cohérence du texte, nous n'avons pas voulu modifier l'ordre des phrases. Nous aurions pu choisir de sélectionner seulement les phrases avec les pronoms personnels par exemple.

ii. Faire une liste hiérarchique des arguments trouvés selon leurs fonctions grammaticales (le sujet, ensuite le complément direct et en dernier le complément indirect).

iii. Trouver un SN avec des traits d'accord correspondants dans la liste.

2. **Retenir l'antécédent.** Une fois qu'un SN avec les bons traits grammaticaux a été trouvé, l'analyseur le conserve en tant que référent ou antécédent du pronom concerné. Cette information est ensuite conservée lors du transfert et finalement utilisée pour la génération du pronom dans la langue cible, le français.

En résumé, nous assumons une interprétation simplifiée des principes A et B de la Théorie de Liage. Lorsqu'un pronom non impersonnel est trouvé, le système commence la discrimination des antécédents à l'intérieur de la phrase. À ce point là, le **Principe B** bloque le lien entre un SN gouvernant et un pronom qui se trouve à l'intérieur de la phrase. S'il s'agit d'un pronom réfléchi ou réciproque, c'est le **Principe A** qui s'applique et qui intervient. Pour les autres pronoms référentiels de 3ème personne, l'antécédent est recherché dans la phrase précédente à l'aide des traits d'accord et de la hiérarchie des arguments.

En ce qui concerne les pronoms sujet nuls (non réalisés lexicalement) de l'espagnol, nous adoptons le point de vue de la théorie générative qui postule la présence d'un pronom abstrait *pro*. Notre composant de RA s'applique à ce pronom *pro* lors que ce dernier est de 3ème personne, de la même manière qu'un pronom réalisé.

3.1 Évaluation de résultats obtenus

L'évaluation ci-dessous prend en considération tous les pronoms, y compris ces qui ne sont pas encore pris en compte par notre procédure de RA. Le Tableau 2 montre les résultats de traduction obtenus avant et après la mise en place du composant de RA. Nous avons considéré la traduction correcte quand le pronom était généré avec les traits grammaticaux correspondants à ceux de son antécédent ; autrement, la traduction était considérée incorrecte.

On peut apprécier une amélioration significative pour les traductions des pronoms personnels nuls (χ^2 (1, N = 156) = 28.59, p<.05), qui passent de 4.0% des traductions correctes à 17.6%. En effet, la correcte identification de l'antécédent des pronoms nuls s'est élevé de 14.1% à 47.4% (11/78 avant et 37/78 après). Pour les autres types de pronoms, les chiffres sont stables, vu que, comme mentionné au début de cette section, notre implémentation ne touche que les pronoms personnels.

Toutefois, on observe aussi que le nombre des traductions correctes des pronoms relatifs a diminué sensiblement (χ^2 (1, N = 86) = 0.32, p<.05). Ces chiffres inattendus sont dus à l'insertion des pronoms lors du transfert – une erreur qui devra être corrigée –, comme l'illustre l'exemple (3). L'introduction d'un pronom personnel additionnel dans la traduction du verbe a empêché la génération du pronom relatif correct. Autrement dit, vu la présence d'un pronom sujet, le système a généré le pronom relatif avec le cas accusatif (*que*) au lieu du pronom avec le cas nominatif (*qui*).

Pronom	Sans RA		Avec RA	
	Correct	Incorrect	Correct	Incorrect
Personnel				
Nul	4.0	30.2	17.6	16.7
OD	2.6	1.3	2.6	1.3
OI	1.3	0.9	1.3	0.9
Relatif	26.8	9.2	23.3	12.8
Possessif	14.9	4.0	13.2	5.7
Démonstratif	2.2	0.0	2.2	0.0
Indéfini	0.9	1.8	1.4	1.3

TABLE 2 – Comparaison de résultats de traduction avant et après le composant de RA (%).

(3) a. *Espagnol* Impedían ayer acercarse a la zona a los grupos radicales, **que** intentaban volver a bloquear el acceso a la conferencia de la OMC como el primer día.

b. *Traduction ITS* Empêchait hier s'approcher de la zone aux groupes radicaux, **ils que** tentaient de se remettre à bloquer l'accès à la conférence de l'OMC.

c. *Référence* Hier, ils empêchaient aux groupes radicaux **qui** tentaient de bloquer l'accès à nouveau à la conférence de l'OMC de s'approcher.

Pour ce qui est des pronoms possessifs, malgré une analyse syntaxique toujours correcte, leur traduction ne l'est pas. Effectivement, tant en espagnol comme en français, l'accord grammatical est fait à la fois selon le possesseur et selon ce qui est possédé. Pourtant, le français utilise différentes formes pronominales pour toutes les combinaisons possesseur-possédé, alors qu'en espagnol il n'y pas de différence pour la troisième personne. Dans d'autres mots, c'est la même forme tant pour un possesseur singulier que pluriel (4). À l'heure actuelle, le composant de RA ne prend pas en considération ce problème.

(4) a. *Espagnol* Los grupos satánicos españoles utilizan cada vez más Internet para difundir **sus** ideales.

b. *Traduction ITS* Les groupes sataniques espagnols utilisent de plus en plus internet pour diffuser **ses** idéaux.

c. *Référence* Les groupes sataniques espagnols utilisent de plus en plus Internet pour diffuser **leurs** idéaux.

4 Conclusion et perspectives de recherche

Dans ce papier nous avons montré l'état de notre recherche pour le traitement des pronoms avec les système ITS-2. Après avoir implémenté un composant de RA qui reprend partiellement les principes de la théorie de liage, nous avons mené une évaluation pour la paire de langues

espagnol-français. En accord avec les principes de la théorie du liage, la traduction des pronoms personnels de sujet s'est améliorée avec le composant.

Pourtant, d'autres pistes de recherche devront être poursuivies pour un traitement adéquat des autres catégories de pronoms. Effectivement, tel qu'il est montré dans le tableau 1, l'ensemble des pronoms relatifs, possessifs, démonstratifs et indéfinis constituent 60.1% des pronoms dans le corpus. Cela représente un nombre non négligeable de pronoms qui ne sont pas pris en considération par la définition stricte d'anaphore selon la théorie du liage.

En ce sens, il existe une autre théorie linguistique pertinente pour l'interprétation anaphorique dans une perspective discursive, la SDRT. Selon cette théorie, la signification transférée par les phrases est le produit de l'interaction de plusieurs relations discursives, par exemple, *Narration, Élaboration, Explication, Contexte, Évidence, Conséquence* et *Contraste*. Des référents de discours sont introduits par les phrases, et donc, sont également soumis à ces relations. De cette façon, les référents introduits par des pronoms peuvent être co-identifiés avec des référents du discours déjà accessibles, en résolvant ainsi les anaphores. L'accessibilité, pour sa part, est limitée par la structure discursive créée par les relations.

Même si cette théorie a été développé activement sur le plan théorique, les propositions pour son implémentation sont rares. À notre connaissance, Asher *et al.* (2004) est le seul travail qui essaie d'implémenter directement la SDRT. Nous pensons que cette théorie est une piste de recherche de grande valeur pour un traitement linguistique approprié et inclusif des pronoms indéfinis, démonstratifs, possessifs et relatifs.

Finalement, des travaux qui utilisent une architecture similaire à la nôtre, comme c'est le cas de Trouilleux (2002), ou qui ont un cadre théorique compatible avec notre chemin de recherche, comme c'est le cas de Bos (2008), seront pris en considération.

Références

ASHER, N., DENIS, P., KUHN, J., LARSON, E., MCCREADY, E., PALMER, A., REESE, B. et WANG, L. (2004). Extracting and using discourse structure to resolve anaphoric dependencies : Combining logico-semantic and statistical approaches. *In Actes de TALN'04, Workshop SDRT*, Fès.

BOS, J. (2008). Wide-coverage Semantic Analysis with Boxer. *In Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 277–286. Association for Computational Linguistics.

BRENNAN, S. E., FREIDMAN, M. W. et POLLARD, C. J. (1987). A centering approach to pronouns. *In Proceedings of the 25th Annual Meeting on Association for Computational Linguistics*, pages 155–162.

BROSCHEIT, S., POESIO, M., PONZETTO, S. P., RODRÍGUEZ, K. J., ROMANO, L., URYUPINA, O., VERSLEY, Y. et ZANOLI, R. (2010). Bart : A multilingual anaphora resolution system. *In Proceedings of the 5th International Workshop on Semantic Evaluation*.

BÜRING, D. (2005). *Binding Theory*. Cambridge University Press.

CHOMSKY, N. (1981). *Lectures on Government and Binding : The Pisa Lectures*. Mouton de Gruyter.

HARDMEIER, C. et FEDERICO, M. (2010). Modelling pronominal anaphora in statistical machine translation. *In Proceedings of the 7th International Workshop on Spoken Language Translation*, pages 283–289.

HOBBS, J. (1986). *Readings in Natural Language Processing*, chapitre Resolving Pronoun References. Morgan Kaufmann Publishers Inc.

KAMP, H. et REYLE, U. (1993). *From Discourse to Logic. Introduction to Modeltheorectic Semantics of Natrual Language, Formal Logic and Discourse Representation Theory*. Kluwer Academic Publishers.

LAPPIN, S. et LEASS, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

LASCARIDES, A. et ASHER, N. (2007). Segmented discourse representation theory : Dynamic semantics with discourse structure. *In Computing Meaning : Volume 3*, pages 87–124. Kluwer Academic Publishers.

LE NAGARD, R. et KOEHN, P. (2010). Aiding pronoun translation with coreference resolution. *In Proceedings of the Joint 5th Workshop on Statistical Machine Translation*, pages 258–267.

MANN, W. C. et THOMPSON, S. A. (1988). Rhetorical structure theory : Towards a functional theory of text organization. *Text*, 8(3):243–281.

MITKOV, R., EVANS, R. et ORASAN, C. (2002). A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. *In Proceedings of the CICLing-2000*.

NERIMA, L. et WEHRLI, E. (2013). Résolution d'anaphores appliquée aux collocations : une évaluation préliminaire. *In Actes de TALN'13*, Les Sables d'Olon.

RECASENS, M. et MARTÍ, M. A. (2010). Ancora-Co : Coreferentially annotated corpora for Spanish and Catalan. *Language Resources and Evaluation*, 44(4):315–345.

REINHART, T. (1983). *Anaphora Resolution and Semantic Interpretation*. Croom Helm.

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.

TAULÉ, M., MARTÍ, M. A. et RECASENS, M. (2008). Ancora : Multilevel annotated corpora for Catalan and Spanish. *In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

TROUILLEUX, F. (2002). A Rule-based Pronoun Resolution System for French. *In Proceedings of Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002)*, Lisbon, Portugal. Benjamins.

WEHRLI, E. (2007). Fips, a "deep" linguistic multilingual parser. *In Proceedings of the Workshop on Deep Linguistic Processing*, pages 120–127. Association for Computational Linguistics.

WEHRLI, E., NERIMA, L. et SCHERRER, Y. (2009). Deep linguistic multilingual translation and bilingual dictionaries. *In Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 90–94.