Repérage de citations, classification des styles de discours rapporté et identification des constituants citationnels en écrits journalistiques

Fabien Poulard, Thierry Waszak, Nicolas Hernandez, Patrice Bellot

TALN 2008, Avignon, 9–13 juin 2008

Repérage de citations, classification des styles de discours rapporté et identification des constituants citationnels en écrits journalistiques

Fabien Poulard[1] Thierry Waszak[2] Nicolas Hernandez [1] Patrice Bellot[2]
(1) LINA UMR 6241 / Université de Nantes
(2) LIA / Université d'Avignon
{fabien.poulard,nicolas.hernandez}@univ-nantes.fr,
{thierry.waszak,patrice.bellot}@univ-avignon.fr

Résumé. Dans le contexte de la recherche de plagiat, le repérage de citations et de ses constituants est primordial puisqu'il peut amener à évaluer le caractère licite ou illicite d'une reprise (source citée ou non). Nous proposons ici une comparaison de méthodes automatiques pour le repérage de ces informations et rapportons une évaluation quantitative de celles-ci. Un corpus d'écrits journalistiques français a été manuellement annoté pour nous servir de base d'apprentissage et de test.

Abstract. In the application context of reported content, that includes plagiarism and impact of textual information searched, citations finding and its fundamentals is essential as it may help estimating legal value of a citation (with or without specifying original source). We propose here a comparison between automatic methods for finding up those elements and we quantitatively evaluate them. A French journalistic corpus has been manually annotated to be used as learning base and for testing.

Mots-clés : détection de citations, classification des styles de discours rapporté, identification du locuteur, techniques par apprentissage et base de règles, écrits journalistiques.

Keywords: detection of citations, reported speech style classification, source identification, machine learning and rules-based techniques, news corpus.

1 Introduction

« Qui dit quoi sur qui (quoi) et de quelle manière ? » L'intérêt pour cette question, et plus largement les informations de type *attribution*, n'est pas nouveau et l'on constate même un certain regain depuis quelques années avec les travaux cherchant à mesurer la subjectivité (opinion, sentiment, émotion) engagée dans l'attribution.

Cet intérêt se retrouve dans plusieurs disciplines (linguistique, sociologie, traitement automatique des langues, professions de la documentation...). Pour certains il s'agit de mesurer l'impact du travail d'un chercheur (mesures bibliométriques) (Hirsch, 2005), pour d'autres d'améliorer les systèmes existants avec de nouvelles formes de résumé (orienté selon une opinion) (Stoyanov & Cardie, 2006), de recherche d'information (indexation des documents par les ci-

tations) (Ritchie *et al.*, 2006), de question-réponse (« Que pense X au sujet de Y ? ») (Somasundaran *et al.*, 2007). Finalement pour d'autres encore il s'agit de collecter des opinions sur différents sujets (Wilson *et al.*, 2005), potentiellement à des fins commerciales (suivi de l'impact d'un produit) (Dave *et al.*, 2003). Ces travaux couvrent plusieurs tâches : du repérage de la présence d'attribution à l'identification de l'étendue de ses constituants en passant par la mesure de sa fonction rhétorique, de sa subjectivité, de sa polarité et de son contexte. Néanmoins, cette couverture est inégale selon la langue, le genre de texte et la motivation applicative.

Le contexte applicatif du présent article est celui de la détection de plagiats et du suivi d'impact à partir d'un écrit original, dans des textes journalistiques francophones. Ce travail s'inscrit dans le cadre du projet ANR PIITHIE[1] (Plagiats et Impacts de l'Information Textuelle recHerchée dans un contexte InterlinguE). Dans cette perspective, le repérage de citations et de ses constituants (l'identification de la source et son propos) est primordial puisqu'il peut permettre l'évaluation du caractère licite ou illicite d'une reprise. En effet, la détection de plagiat se fait notamment par identification de similarités entre deux textes. La présence commune de citations a tendance à augmenter artificiellement la similarité entre deux textes qui ne sont pas plagiat l'un de l'autre. À l'inverse, en faisant l'hypothèse que les citations ne sont pas modifiées dans un texte plagié, les citations constituent de bons points de repères pour identifier de potentiels plagiats en analysant le texte environnant et seulement celui-ci.

Dans cet article, nous proposons des approches pour le repérage des citations, leur classification en discours direct (DD) ou discours indirect (DI), ainsi que pour l'identification des entités nommées dont on rapporte un discours, et les segments textuels entre guillemets porteurs de DD. L'intérêt pour les DD s'explique par le fait que l'identification d'une reprise *verbatim* facilite le suivi d'impact. À ces fins, nous avons construit et annoté un corpus d'écrits journalistiques.

1.1 Travaux existants en repérage d'attribution

Notre problématique diffère de celle qui a pu être traitée dans les travaux menés en anglais sur des textes scientifiques (Teufel *et al.*, 2006) où les sources sont clairement identifiées (l'auteur ou les références bibliographiques) et les formes de reprise différentes, puisque pour ce genre les citations constituent un positionnement de l'auteur écrivain qui nourrit son discours, alors que pour les écrits journalistiques, le discours rapporté constitue l'essence même de l'article.

Nos considérations seraient davantage à rapprocher des travaux en reconnaissance d'attribution et de capture de subjectivité réalisés sur les journaux et les blogs (Bethard *et al.*, 2004; Choi *et al.*, 2005; Stoyanov & Cardie, 2006; Somasundaran *et al.*, 2007). Nous nous différencions néanmoins sur plusieurs points. Outre le fait que ces travaux ont été dirigés sur de l'anglais, ils reposent avant tout sur des ressources lexicales importantes (pour mesurer la subjectivité), syntaxiques (arbre de dépendance ou fonction grammaticale) et sémantiques pour les travaux ayant cherché à identifier les syntagmes sources et les propositions porteuses d'un discours repris. En comparaison nos approches visent à exploiter essentiellement des marques de surfaces de nature typographique, morphologique et positionnelle. En cela, nous suivons l'approche décrite par (Giguet & Lucas, 2004). Ces marques sont néanmoins complétées par une ressource lexicale de verbes de parole produite par (Mourad & Desclés, 2004).

À notre connaissance, (Mourad & Desclés, 2004) et (Giguet & Lucas, 2004) sont les seuls

[1]Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche, projet PIITHIE portant la référence 2006 TLOG 013 03. www.piithie.com

travaux réalisés en français sur le repérage de citations. Ces deux techniques reposent sur des règles. Celle de (Mourad & Desclés, 2004) se fonde sur l'exploration contextuelle qui distingue les marques ayant un rôle d'embrayeur de celles jouant le rôle de confirmateur. Le modèle de (Giguet & Lucas, 2004) s'appuie dans un premier temps sur la co-présence de marques pour reconnaître les constituants source, relateur ou discours, puis sur la reconnaissance du motif SRD (*source + relateur + discours rapporté*) ou de son inverse, DRS, pour repérer une citation. Par la suite, nous nous intéressons à l'expression linguistique référant à la source que nous appelons « expression locuteur ». Nous étendons ce modèle en ne posant pas de contraintes sur les motifs possibles et en considérant diverses techniques de reconnaissance notamment par apprentissage supervisé. Par ailleurs nous apportons une évaluation quantitative de nos méthodes.

1.2 Difficulté des tâches de repérage de citations et de ses constituants

Les problèmes du repérage sont complexes puisqu'une citation peut se caractériser par différents types d'information. Dans sa typologie, (Jackiewicz, 2006), décrit les citations en fonction du type d'information accompagnant le discours repris (paramètres de la situation d'énonciation (par qui, quand, où et pour qui), intentions du locuteur d'origine, informations à l'attention de celui à qui le discours est rapporté). (Giguet & Lucas, 2004) parlent d'invariants prototypiques et considèrent en plus du discours repris et de sa source, l'élément relateur qui marque le lien entre ces derniers. Dans un contexte d'analyse de la subjectivité, (Prasad *et al.*, 2006) rendent compte d'un important travail sur l'annotation des attributions en fonction de leur source, leur degré de factualité, leur polarité et leur contexte d'existence. Notre définition d'expression locuteur se rapproche des cas de sources indirectes qu'ils considèrent.

Outre ces différents types d'informations, il existe plusieurs manières[2] de restituer un discours en provenance d'une situation d'énonciation initiale. En effet, les reprises de discours peuvent prendre différentes formes d'intégration, du style direct (reprise littérale traditionnellement marquée par des guillemets (cf. exemples 1 et 2) au style indirect (discours intégré dans l'énonciation du reprenant (cf. exemples 3) en passant par des formes intermédiaires avec des îlots de reprises verbatims (que nous appelons style « à composantes ») (cf. exemple 4). De surcroît, outre le fait qu'il puisse être difficile de délimiter un discours repris, celui-ci peut être fragmenté (cf. exemple 2) voire étalé sur plusieurs phrases. À cela on peut encore ajouter le fait que les informations caractérisant une citation peuvent prendre différentes formes morpho-syntaxiques : e.g. relateur verbal ou prépositionnel (cf. exemples 1 et 3), source marquée par un pronom ou une entité nommée (cf. exemples 2 et 4), se combiner selon différentes configurations syntaxiques (cf. exemples 1 et 2), ou se retrouver distribuées dans une fenêtre transphrastique.

(1) **Le quotidien économique** *souligne :* "Si le rapport ne veut pas associer ces montants à l'idée d'une nouvelle 'cagnotte' budgétaire, ni au débat électoral sur le niveau de prélèvements obligatoires, le montant est équivalent au déficit budgétaire de l'État, à savoir 36,5 milliards d'euros l'an dernier."
(2) "En 2003 *, explique-t-***il**, j'ai fait effectuer douze tests sur des vols en France, dans onze cas sur douze, des armes et explosifs ont pu être introduits dans les avions".
(3) *D'après* **sa mère**, Julien faisait une sieste dans l'appartement familial lorsqu'il a disparu.
(4) **Le Figaro** *estime*, **lui**, *que* "les techniciens et les cadres sont en première ligne", notamment ceux de la "Central Entity" de Toulouse.

[2]Le groupe de recherche *Ci-dit* rend compte de nombreuses descriptions exemplifiées de ces formes en corpus de presse et de littérature. `www.ci-dit.com`

2 Corpus

Nous avons constitué un corpus avec trois objectifs en tête. Tout d'abord, pour obtenir des actes langagiers réels illustrant les différents concepts décrits à la section précédente. Ensuite, pour servir de support d'apprentissage supervisé. Finalement, pour nous permettre d'évaluer les performances de nos méthodes. Nous présentons dans cette section la constitution et l'organisation du corpus dans un premier temps, puis nous décrivons statistiquement sa composition.

Notre corpus est constitué d'articles extraits des éditions en ligne de journaux francophones (La Tribune, Challenges, Le Soir, Le Figaro, Libération, L'Humanité, Le Monde, AFP, Reuters). Tous ces articles ont été collectés séparément dans un premier temps et finalement fusionnés en un seul corpus. Le contenu de chacun des 108 articles (env. 70 000 mots) a été stocké sous forme d'un arbre XML dans un fichier, accompagné de métadonnées sur le nom du journal, la date de publication et les auteurs. Nous avons annoté manuellement, en nous inspirant du schéma SRD de (Giguet & Lucas, 2004), 846 expressions locuteur et 938 discours repris, en précisant pour ces derniers le style de discours rapporté associé : direct, indirect ou à composantes. On trouve en moyenne 9 discours repris par article, avec un minimum de 5,4 pour les articles du quotidien "Le Soir". Si la taille des discours repris est d'environ 25 mots, le rapport entre les mots appartenant à du discours repris par rapport au discours englobant varie de 17% pour Challenges à près de 70% pour Reuters.

Les styles de discours rapporté modifiant peu le discours source sont les plus fortement utilisés. En effet, près de 80% des discours repris que nous avons annotés sont rapportés au style direct ou à composantes. D'une manière générale, les expressions locuteurs sont moins nombreuses que les discours repris. Toutefois, 9 discours repris sur 10 sont accompagnés d'une expression locuteur. Les journalistes s'efforcent donc d'accompagner les discours qu'ils rapportent d'une expression faisant référence à l'énonciateur source. (Poulard, 2008) décrit quantitativement et qualitativement un sous-ensemble du corpus avec plus de précision.

3 Propositions de méthodes de repérage

Nous avons utilisé le corpus décrit dans la section précédente pour expérimenter plusieurs approches que nous présentons dans cette section. Chaque méthode nécessite un prétraitement : étiquetage grammatical, lemmatisation[3] et découpage en phrases préservant les segments entre guillemets. Les deux premières méthodes présentées cherchent à repérer les segments de discours repris au style direct et indirect, à l'aide d'automates pour la première et statistiquement pour la seconde. La troisième méthode tente de repérer statistiquement les constituants citationnels à partir de candidats prédéterminés.

3.1 Classification des discours à l'aide d'automates à états finis

Notre but est de classer des segments de texte comme étant ou non du discours repris direct ou indirect. L'idée est de repérer les éléments du schéma SRD proposé par (Giguet & Lucas, 2004) et de les mettre en relation grâce à l'utilisation d'automates à états finis. Nous sommes ici dans une approche linguistique nécessitant l'écriture de règles représentées par des automates.

[3]Nous avons utilisé TreeTagger pour le découpage en mots, l'étiquetage et la lemmatisation.

Le principe des automates à états finis (*Finite State Machine* – FSM) étant d'accepter ou de rejeter des suites de symboles, nous pouvons, dans un premier temps, les utiliser pour identifier dans les textes des expressions locuteur, des relateurs ou des segments de discours repris potentiels. En effet, en reprenant l'exemple (1) de la section 1, l'expression locuteur « *Le quotidien économique* » (qui peut être décomposée comme étant un déterminant suivi d'un nom et d'un adjectif) pourra être identifiée à l'aide d'un FSM spécifique (les FSM se basant sur les étiquettes morpho-syntaxiques et sur les lemmes des mots). De même, le relateur « *souligne :* » est identifié ainsi que le segment de texte entre guillemets qui le suit (discours repris direct potentiel). Pour cette approche, il nous a donc fallu concevoir tous les FSM possibles pouvant représenter les expressions locuteur, relateurs et segments de discours repris potentiels, soit environ 900 FSM représentant 39 catégories différentes comme les expressions locuteur composées d'un déterminant suivi d'un nom, d'un nom propre ou les relateurs composés d'un auxiliaire suivi d'un verbe d'expression, ...

Afin d'identifier D comme faisant partie du schéma SRD, nous procédons en trois étapes : un premier niveau de FSM identifie toutes les expressions locuteur, les relateurs et les segments de discours repris potentiels ; puis un deuxième niveau associe les expressions locuteur avec les relateurs (exemple de la figure 1) ; finalement, un dernier niveau permet de compléter le schéma SRD. S'il existe un FSM permettant d'associer les éléments S, R et D, en considérant tous les ordres possibles, on décide qu'il s'agit réellement d'un discours repris. Précisons que le dernier niveau est composé de deux catégories de FSM, ceux identifiant le discours direct et ceux identifiant le discours indirect.

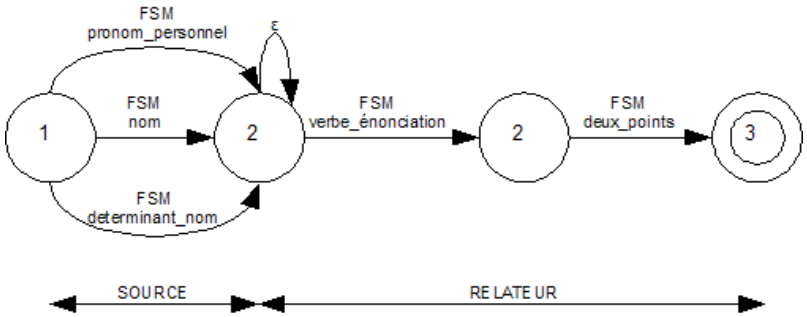

FIG. 1 – Exemple d'un FSM permettant d'associer une expression locuteur avec un relateur.

3.2 Classification des discours par apprentissage

En comparaison avec la précédente méthode, nous présentons ici, une approche statistique de repérage des segments de discours repris. Il s'agit d'effectuer un apprentissage à partir du corpus annoté manuellement pour aboutir à un modèle prédictif des classes : nous voulons classer les segments de texte comme étant soit du discours direct, soit du discours indirect, soit n'étant pas du discours repris.

Afin d'obtenir des segments de discours repris candidats, nous utilisons le premier niveau des FSM évoqué précédemment. Ces FSM nous permettent également de retrouver certains constituants (comme les verbes d'énonciation, les syntagmes prépositionnels, ...) qui vont permettre d'extraire les attributs pour l'apprentissage. En partant de nos observations et en se fondant sur (Mourad & Desclés, 2004), nous avons pu dégager les éléments caractéristiques à retenir pour

l'élaboration d'un système statistique. Ces caractéristiques correspondent à des marques lexicales, syntaxiques et typographiques. Ainsi, nous considérons par rapport au cadre phrastique contenant le discours repris candidat D deux types d'attributs : ceux recherchés à l'intérieur du cadre phrastique et ceux ne s'y limitant pas. Pour les premiers, nous considérons les distances en nombre de mots entre D et le plus proche verbe d'énonciation, entre D et le plus proche verbe d'énonciation de discours indirect (sous ensemble des verbes d'énonciation) et enfin entre D et le plus proche syntagme prépositionnel. Pour les seconds, nous recherchons la présence, dans les trois et neuf mots du contexte gauche ou droit de D, un indice du type : verbe, syntagme prépositionnel, conjonction *que* (attributs booléens). Nous utilisons également comme attributs les étiquettes morpho-syntaxiques associées à D (qui permettent notamment d'identifier le changement de temps des verbes dans D) ainsi que le nombre de mots composant D.

Les modèles ont été construits et testés à l'aide d'arbres de décision, de machines à support de vecteurs (SVM) et des réseaux Bayésiens.

3.3 Identification des constituants

Nous contournons la difficulté de déterminer les bornes des constituants *discours repris* et *expression locuteur* en identifiant des segments appartenant à l'un, puis à l'autre.

3.3.1 Appartenance au discours repris

Environ 80%[4] des discours repris dans notre corpus contiennent un passage entre guillemets, et 82% de ces derniers appartiennent à un discours repris. Nous faisons donc l'hypothèse que les segments entre guillemets sont candidats à appartenir à un *discours repris*. Ce choix se justifie par leur extraction aisée, en couplant les guillemets dans une pile par exemple. À partir des candidats extraits de notre corpus et caractérisés automatiquement selon les marques présentées ci-dessous, nous avons produit un modèle prédictif dont les performances sont présentées à la section 4.2.

Nous caractérisons automatiquement des segments entre guillemets dans leur contexte phrastique afin de les classer en deux catégories : ceux appartenant à du discours repris et les autres. Nous considérons uniquement des marques de surface ou qui peuvent être calculées aisément. Ainsi, nous prenons en compte les marques discursives classiques (pronoms et adjectifs possessifs aux premières personnes, verbes conjugués au sein du candidat), mais également les éléments marquant une rupture dans l'énonciation (les incises *[...]* ou *(...)*, guillemets au sein des candidats). Nous y ajoutons les *relateurs* de (Giguet & Lucas, 2004) : verbe + que, verbes d'énonciation (Mourad & Desclés, 2004). Finalement, nous avons considéré la taille en mots des candidats et la présence dans le contexte phrastique de pronoms à la troisième personne. Cette dernière marque permet de contre-balancer la présence des pronoms aux premières personnes dans une énonciation qui n'est pas à la troisième personne.

Les marques sus énoncées nous ont permis de caractériser automatiquement les 1004 segments entre guillemets de notre corpus, dont 820 appartiennent à un *discours repris*. La classification de ces candidats nous permet d'obtenir des segments appartenant à du discours repris.

[4] 508 discours direct et 259 discours à composantes pour 938 discours repris au total.

3.3.2 Appartenance aux expressions locuteur

Neuf discours repris sur dix sont accompagnés d'une expression locuteur dans notre corpus. Selon (Giguet & Lucas, 2004), les expressions locuteur apparaissent en premier lieu sous des formes très déterminées, reprises par la suite sous forme de "réductions lexicales", "d'anaphores". Nous faisons l'hypothèse que ces formes très déterminées sont constituées d'entités nommées, et choisissons ces dernières comme candidates. Nous utilisons Némesis (Fourour, 2004) pour leur extraction. À partir des candidats extraits de notre corpus annoté et caractérisés automatiquement selon les marques présentées ci-dessous, nous avons pu produire un modèle prédictif dont les performances sont présentées à la section 4.2.

Nous caractérisons les entités nommées afin de les classer en deux catégories : celles appartenant à une expression locuteur et les autres. Les marques retenues correspondent à la mise en place du scénario d'énonciation : syntagmes prépositionnels *Selon X*, *D'après X*, ..., proximité d'un verbe d'énonciation ou encore la taille en mot.

Les marques sus énoncées nous ont permis de caractériser automatiquement les 1703 entités nommées de notre corpus, dont 291[5] appartiennent à un *discours repris*.

4 Évaluations

Nous avons évalué les méthodes précédemment présentées par validation croisée sur notre corpus à l'aide de la plateforme Weka (Witten & Frank, 2005), en considérant les métriques classiques (précision, rappel, F-mesure). La première section présente l'évaluation des méthodes de repérage des segments de discours repris au style direct et indirect, et la deuxième section évalue la méthode de recherche des constituants.

4.1 Méthode pour le repérage du discours direct et indirect

Notons tout d'abord que pour le calcul de la précision et du rappel, dans le cas du discours direct, est considéré correct un segment de texte dont les frontières coïncident exactement avec le segment de référence. Dans le cas du discours indirect, nous choisissons de considérer un segment comme correct simplement s'il y a chevauchement avec le segment de référence. Cela nous permet notamment de mesurer la couverture du système et reflète la difficulté de segmentation du discours indirect dont nous mesurons la qualité en utilisant la mesure WindowDiff définie par (Pevzner & Hearst, 2002). Notons que cette mesure indique une segmentation optimale lorsqu'elle vaut 0.

Dans cette évaluation, nous considérons dans le corpus présenté dans la section 2 comme discours direct, les segments de discours repris direct mais aussi chaque segment de texte entre guillemets à l'intérieur des segments de discours repris « à composantes ». Ce qui représente 863 segments de discours direct et 171 segments de discours indirect.

Dans le cas de la première méthode à base de FSM, nous utilisons l'ensemble de ce corpus comme corpus de test ; les FSM ayant été conçu à partir d'un autre corpus de développe-

[5] À supposer que chaque expression locuteur compte au plus une entité nommée, ces dernières couvriraient 34% des expressions locuteur du corpus.

ment contenant 261 segments de discours direct et 82 segments de discours indirect. En ce qui concerne l'approche statistique, une validation croisée (à l'aide de Weka) a été effectuée sur le corpus. Les résultats sont présentés dans le tableau 1.

Méthode	Discours direct			Discours indirect		
	Précision	Rappel	F-Mesure	Précision	Rappel	F-Mesure
FSM	0,89	0,93	0,91	0,58	0,72	0,68
Réseaux Bayésiens	0,86	0,93	0,91	0,58	0,68	0,64
SVM	0,92	0,88	0,89	0,70	0,52	0,57
Arbres de décision	0,91	0,88	0,89	0,66	0,52	0,56

TAB. 1 – Évaluation des méthodes pour les discours directs et indirects

Nous obtenons des résultats très satisfaisants pour le discours direct. En ce qui concerne les 10% de pertes, notons que la majeure partie des segments de discours direct non rapportés ne peuvent pas être rattachés à une expression locuteur. Comme prévu, le repérage du discours indirect pose davantage de difficultés. Les résultats montrent toutefois que les systèmes à base de FSM et des réseaux Bayésiens sont capables d'identifier la présence d'environ 70% des citations indirectes. En ce qui concerne le WindowDiff pour la méthode à base de FSM nous obtenons une mesure de 0,23. Cela reflète le fait que la segmentation des discours indirects n'est pas optimale. En effet, elle a été faite en suivant des heuristiques : par exemple, on suppose qu'un segment de discours indirect commençant par la conjonction *que* se termine avec la phrase. Ce qui n'est pas toujours le cas. Les résultats obtenus par les FSM sont très proches de ceux des réseaux Bayésiens. On peut remarquer que les SVM, tout comme les arbres de décisions, permettent d'avoir une meilleure précision. Toutefois cela est obtenu au détriment du rappel, dont la chute est importante pour le discours indirect.

4.2 Méthode avec apprentissage supervisé, recherche des constituants

Nous utilisons une classification selon la classe majoritaire comme approche de base, le rappel et la précision pour la seconde classe est donc toujours nul. Nous avons fait ce choix faute de résultats expérimentaux dans la littérature. Nous présentons ci-dessous l'évaluation du modèle pour la classification des segments entre guillemets, puis l'évaluation de celui pour la classification des entités nommées en expression locuteur.

Pour rappel, nous considérons les segments entre guillemets comme sous-ensembles candidats de discours repris. Nous séparons les segments entre guillemets en deux classes : ceux qui font partie d'un discours repris, et ceux n'appartenant à aucun. La première classe est majoritaire, l'approche de base obtient une F-mesure de 0,899. Nous obtenons les meilleurs résultats avec l'algorithme *ADTree*[6] (*cf. Tableau 2*). Le modèle est alors capable de repérer 94% des segments entre guillemets appartenant à un discours repris avec une précision de 89,4%, soit un gain de 7,7 points par rapport à l'approche de base.

À l'instar des segments entre guillemets, nous considérons les entités nommées comme sous-ensembles candidats d'expressions locuteur. Nous les séparons en deux classes : celles prenant part à une expression locuteur, et les autres. Cette deuxième classe est majoritaire, l'approche de base offrant une précision de 82,9%. Nous nous intéressons aux entités nommées de la première

[6]Algorithme de type arbre de décision.

classe, les résultats les plus probants sont obtenus avec les réseaux Bayésiens (*cf. Tableau 2*). Nous récupérons alors près de 50% des entités nommées constituant les expressions locuteur, et parmis celles que nous sélectionnons, 45% appartiennent effectivement à des expressions locuteur. Les entités nommées représentent 34% des expressions locuteurs de notre corpus, nous pouvons donc extraire ainsi 17% des expressions locuteur du corpus.

Méthode	Classe *discours repris*			Classe *non discours repris*		
	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
Approche de base	0,817	1	0,899	0	0	0
Réseaux Bayésiens	0,893	0,884	0,888	0,505	0,527	0,516
SVM	0,817	1	0,899	0	0	0
Arbre (ADTree)	**0,894**	**0,94**	0,917	**0,655**	0,505	0,571
Méthode	Classe *expression locuteur*			Classe *non expression locuteur*		
	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
Approche de base	0	0	0	0,829	1	0,907
Réseaux Bayésiens	0,458	**0,474**	0,466	**0,891**	0,885	0,888
SVM	0	0	0	0,829	1	0,907
Arbre (J48)	0,576	0,247	0,346	0,861	0,962	0,909

TAB. 2 – Évaluation des modèles pour le repérage des constituants.

5 Conclusion et discussion

Dans cet article, nous avons proposé et évalué différentes méthodes dédiées à l'analyse du discours rapporté en corpus de presse. À cette fin, nous avons construit et annoté manuellement un corpus que nous souhaitons rendre public (discussion avec les éditeurs en cours).

Notre approche a montré des résultats significatifs. Comme nous l'avons déjà mentionné, à notre connaissance, il n'y a pas d'évaluation disponible pour le français pour les tâches de détection de citations, d'identification des composants citationnels, ou encore de classification des styles de discours rapporté (Mourad & Desclés, 2004), (Giguet and Lucas, 2004). En anglais, des travaux similaires ont été menés dans le contexte de l'analyse de la subjectivité. Néanmoins les résultats ne peuvent pas être comparés car le discours rapporté est un cas particulier d'attribution. De plus, nos intérêts diffèrent souvent. Alors que nous n'observons que les sources indirectes (les seules présentes dans les articles de presse), (Bethard *et al.*, 2004) s'intéressent à l'identification des sources directes (l'auteur), tandis que (Choi *et al.*, 2005) ne distinguent pas les deux.

En ce qui concerne nos travaux futurs sur l'analyse de citations, nous projetons de ne pas nous limiter aux entités nommés et de considérer aussi les pronoms personnels comme expression locuteur candidate. Nous pourrons par exemple suivre les travaux de (Stoyanov & Cardie, 2006) pour utiliser des méthodes de résolution d'anaphores appliquées à des cas d'attribution. Par ailleurs, nous souhaitons évaluer le recoupement entre les expressions que nous extrayons et celles annotées soit par la mesure WindowDiff soit par une technique proposée par (Choi *et al.*, 2005). Enfin, nous chercherons à adapter nos techniques afin d'être à même de pouvoir traiter des corpus de presse anglophones.

Remerciements

Nous remercions Annie Tartier et les relecteurs pour leurs conseils éclairés.

Références

BETHARD S., YU H., THORNTON A., HATZIVASSILOGLOU V. & JURAFSKY D. (2004). Automatic extraction of opinion propositions and their holders. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text : Theories and Applications*.

CHOI Y., CARDIE C., RILOFF E. & PATWARDHAN S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. In *HLT - EMNLP*, Vancouver, Canada.

DAVE K., LAWRENCE S. & PENNOCK D. M. (2003). Mining the peanut gallery : Opinion extraction and semantic classification of product reviews. In *Twelfth International World Wide Web Conference (WWW'03)*.

FOUROUR N. (2004). *Identification et catégorisation des entités nommées dans les textes français*. PhD thesis, Université de Nantes.

GIGUET E. & LUCAS N. (2004). La détection automatique des citations et des locuteurs dans les textes informatifs. In (López-Muñoz *et al.*, 2004), p. 410–418.

HIRSCH J. E. (2005). An index to quantify an individual's scientific research output. In *Proceedings of the National Academy of Sciences*, volume 102, p. 16569–16572.

JACKIEWICZ A. (2006). Relations intersubjectives dans les discours rapportés. *Journal TAL*, **47**(2), 65–87.

J.-M. LÓPEZ-MUÑOZ, S. MARNETTE & L. ROSIER, Eds. (2004). *Le discours rapporté dans tous ses états : question de frontiéres*, Paris. L'Harmattan.

MOURAD G. & DESCLÉS J.-P. (2004). Identification et extraction automatique des informations citationnelles dans un texte. In (López-Muñoz *et al.*, 2004).

PEVZNER L. & HEARST M. (2002). A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*.

POULARD F. (2008). Analyse quantitative et qualitative de citations extraites d'un corpus journalistique. In *RECITAL'08 (À paraître)*, Avignon.

PRASAD R., DINESH N., LEE A., JOSHI A. & WEBBER B. (2006). Attribution and its annotation in the penn discourse treebank. *Journal TAL*, **47**(2), 43–64.

RITCHIE A., TEUFEL S. & ROBERTSON S. (2006). How to find better index terms through citations. In *"Can Computational Linguistics Improve Information Retrieval ?" Workshop at ACL/COLING*, Sydney, Australia.

SOMASUNDARAN S., WILSON T., WIEBE J. & STOYANOV V. (2007). Qa with attitude : Exploiting opinion type analysis for improving question answering in on-line discussions and the news. In *International Conference on Weblogs and Social Media (ICWSM'07)*.

STOYANOV V. & CARDIE C. (2006). Toward opinion summarization : Linking the sources. In *COLING-ACL 2006 Workshop on Sentiment and Subjectivity in Text*.

TEUFEL S., SIDDHARTHAN A. & TIDHAR D. (2006). Automatic classification of citation function. In *EMNLP*, Sydney, Australia.

WILSON T., HOFFMANN P., SOMASUNDARAN S., KESSLER J., JANYCEWIEBE, CHOI Y., CARDIE C., RILOFF E. & PATWARDHAN S. (2005). Opinionfinder : A system for subjectivity analysis. In *HLT - EMNLP*, Vancouver, Canada.

WITTEN I. H. & FRANK E. (2005). In *Data Mining : Practical machine learning tools and techniques*. Morgan Kaufmann, 2nd edition edition.