Résolution d'anaphores appliquée aux collocations: une évaluation préliminaire

Luka Nerima Eric Wehrli
LATL, Université de Genève, 2 rue de Candolle, 1211 Genève 4
luka.nerima@unige.ch, eric.wehrli@unige.ch

RÉSUMÉ

Le traitement des collocations en analyse et en traduction est depuis de nombreuses années au centre de nos intérêts de recherche. L'analyseur Fips a été récemment enrichi d'un module de résolution d'anaphores. Dans cet article nous décrivons comment la résolution d'anaphores a été appliquée à l'identification des collocations et comment cela permet à l'analyseur de repérer une collocation même si un de ses termes a été pronominalisé. Nous décrivons aussi la méthodologie de l'évaluation, notamment la préparation des données pour le calcul du rappel. Dans la tâche d'identification des collocations pronominalisées, Fips montre des résultats très encourageants : la précision mesurée est de 98% alors que le rappel est proche de 50%. Dans cette évaluation nous nous intéressons aux collocations de type verbe-objet direct en conjonction avec les pronoms anaphoriques à la 3e personne. Le corpus utilisé est un corpus anglais d'environ dix millions de mots.

ABSTRACT

Anaphora Resolution Applied to Collocation Identification: A Preliminary Evaluation

Collocation identification and collocation translation have been at the center of our research interests for several years. Recently, the Fips parser has been enriched by an anaphora resolution mechanism. This article discusses how anaphora resolution has been applied to the collocation identification task, and how it enables the parser to identify a collocation when one of its terms is pronominalized. We also describe the evaluation methodology, in particular the preparation of data for the calculation of the recall. In the task of pronominalized collocation identification, Fips shows encouraging results: the measured precision is 98% while recall approaches 50%. In this paper we focus on collocations of the type verb-direct object and on a widespread type of anaphora: the third personal pronouns. The corpus used is a corpus of approximately ten million English words.

MOTS-CLÉS : Analyse, résolution d'anaphores, pronoms personnels, collocations, corpus
KEYWORDS : Parsing, anaphora resolution, personal pronoun, collocations, corpus

1 Introduction

Tant le traitement des pronoms anaphoriques que celui des collocations sont considérés comme des problèmes majeurs en traitement automatique des langues en général et en traduction automatique en particulier. De très nombreux travaux ont été consacrés à ces deux thèmes (voir en particulier Mitkov, 2002, ou Poesio et al. 2010, pour la résolution d'anaphores, Seretan, 2011, pour l'identification de collocations), mais à notre connaissance,

rares sont les recherches qui ont porté sur l'intersection de ces deux domaines, à savoir le traitement de collocations dans lesquelles un des deux termes de la collocation est un pronom anaphorique. Dans ce papier, nous présentons une recherche en cours sur les collocations de type verbe-objet direct en anglais, où l'objet est un pronom anaphorique, comme dans l'exemple *Paul will break it* lorsque le pronom anaphorique *it* renvoie au mot *record*, ce qui nous donne une occurrence de la collocation *break-record* (*battre-record*). Le processus d'identification des collocations a été décrit à plusieurs reprises (voir en particulier Wehrli & al. 2010) et ne sera pas repris dans cet article.

2 Résolution d'anaphores

Comme premier pas en direction d'un traitement des anaphores, nous avons développé une procédure qui permet à l'analyseur Fips (Wehrli, 2007) de traiter les pronoms personnels de 3e personne, de loin le type le plus fréquent d'anaphores. Selon Tutin (2002), les pronoms personnels constituent entre 60 et 80% des expressions anaphoriques relevées dans un large corpus du français. Russo et al. (2011) rapportent des résultats assez semblables pour l'anglais, l'italien, l'allemand et le français.

Notons, par ailleurs, que notre traitement des pronoms anaphorique ne prend en considération que les pronoms de troisième personne dont l'antécédent est dans la phrase ou dans la phrase précédente. Selon Laurent (2001), ces deux cas représentent près de 90% des pronoms anaphoriques. La procédure de résolution d'anaphores (RA) reprend dans les grandes lignes celle présentée par Lappin et Leass (1994), adaptée aux spécificités de notre grammaire et de nos représentations.

La première tâche de la procédure de RA consiste à distinguer parmi tous les pronoms de 3e personne les occurrences anaphoriques des occurrences non-anaphoriques, telles que l'usage impersonnel du pronom *it*, comme dans les exemples suivants :

(1) a. It is raining

b. It turned out that Bill was lying.

c. To put it lightly.

d. It is said that they have been cheated.

C'est essentiellement sur la base des informations lexicales (p. ex. verbes "météorologiques") et des informations grammaticales (p. ex. structure impersonnelle) que l'identification des emplois impersonnels du pronom *it* est réalisée.

L'étape suivante concerne les anaphores au sens strict de la théorie du liage de Chomsky (1981), qui stipule [principe A] que les pronoms réfléchis et réciproques doivent être liés dans leur catégorie gouvernante. Notre interprétation quelque peu simplifiée de ce principe est d'exiger que les pronoms réfléchis et réciproques renvoient au sujet de la proposition minimale qui contient le pronom.

Enfin, dans la 3e étape de notre procédure, nous considérons les pronoms référentiels tels que (*he, him, it, she, her, them,* etc.). Nous fondant à nouveau sur les inspirations de la théorie du liage [principe B], nous considérons qu'un pronom ne peut pas renvoyer à un antécédent à l'intérieur de sa proposition minimale. C'est ainsi que *him* dans l'exemple (2)

ci-dessous ne peut pas renvoyer à *Paul*.

(2) *Paul$_i$ likes him$_i$

L'exemple 3 est intéressant car la coréférence de *her* et *Mary* y est impossible, mais pas celle de *him* et *Paul*.

(3) Paul persuaded Mary to talk to her / him

Ce contraste s'explique aisément si l'on se souvient que dans l'analyse chomskyenne, les compléments infinitifs sont des propositions dotés d'un sujet abstrait soumis au processus de contrôle. Dans notre exemple, les propriétés lexicales du verbe *persuade* établissent que le contrôleur du sujet vide de la proposition enchâssée est l'argument objet du verbe *persuade*. Autrement dit, la structure est la suivante :

(4) Paul persuaded Mary [S [NP e] [VP to talk [PP to [NP her / him]]]]

Him et *her* ne peuvent renvoyer au sujet de la proposition infinitive, lui-même lié (contrôlé) par *Mary*. Cela exclut la coréférence entre *her* et *Mary*. Par contre, rien dans cette structure n'empêche une lecture coréférentielle de *him* et *Paul*.

Notons, enfin, que la théorie du liage (ou notre interprétation simplifiée de cette dernière) ne constitue pas une méthode de résolution d'anaphore à proprement parler. En effet, elle ne dit pas quel est l'antécédent d'un pronom, mais uniquement quels sont les candidats potentiels qui doivent être exclus pour violation d'un des principes de la théorie.

Notre procédure de RA, tout comme la procédure de Lappin et Leass, constitue une liste de candidats - dans notre cas, les syntagmes nominaux arguments - et lorsqu'un pronom est rencontré, sélectionne dans cette liste le "meilleur" candidat, sur la base (i) des règles d'accord (nombre, genre), (ii) des principes du liage (qui excluent certains candidats) et (iii) d'une heuristique inspirée de la Centering Theory (cf. Grosz et al. 1995; Kibble, 2001). Selon cette heuristique, la préférence est donnée au premier argument sujet et en second lieu, à l'argument non-sujet le plus proche.

3 Evaluation

Dans ce travail, nous nous intéressons à évaluer la performance de l'analyseur Fips à identifier dans un corpus des collocations de type verbe-objet direct dont l'objet a été pronominalisé. Nous nous sommes focalisés sur les deux phénomènes les plus fréquents, illustrés par les phrases suivantes construites à partir de la collocation *dépenser de l'argent* :

(5) a. Je vous ai donné de l'*argent*, vous pouvez *le* dépenser.

b. L'argent est là. Alors pourquoi n'a-t-il pas été dépensé ?

Dans la phrase (5a.) le mot *argent* est repris par le pronom *le* et joue le rôle d'objet de la collocation. Dans l'exemple (5b.), la collocation est au passif et le pronom sujet *il* correspond à l'objet direct « profond » du verbe *dépenser*. A noter que le référent anaphorique ne se trouve pas forcément dans la phrase elle-même mais peut se trouver dans une phrase voisine, la précédente dans la plupart des cas. Pour l'instant, Fips ne traite que les pronoms anaphoriques dont l'antécédent se trouve dans la phrase elle-même ou dans la phrase précédente.

3.1 Expérimentation

L'analyseur Fips dispose d'une base de données lexicale comprenant pour chaque langue un lexique de collocations. Pour le français, par exemple, ce lexique contient environ 16'000 entrées et pour l'anglais environ 9'000. Dans cette étude et dans la suite de l'article, nous ne considérons que ces collocations, c'est-à-dire celles qui sont lexicalisées. Dans cette expérience, nous nous intéressons à mesurer la performance (précision et rappel) de Fips dans la tâche d'identification des collocations qui sont à la fois (1) de type verbe-object direct, (2) lexicalisées, et (3) dont l'objet a été pronominalisé. Dans ce travail nous nous sommes limités à l'anglais

3.2 Corpus et méthodologie d'évaluation

Le corpus utilisé pour cette évaluation est constitué d'environ 10'000 articles parus dans le journal « The Economist » entre les années 2003 à 2010, totalisant environ 10 Mio de mots. Nous avons utilisé l'outil FipsCoView (Seretan & Wehrli, 2011) basé sur l'analyseur Fips pour extraire les collocations. Trente et une collocations (type) et quarante huit occurrences (token) répondant aux critères décrits dans la section précédente ont été repérées par Fips. Nous avons déterminé manuellement la précision de cette extraction.

Pour le rappel nous avons procédé comme suit : nous avons retenu les 18 collocations les plus fréquentes (parmi les 31) et à l'aide d'expressions régulières assez simples[1] nous avons recherché toutes les occurrences pronominalisées des 18 collocations et extrait les phrases susceptibles de les contenir. Le résultat de cette extraction a ensuite été filtré à la main par un annotateur[2]. Nous avons ainsi obtenu une cinquantaine de phrases constituant le corpus de référence (ou paires de phrases dans le cas où l'antécédent se trouve dans la phrase précédente).

Voici quelques exemples de phrases illustrant les phénomènes de pronominalisation les plus fréquents, tirées du corpus de référence:

(6) *to spend money* :

a. The explosion of the IT business and its offshoots has helped produce a new breed of young professionals with *money* in their pockets and their own ideas on how to *spend it*.
b. Lots of EU *money* is flowing to Poland and the rest. *It* must be *spent* fast.

to solve a problem :

c. Africa has, to put it mildly, a lot of *problems*; even a hyperpower cannot *solve them* all.

[1] Les expressions régulières (ER) recherchent à l'intérieur d'une phrase ou dans deux phrases contiguës la présence de trois éléments lexicaux : le verbe et l'objet de la collocation ainsi qu'un pronom référentiel. L'ER prend en compte toutes les formes fléchies des ces trois éléments lexicaux. Par exemple pour le verbe *to spend* elle accepte les formes: *spend, spends, spending* et *spent*. L'ER impose aussi que l'antécédent de l'objet direct apparaisse avant le verbe.

[2] Comme nous ne prenons en compte que des collocations lexicalisées, c'est-à-dire validées par un lexicographe au moment de leur insertion dans notre lexique, la tâche de juger une collocation pronominalisée est relativement simple. Il ne nous a dès lors pas paru nécessaire d'effectuer ce filtrage par plusieurs annotateurs et de comparer leurs jugements.

to make a decision :

d. But this time, the *decision* seems genuine: even senior party members appeared astonished at the announcement, and Dr Mahathir himself wept as he *made it*.

L'exemple (6a) montre le cas de la pronominalisation de l'objet par *it*, (b) une collocation au passif et dont l'antécédent se trouve dans la phrase précédente, (c) l'objet est un pronom au pluriel, (d) l'antécédent est éloigné du pronom, 20 mots les séparent.

3.3 Résultats

A noter que par commodité, nous avons refait une analyse avec Fips sur le corpus de référence pour déterminer le rappel. En terme de précision et de rappel de l'analyseur Fips, nous obtenons les résultats reportés dans la Table 1. Les résultats sont donnés séparément pour chacun des deux types de pronominalisation, objet direct pronominalisé et collocation au passif. Le nombre de phrases du corpus de référence est indiqué entre parenthèse dans la première colonne :

Pronominalisation	Précision	Rappel
Objet pronominalisé (40)	97	35
Collocation au passif (12)	100	100
Les deux (52)	98	48

TABLE 1 – Précision et rappel de l'identification des collocations pronominalisées (en %)

La précision est excellente. Nous l'expliquons par le fait que les contraintes sont tellement fortes pour résoudre l'anaphore et pour repérer une collocation que le risque d'erreur est très faible. Il faudrait en effet que Fips calcule un antécédent erroné mais, qui combiné avec le verbe, donne une collocation qui existe dans notre lexique. La probabilité est très faible mais notre évaluation a mis en évidence que ce cas s'est produit une fois, avec l'analyse de la phrase (7) ci dessous :

(7) Concerted international **pressure* then forced it to confess to 18 years of lies. Yet there are already troubling signs that, at a meeting of the IAEA's governing board this week, some governments will be tempted, as America's Colin Powell *puts it*, to declare premature victory.

Le nom *pressure* a été choisi comme étant l'antécédent du pronom *it* (*Colin Powell puts it*) et la collocation *to put pressure* a été identifiée erronément.

Le rappel est plus modeste mais il faut se rappeler que la RA est une tâche très difficile : de nombreux phénomènes linguistiques viennent perturber la résolution comme par exemple les conjonctions de coordination. Dans l'exemple (6a), c'est la coordination *and* qui a empêché la remontée jusqu'à l'antécédent *money* et qui a fait échouer l'identification de la collocation *to spend money*.

On remarquera aussi que le nombre de collocations verbe-objet direct pronominalisées

semble assez faible dans le corpus. Cela suggère que cette configuration se produit rarement. Nous avons aussi observé que certaines collocations sont moins sujettes à la forme passive. Enfin, il faut aussi se rappeler que seulement 18 collocations (types de collocation) ont été recherchées.

4 Conclusions

Dans cet article, nous avons présenté l'application de la résolution d'anaphores à l'identification des collocations par l'analyseur de Fips. Nous nous sommes focalisé sur les collocations de type verbe-objet direct et aux pronoms personnels de la 3e personne. Même si ces deux phénomènes linguistiques réunis ensemble se sont avérés relativement peu fréquents dans le corpus choisi, ils méritent d'être analysés avec soin surtout en situation de traduction : si la collocation n'est pas identifiée, la traduction sera mauvaise voire même incompréhensible. La précision mesurée de l'analyseur Fips dans cet exercice d'identification est très bonne (98%) et le rappel honorable (48%).

La méthodologie pour calculer le rappel s'est avérée très utile : en l'absence de corpus annoté, pouvoir produire un corpus de référence à moindre frais est appréciable. En affinant la méthode nous espérons aussi réduire l'ampleur du nettoyage manuel. Cela nous permettra de mener des évaluations sur le repérage d'un plus grand nombre de collocations. Cela nous aidera aussi à produire, à partir de corpus réels, des données de test pour mettre au point les améliorations de l'analyseur Fips.

Un autre axe pour nos travaux futurs sera de prendre en compte d'autres types de RA et d'appliquer les RA à d'autres types de collocations, par exemple sujet-verbe, verbe-groupe prépositionnel, etc.

Références

CHOMSKY, N. (1981). Lectures on Government and Binding, Foris Publications.

GROSZ, B., A. JOSHI, A. & WEINSTEIN, S. (1995). Centering: A Framework for Modeling the Local Coherence of Discourse, Computation Linguistics, 21:2, 203-225.

KIBBLE, R. (2001). A Reformulation of Rule 2 of Centering Theory", in *Computational Linguistics*, 27:4, Cambridge, Mass., MIT Press.

LAPIN, S., LEASS, J.L. (1994). An Algorithm for Pronominal Anaphora Resolution, *Computational Linguistics*, 20:4, 535-561.

LAURENT, D. (2001). De la résolution des anaphores. Rapport interne, Synapse Développement. Disponible en ligne sur http://www.synapse-fr.com/descr_technique/Resolution_des_anaphores.pdf

MITKOV, R. (2002). *Anaphora Resolution*, Longman.

POESIO, M., PONZETTO, S. et VERSLEY, Y. (2011). Computational Models Of Anaphora Resolution : A Survey. Disponible en ligne sur http://clic.cimec.unitn.it/massimo/Publications/lilt.pdf

RUSSO, L., Y. SCHERRER, J.-PH. GOLDMAN, S. LOAICIDA, L. NERIMA, E. WEHRLI (2011). Etude inter-langues de la distribution et des ambiguïtés syntaxiques des pronoms. *In Actes de TALN-*

2011, Montpellier.

SERETAN, V., WEHRLI, E. (2011). FipsCoView: On-line Visualisation of Collocations Extracted from Multilingual Parallel Corpora, *In Proceedings of the ACL Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, 125-127.

SERETAN, V. (2011). *Syntax-Based Collocation Extraction*, Springer Verlag.

TUTIN, A. (2002). A Corpus-based Study of Pronominal Anaphoric Expressions in French. In *Proceedings of DAARC 2002*, Lisbonne, Portugal.

WEHRLI, E. (2007). Fips, a "deep" linguistic multilingualparser. *In Proceedings of the ACL 2007 Workshop on Deep Linguistic processing*, pp. 120-127, Prague, Czech Republic.

WEHRLI, E., SERETAN, V., et NERIMA, L. (2010). Sentence Analysis and Collocation Identification. *In Proceedings of the Workshop on Multiword Expressions, Coling-2010*, Beijing, pp. 27-35.