TALN 2011, Montpellier, 27 *juin* –1er *juillet 2011*

Identifier la cible d'un passage d'opinion dans un corpus multithématique

Matthieu Vernier, Laura Monceaux, Béatrice Daille
Université de Nantes, LINA, 2, rue de la Houssinière 44322 Nantes
{Matthieu.Vernier, Laura.Monceaux, Beatrice.Daille}@univ-nantes.fr

Résumé. L'identification de la cible d'une d'opinion fait l'objet d'une attention récente en fouille d'opinion. Les méthodes existantes ont été testées sur des corpus monothématiques en anglais. Elles permettent principalement de traiter les cas où la cible se situe dans la même phrase que l'opinion. Dans cet article, nous abordons cette problématique pour le français dans un corpus multithématique et nous présentons une nouvelle méthode pour identifier la cible d'une opinion apparaissant hors du contexte phrastique. L'évaluation de la méthode montre une amélioration des résultats par rapport à l'existant.

Abstract. Recent works on opinion mining deal with the problem of finding the semantic relation between sentiment expressions and their target. Existing methods have been evaluated on monothematic english corpora. These methods are only able to solve intrasentential relationships. In this article, we focus on this task apply to french and we present a new method for solving intrasentential and intersentential relationships in a multithematic corpus. We show that our method is able to improve results on the intra- and intersential relationships.

Mots-clés : Fouille d'opinions, Identification des cibles, Méthode RankSVM.

Keywords: Opinion mining, Targeting sentiment expressions, RankSVM.

1 Introduction

Le début des années 2000 marque l'éclosion de la fouille d'opinions. Les travaux pionniers se sont principalement intéressés à la catégorisation globale de documents d'opinion, soit selon leur polarité (Turney, 2002; Torres-Moreno *et al.*, 2007), soit selon leur subjectivité (Wiebe & Riloff, 2005). Depuis, un très grand nombre de travaux traitent de données textuelles d'opinion dans des axes scientifiques et des domaines applicatifs très différents. Plus récemment, le recul sur dix ans de travaux permet selon nous de segmenter le domaine en cinq problématiques :

– **extraire les mots d'opinions** d'une langue pour construire des ressources et améliorer leur qualité (Baccianella *et al.*, 2010; Mathieu, 2006) ;
– **catégoriser globalement un document** selon l'opinion (Torres-Moreno *et al.*, 2007; Pang & Lee, 2008) ;
– **catégoriser des passages d'opinions** dans un document qui exprime des opinions hétérogènes (Wilson, 2008) ;
– **identifier la source** [1] d'une opinion (Choi *et al.*, 2005; Ruppenhofer *et al.*, 2008) ;
– **identifier la cible** [2] d'une opinion (Kessler & Nicolov, 2009; Jakob & Gurevych, 2010).

1. l'énonciateur d'une opinion.
2. le sujet sur lequel porte l'opinion.

L'identification de la cible d'un passage d'opinion fait partie des axes les plus récemment abordés dans la littérature du domaine. Aucun travail ne s'y est intéressé pour le traitement du français et, à notre connaissance, seuls deux travaux anglophones majeurs y consacrent une étude spécifique (Kessler & Nicolov, 2009; Jakob & Gurevych, 2010). Pourtant, afin d'analyser des textes dont le contenu est de plus en plus hétérogène (blogs, forums, etc.), cette problématique correspond à un besoin particulièrement criant. Elle nécessite de combiner des techniques connues en traitement automatique des langues (analyse syntaxique, résolution d'anaphore pronominale et nominale, segmentation thématique, etc.) pour proposer une solution souple et efficace (voir §. 2).

Dans cet article, nous nous intéressons spécifiquement à l'identification de la cible d'une opinion (voir §. 2). Nous renvoyons à (Vernier & Monceaux, 2010) pour la présentation d'une méthode et d'une ressource pour la détection automatique des passages d'opinions pour le français. Dans nos travaux précédents, nous avons développé le premier corpus francophone (*corpus Blogoscopie*) dans lequel les passages d'opinions et leur cible sont annotés (voir §. 3.1) (Dubreil *et al.*, 2008). Une étude statistique sur ce corpus permet de structurer le problème et de proposer différentes pistes de solutions (voir §. 3.2). Nous présentons quatre méthodes (dont deux méthodes *baseline*) pour identifier la cible d'un passage d'opinion (voir §. 4). Nous expérimentons ces quatre méthodes sur une sous-partie du corpus Blogoscopie et discutons des résultats obtenus (voir §. 5).

2 Une problématique récente peu explorée

L'identification de la cible d'une opinion a souvent été ignorée ou considérée comme un aspect de second plan. Par exemple, les travaux de catégorisation de l'opinion au niveau du document considèrent qu'un texte est monothématique et n'évalue qu'un seul objet donné (*un film, un livre, un appareil photo,* etc.). Ce n'est que récemment que quelques travaux anglophones ont directement axés leur étude sur cette problématique (Kessler & Nicolov, 2009; Jakob & Gurevych, 2010). Ils ont introduit le terme *target*, traduit ici par *cible*, pour désigner l'objet concerné par un passage d'opinion. Dans la pratique, ils ont limité cette cible à quelques catégories d'objets :
– un produit particulier ou une marque (*iPhone, EOS 5D de Canon, Apple, Matrix*, etc.) ;
– une caractéristique d'un produit (*la durée de vue de la batterie*, *le scénario d'un film*, etc.).
En fait, de part le large éventail d'objets qu'elle peut couvrir, la notion de cible est complexe à circonscrire. Elle est d'ailleurs souvent laissée floue dans les travaux du domaine. Dans notre cadre, nous considérons que tout objet peut être la cible d'une opinion. Nous tâchons de mieux appréhender cette notion en présentant quatre facteurs qui rendent complexe la problématique d'identification de la cible d'une évaluation (§. 2.1). Plusieurs approches ont été proposées mais celles-ci ne résolvent que partiellement la tâche et n'ont été appliquées qu'à l'anglais. Nous précisons leurs limites et motivons le travail présenté dans cet article (§. 2.2).

2.1 Une tâche complexe

L'identification de la cible d'une opinion consiste à relier un passage d'opinion avec l'objet du monde qu'il évalue. Dans (1), il y a ainsi trois passages d'opinions qui évaluent un objet cible unique. Néanmoins, cette tâche s'avère particulièrement complexe de part quatre facteurs que nous précisons ci-dessous.

(1) L'équipe de France va mal . Entre honte et déception .

Différentes formes textuelles pour un unique objet du monde Premier facteur, comme l'a introduit Benveniste (Benveniste, 1966), il importe de distinguer les signes linguistiques et les objets du monde auxquels ils font références. Un même objet peut être représenté dans le texte par différentes expressions nominales ou pronominales.

Dans (2) et (3), on parle de l'objet du monde *équipe de France de football* par une variante nominale métonymique (*les Bleus*) et par une anaphore pronominale (*elle*). Dans un but applicatif, il est nécessaire de regrouper les opinions qui portent sur le même objet et de nommer l'objet évalué le plus précisément possible. On ne pourra ainsi pas considérer le pronom *elle* comme la cible de l'évaluation *séduisante*. *Les Bleus* est une réponse intermédiaire plus acceptable pour identifier l'objet cible réellement évalué (*l'équipe de France de football*).

(2) [...] lors du fiasco des Bleus en juin.

(3) Elle a enfin été séduisante [...].

Les relations méronymiques entre objets Les objets sont potentiellement liés à d'autres objets par des relations méronymiques [3]. Dès lors, même si évaluer un méronyme A d'un mot B peut être une façon d'évaluer indirectement B, il importe de considérer le méronyme A comme la cible exacte de l'évaluation. Ainsi, dans (4) et (5), on évalue tout d'abord *la défense* et *le marquage sur les corners* de *l'équipe de France* et non *l'équipe de France* dans sa globalité.

(4) Disons-le clairement : la défense française n'a pas été très bonne , ce soir.

(5) Le marquage sur les corners reste encore approximatif .

Présence de plusieurs objets candidats autour de l'opinion Le troisième facteur est la présence de plusieurs objets distincts dans le contexte d'un même passage d'opinion. Dans (6), les deux passages d'opinions portent sur des cibles distinctes qu'il faut pouvoir déterminer parmi les quatres objets présents dans la phrase : *la dernière finale de la coupe du monde, l'équipe de France, le coursier, les pizzas*. L'objet le plus proche de l'opinion n'est pas nécessairement sa cible.

(6) [...] pizzas commandées lors de la dernière finale de coupe du monde avec l'équipe de France , coursier courageux , pizzas aussi banales que le match .

Proximité aléatoire entre l'opinion et sa cible Dernier facteur de complexité, la cible évaluée ne se situe pas toujours à proximité du passage d'opinion. Dans (7), le passage d'opinion *de la provocation* [4] porte sur *la ligne du sélectionneur* qui se trouve dans la phrase précédente. Le pronom *ce* en est une anaphore. Néanmoins, la présence de nombreux autres objets dans le contexte (*un iota, le jeu, les joueurs, la concrétisation de cette vue*) rend complexe l'identification de la cible réelle de l'opinion pour une approche automatique.

(7) force est de constater que la ligne du sélectionneur n'a jamais bougé d'un iota : le jeu appartient aux joueurs. Ce qui a souvent pu passer pour de la provocation n'est en fait [...]

Pour résoudre la problématique d'identification de la cible d'une opinion, il convient donc de considérer ces quatre facteurs. Le processus d'identification consiste alors à :

– repérer les différents objets impliqués dans un énoncé. Ceux-ci sont généralement représentés textuellement par des groupes nominaux ou pronominaux (tâche T_1) ;
– identifier l'objet cible qui est directement concerné par une opinion (tâche T_2) ;
– résoudre les relations d'anaphores nominales ou pronominales entre les objets (tâche T_3) ;
– résoudre les relations méronymiques entre les objets (tâche T_4).

3. Un méronyme A d'un mot B est un mot dont le signifié désigne une sous-partie du signifié de B.
4. Ce passage est considéré comme une opinion même s'il s'agit d'un discours rapporté.

2.2 Travaux liés et motivations

Pour résoudre l'identification de la cible d'une opinion, quatre types d'approches ont été proposées dans les travaux du domaine. Les solutions qu'elles proposent ne sont que partielles et nécessitent d'être améliorées. De plus, elles n'ont été évaluées que sur l'anglais et sur des corpus monothématiques.

Quatre approches proposées La **proximité** est la première hypothèse considérée pour relier une cible et une opinion. Ainsi, dans les travaux de Grefenstette *et al.* (2004) (textes journalistiques sur la politique) et Mishne & Glance (2006) (blogs sur le cinéma), les auteurs considèrent un mot ou une expression w comme objet cible. Toutes les opinions d'un document qui co-occurent avec w dans une certaine fenêtre de mots, sont considérées comme des opinions qui portent sur w. Cette tâche n'étant pas le coeur de leur travail, l'efficacité de cette méthode n'a pas été évaluée. Les **dépendances syntaxiques** entre une opinion et une cible constitue la deuxième méthode pour résoudre cette problématique. Hu & Liu (2004) se sont restreints aux chemins syntaxiques proches entre un adjectif et une cible dans un corpus de film. De la même façon, Bethard *et al.* (2004) se sont intéressés uniquement aux verbes d'opinion. Bloom *et al.* (2007) ont développé manuellement des listes de chemins de dépendances syntaxiques entre une opinion et sa cible dans des corpus de produits commerciaux. La précision obtenue varie autour de 0,70. Toutes ces approches ont été entraînées sur des corpus monothématiques ce qui tend à restreindre la richesse des cas observés. De plus, ces méthodes ne sont pas applicables dans les nombreux cas où la cible se situe dans une phrase différente de l'opinion. Kim & Hovy (2006) ont ainsi montré que dans une majorité des cas l'étude des dépendances syntaxiques ne suffit pas à déterminer la bonne cible et qu'elle est souvent confondue avec la source de l'opinion. En **combinant des informations** lexicales, grammaticales et syntaxiques par apprentissage, Kessler & Nicolov (2009) ont amélioré les résultats des approches syntaxiques sur un corpus monothématique de critiques de voiture. Leur méthode est la méthode état de l'art pour l'anglais, néanmoins elle se restreint toujours au grain phrase et ne traite pas les nombreux cas où la cible n'est pas présente dans la même phrase que l'opinion. Ils se sont également limités à l'étude des cas où l'opinion est un adjectif ou un verbe. Jakob & Gurevych (2010) sont à notre connaissance les seuls à sortir du grain phrase. Ils se sont intéressés spécifiquement à la **résolution d'anaphore pronominale** dans un corpus de critiques de film pour mieux identifier la cible exacte.

Motivations Dans ce travail, nous souhaitons d'une part adapter les travaux existants au français et vérifier ainsi les conclusions obtenues. D'autre part, nous souhaitons améliorer l'existant sur les trois aspects suivants :

– notre approche est **multithématique**. En effet, l'aspect monothématique des études existantes (souvent ciblées sur un genre particulier : les critiques d'objets commerciaux) tend à simplifier les cas linguistiques rencontrés ;
– notre approche **ne se limite pas à quelques catégories grammaticales** d'opinion. Les opinions peuvent ici être des groupes nominaux, adjectivaux, verbaux ou adverbiaux ;
– nous utilisons **le texte comme grain d'étude**. Les cas où la cible n'est pas présente dans la même phrase que l'opinion sont pris en compte.

Pour les besoins de notre travail, un corpus de référence où les passages d'opinions et les cibles sont annotées est nécessaire. Nous présentons dans la section suivante le corpus utilisé.

3 Corpus multithématique de référence

Actuellement, le seul corpus francophone disponible où les passages d'opinions et les cibles d'opinions sont annotés est le corpus de blogs *Blogoscopie* [5] (Dubreil *et al.*, 2008). Pour le problème présenté, les blogs sont intéressants car ils présentent une richesse linguistique plus grande que les textes de critiques d'objets commerciaux

5. Le corpus est disponible à l'adresse : http ://www.lina.univ-nantes.fr/Ressources.html

(beaucoup d'objets différents dans un même texte, syntaxes des phrases plus variées, etc.). Nous présentons brièvement le corpus Blogoscopie (§. 3.1) et analysons dans une deuxième partie des données statistiques sur le problème d'identification d'une cible d'opinion (§. 3.2).

Nature du corpus Le corpus Blogoscopie est constitué de 200 billets de blogs, et de 614 commentaires associés à ces billets, tous issus de la plateforme de blogs OverBlog [6]. Il est constitué d'environ 100 000 mots. La création du corpus Blogoscopie est basée sur un critère de représentativité thématique. Ainsi :
– 110 billets et 458 commentaires ont été sélectionnés à partir des 33 catégories de la plateforme OverBlog : *actualité, artiste, cinéma, consommation, console, croyance, détente, économie, gastronomie,* etc.
– 90 billets et 156 commentaires ont été sélectionnés à partir de dix mots-clés (9 billets par mot-clé) : *Beaujolais, développement durable, grève SNCF, Harry Potter, loi LRU, énergie nucléaire, Raymond Domenech,* etc.

L'exemple (8) présente une version allégée de l'annotation d'un passage du corpus Blogoscopie. Dans ce corpus, trois types d'objets principaux ont été annotés :
– 6 876 **objets**. Ces objets peuvent être dits :
  – « principaux » (OP) : si le sujet général d'un document concerne principalement cet objet. Un même document peut avoir plusieurs objets principaux ;
  – « associés » (OA) : si cet objet est une sous-partie d'un objet concerné.
– 4 909 **opinions** ;
– 4 129 **couples cible-opinion**.

(8) Le sommet a été atteint avec **<objet>** le nucléaire **</objet>** : **<objet>** M. Sarkozy **</objet>** **<opinion cible="M. Sarkozy">** a abusé **</opinion>** **<objet>** l'opinion publique **</objet>** en annonçant qu'il n'y aurait pas de **<objet>** "nouveaux sites" **</objet>**.

Précisons que dans ce corpus :
– les annotateurs n'ont pas pris en compte les anaphores pronominales ;
– la forme textuelle de l'objet cible la plus proche de l'opinion est considérée comme la cible. Lorsqu'une anaphore pronominale est présente, l'antécédent nominal le plus proche est considéré comme la cible.

Segmentation du corpus et statistiques Pour nos expérimentations, nous segmentons le corpus en deux :
– une partie « entraînement » constituée de 3 909 couples opinion-cible et de 5 584 objets (160 documents) ;
– une partie « test » constituée de 1 000 couples opinion-cible et de 1 292 objets (40 documents).

Nb OP / billet	1	2	3	4	5+
billets concernés	17 %	**35 %**	**19 %**	14 %	15 %

TABLE 1 – Nombre d'objets principaux par billet

Nb OA / billet	1-5	5-10	10-15	15+
billets concernés	**30 %**	23 %	15 %	**32 %**

TABLE 2 – Nombre d'objets associés par billet

Des statistiques sur la partie « entraînement » du corpus donnent un aperçu de la tâche. Le tableau 1 montre qu'une majorité de billets (et leurs commentaires) parle en général d'au moins deux ou trois objets principaux différents. Le tableau 2 montre que le nombre d'objets associés est très variable. Dit autrement, si dans une majorité de cas un blogueur articule le sujet de son billet autour de deux ou trois objets principaux, en revanche, on peut difficilement prédire le nombre d'objets associés qu'il va évoquer. Pour identifier la cible d'une opinion, il s'agit donc de la trouver parmi une liste d'objets pouvant aller de 2 à plus de 20. Le tableau 3 montre que la position de la cible est variable : soit très proche, soit très éloignée (voire dans une phrase différente dans 44 % des cas). Dans 47 % des cas, il y a au moins un objet intercallé entre l'opinion et sa cible.

6. http ://www.over-blog.com

Position relative de la cible	avant l'opinion		après l'opinion	
opinions concernées	**61 %**		39 %	
Distance - en nombre de mots -	0	1-4	5-10	11+
opinions concernées	**21 %**	**20 %**	11 %	**48 %**
Distance - en nombre de phrases -	0	1	2	3+
opinions concernées	**56 %**	9 %	6 %	19 %
Distance- en nombre de objets intercallés -	0	1	2	3+
opinions concernées	**53 %**	7 %	12 %	**28 %**

TABLE 3 – Étude de la position de la cible par rapport à son opinion

4 Méthodes utilisées

D'après les statistiques du corpus « entraînement » Blogoscopie, nous distinguons deux niveaux de complexité :
– la cible et l'opinion sont dans la même phrase (intraphrastique) : 56 % des cas du corpus d'entraînement ;
– la cible et l'opinion sont dans une phrase différente (interphrastique) : 44 % des cas du corpus d'entraînement.
Dès lors, nos objectifs sont les suivants :
– **évaluer les méthodes existantes** qui traitent uniquement les cas **intraphrastiques** afin de les valider sur le français, sur un corpus multithématique et en prenant davantage en compte les différentes catégories grammaticales des opinions ;
– **développer une méthode** qui traite les cas **intra- et interphrastiques** simultanément.
Pour le premier objectif, nous ré-implémentons la méthode état-de-l'art (**M-RankSVM**, voir §. 4.1.1) et nous la comparons à une méthode *baseline* (**M-Syntaxe**, voir §. 4.1.2). Pour le second objectif, nous proposons une méthode basée sur la saillance d'une cible (**M-Saillance**, voir §. 4.2.1). Celle-ci ré-utilise les résultats de **M-RankSVM** comme une amorce. Nous la comparons avec une méthode *baseline* (**M-Proximite**, voir §. 4.2.2). Pour ces quatre méthodes :
– les passages d'opinions sont annotés manuellement (annotations présentes dans le corpus Blogoscopie) ;
– le TreeTagger (Schmid, 1994) donne la catégorie grammaticale ;
– un algorithme de découpage en syntagmes non-récursifs, qui réimplémente la méthode de (Vergne & Giguet, 1998), permet d'identifier les groupes nominaux et pronominaux d'un texte. Ces expressions nominales et pronominales sont considérées comme l'ensemble des objets principaux et associés du texte.

4.1 Méthodes applicables à un niveau intraphastique

4.1.1 Approche par combinaison d'indices morpho-syntaxiques

La méthode état de l'art de Kessler & Nicolov (2009) a été évaluée sur un corpus monothématique anglais. Elle ne s'intéresse qu'aux opinions de type verbe et adjectif uniquement. Elle consiste à combiner des informations lexicales et sémantiques pour pallier les limites des approches purement syntaxiques. Nous adaptons leur approche pour traiter un corpus multithématique français sans restriction grammaticale (par la suite, **M-RankSVM**). La méthode prend en entrée : une phrase s à analyser, un passage d'opinion $opinion$ présent dans s, un ensemble d'objets o présents dans s et une fonction d'ordonnancement permettant de classer les objets de s. Cette fonction

d'ordonnancement est extraite par apprentissage sur le corpus « entraînement ».

Pour obtenir la fonction d'ordonnancement, nous utilisons l'algorithme RankSVM issue de SVMLight (Joachims, 2002). Le corpus « entraînement » Blogoscopie permet de fournir à la méthode :

– 3 909 exemples « positifs » correspondant aux couples « opinion-cible » présents dans le corpus. Ces exemples ont un rang 1 ;
– 4 000 exemples « négatifs » correspondant à des couples « objet non cible-opinion » présents dans le corpus. Ces exemples ont un rang 0 ;

Chacun de ces exemples est décrit par les caractéristiques présentées dans le tableau 4. À partir de ces exemples, RankSVM entraîne un modèle ayant pour objectif de déterminer une fonction de préférence qui maximise le *rang*. À l'image de Kessler & Nicolov (2009), nous utilisons les paramètres par défaut de RankSVM. Lors de la phase de test, cette méthode effectue un ordonnancement de tous les objets qui apparaissent dans la même phrase que le passage d'opinion en s'appuyant sur le modèle appris lors de la phase d'entraînement. L'objet classé premier est considéré comme la cible de l'opinion.

Caractéristique	**Exemple**
Chemin lexical lemmatisé entre l'opinion et sa cible	*pour*
Chemin grammatical entre l'opinion et sa cible	Prep.
Catégorie(s) grammatical(es) de l'opinion	DET NOM ADJ
Catégorie(s) grammatical(es) de la cible	DET NOM
Chemin de dépendance	↓ PP, ↓ DP
Nombre d'objets entre l'opinion et la cible	0
Nombre de mots entre l'opinion et la cible	1
Rang	**1**

TABLE 4 – Caractéristiques utilisées pour décrire chaque instance d'apprentissage. Exemple : « [...] est une défaite majeure pour l'écologie »

Chemin de dépendance(s)	**Fréquence**	**Exemple**
↑ AP	121	C'est une belle image (ex : [$_{TP}$[$_{DP}$ C'] est [$_{VP}$ [$_{DP}$ une [$_{NP}$[$_{AP}$ belle] image]]]])
↓ DP	91	Cette question me taraude
↓ PP, ↓ DP	65	J'ai une véritable fascination pour J.K Rowling
↑ TP, ↑ CP	59	Un événement qui contribue à alourdir le débât
↑ VP, ↓ DP	45	Ce jeu est accessible pour les enfants de 12 ans
↑ AdvP, ↓ DP	29	Cette histoire de travailler plus [...] ne tient [...] pas la route

TABLE 5 – Chemins de dépendances syntaxiques les plus fréquents entre une opinion et une cible dans le corpus d'entraînement Blogoscopie. AP = syntagme adjectival ; DP = syntagme déterminant ; PP = syntagme prépositionnel ; AdvP = syntagme adverbial ; CP = syntagme conjonctionnel ; NP = syntagme nominal ; TP = syntagme temporel ; VP = syntagme verbal

4.1.2 Approche syntaxique (*baseline*)

Nous comparons la méthode **M-RankSVM** avec une méthode qui repose uniquement sur une analyse des dépendances syntaxiques (par la suite, **M-Syntaxe**). Elle ne fonctionne également que dans un cadre intraphrastique. Elle prend en entrée : une phrase s à analyser, un passage d'opinion $opinion$ présent dans s, un ensemble d'objets O présents dans s, un ensemble de chemins de dépendances syntaxiques entre une opinion et une cible avec leur probabilité correspondante. Cet ensemble est déterminé à partir du corpus « entraînement » Blogoscopie.

Nous avons utilisé l'analyseur de dépendance syntaxique FIPS (Wehrli, 2007) sur 500 exemples [7] de couples opinion-cible qui apparaissent dans la même phrase dans le corpus d'entraînement Blogoscopie. Pour ces exemples, FIPS fournit un chemin de dépendance (visualisable sous forme d'arbre syntaxique (Fig. 1)). Nous classons par ordre de fréquence décroissante (voir tableau 5), les chemins de dépendances qui vont :
– de la fin d'une opinion vers le début d'une cible, si l'opinion est située avant la cible dans le texte ;
– du début de l'opinion vers la fin d'une cible, si l'opinion est située après la cible dans le texte.
Lors de la phase de test, cette méthode effectue une analyse syntaxique de la phrase contenant l'opinion. Le chemin de dépendance syntaxique le plus fréquent qui permet de relier $opinion$ avec un objet o ($\in O$) désigne l'objet o comme la meilleure cible potentielle. En utilisant un des chemins syntaxiques, il est possible que o soit un groupe pronominal (*il*, *aucun d'entre eux*, etc.).

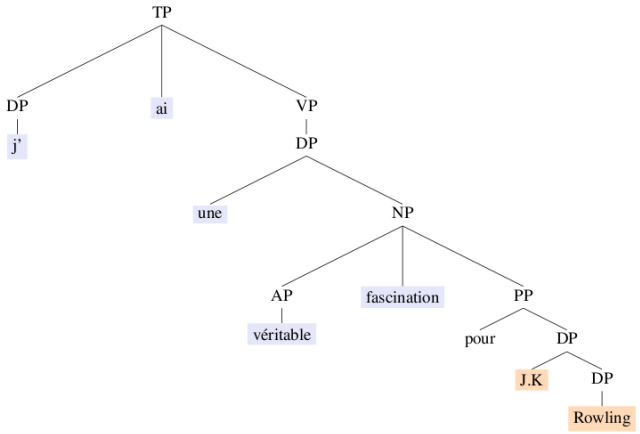

FIGURE 1 – Analyse syntaxique de « *J'ai une véritable fascination pour J.K Rowling* ». L'opinion est séparée de sa cible par la dépendance ↓ **PP,** ↓ **DP**.

4.2 Méthodes applicables à un niveau intra- et interphrastique

Pour le second objectif, nous proposons une méthode (**M-Saillance**) pour traiter les cas intraphrastiques et interphrastiques. Nous comparons cette méthode à une méthode *baseline* (**M-Proximite**).

7. Le nombre de couples est réduit à 500 pour éviter les cas des phrases trop longues qui induisent en erreur l'analyseur FIPS.

4.2.1 Approche par mesure de saillance

Lors de la phase de test, en sortie de la méthode **M-RankSVM**, pour chaque document d, nous disposons d'un ensemble de couples d'opinion-cible déterminés automatiquement. Nous observons les quatre faits suivants :
– O_1 : certaines opinions n'ont **pas de cible affectée** (si la phrase contenant l'opinion ne possède aucun objet) ;
– O_2 : certaines opinions ont pour cible **un groupe pronominal** ;
– O_3 : certains couples cible-opinion ne sont **pas reliés par une dépendance syntaxique couramment observée** ;
– O_4 : certaines des cibles ne sont **pas évaluées ailleurs dans le document**.

Les couples qui vérifient l'une des observations O_1 ou O_2, ou qui vérifient simultanément O_3 et O_4 sont considérées comment des couples **faibles** (cas douteux). Les autres couples sont considérés comme des identifications **fortes** sur lesquelles nous pouvons nous appuyer. Notre méthode (**M-Saillance**) consiste à utiliser les résultats « forts » de **M-RankSVM** comme une amorce pour corriger les couples faibles. Cette méthode a deux objectifs :
– identifier une cible nominale pour les opinions qui n'en ont pas ou qui sont reliées à un groupe pronominal ;
– corriger l'affectaction des autres couples faibles.
Les hypothèses linguistiques de notre méthode sont les suivantes :
– **Hypothèse 1** : s'il n'existe pas de objet nominal pertinent dans la même phrase qu'une opinion, celui-ci existe probablement dans les phrases voisines ;
– **Hypothèse 2** : l'énonciateur d'un discours subjectif articule ses opinions au fil d'un document en évaluant les mêmes objets plusieurs fois. Plus un objet est la cible d'une opinion, plus il est saillant et plus il est probable qu'il soit de nouveau la cible d'une opinion dans le contexte.

Description Soit $opinion$, un passage d'opinion du document d, dont la relation avec sa cible actuelle est **faible**. Soit $segment$, l'ensemble des phrases voisines de l'opinion. $segment$ est limité à 4 phrases avant l'opinion et 4 phrases après. À partir des couples forts opinion-cible issus de **M-RankSVM**, nous classons par ordre décroissant les cibles les plus fréquemment liées à une opinion dans $segment$. Nous classons également les cibles les plus fréquentes dans d. Les cibles sont préalablement lemmatisées et les mots fonctionnels sont retirés (par exemple, *l'entraîneur de l'équipe de France* devient *entraîneur équipe France*). Pour tout objet o présent dans $segment$, nous mesurons une probabilité que o soit la cible de $opinion$ avec la formule suivante :

$$\text{P}(o, opinion) = \frac{\text{NB}(o, cible, segment)}{\text{NB}(o, segment).\text{NB}(cible, segment)} \cdot \frac{\text{NB}(o, cible, document)}{\text{NB}(o, document).\text{NB}(cible, document)} \quad (1)$$

$$\text{Cible}(opinion) = \text{MAX}(\text{P}(o, opinion)) \quad (2)$$

où :
– NB(o, $cible$, $segment$) est le nombre de fois que le objet o est une cible forte d'une opinion quelconque dans l'ensemble des phrases voisines d'$opinion$. De la même façon, NB(o, $cible$, $document$) est le nombre de fois que o est une cible forte d'opinion dans le document ;
– NB(o, $segment$) et NB(o, $document$) sont les nombres de fois où le objet o est présent dans le segment de phrases voisines et dans le document ;
– NB($cible$, $segment$) et NB($cible$, $document$) sont les nombres de cibles d'opinions dans le segment et dans le document.
P(o, $opinion$) représente la saillance de l'objet o dans le segment et le document. Cette mesure donne un score d'association entre le l'objet o et l'opinion émise dans le document et le segment en calculant la dépendance de ces trois variables. La cible de l'opinion choisie par **M-Saillance** est l'objet qui maximise cette probabilité.

4.2.2 Approche par proximité (*baseline*)

La méthode par proximité (par la suite, **M-Proximité**) est applicable au niveau intraphrastique et interphrastique. Elle prend en entrée : le document d à analyser, un passage d'opinion $opinion$ et sa position dans le document, un ensemble d'objets O présents dans d et leur position dans le document. La position dans le document est déterminée par une indexation de l'ensemble des mots du texte : $W = \{w_1, w_2, \ldots, w_n\}$. Le passage $opinion$ et l'ensemble des éléments de O sont délimités par un mot de début ($w_{\text{début}} \in W$) et un mot de fin ($w_{\text{fin}} \in W$).

Pour tout $o \in O$, cette méthode compte le nombre de mots qui séparent l'objet o de $opinion$ dans le document d. La ponctuation n'est pas prise en compte. Si $opinion$ est inclu dans o, la distance en mots est égale à 0. Le choix du meilleur candidat s'effectue sur la base de la plus petite distance en mots. En cas d'égalité, la cible qui précède l'opinion est préférée (nous justifions ce choix empirique par l'étude statistique réalisée précédemment).

$$\forall\, o \in O \text{ se situant avant } opinion, \text{Proximite}(t) = w_{\text{fin}}(o) - w_{\text{début}}(opinion)$$

$$\forall\, o \in O \text{ se situant après } opinion, \text{Proximite}(t) = w_{\text{fin}}(opinion) - w_{\text{début}}(o)$$

$$\text{Cible}(opinion) = \text{MIN}(\text{Proximite}(o))$$

5 Résultats

Type / **Score**	**Intraphrastique** Précision	**Interphrastique** Précision	**Total** Précision
M-Syntaxe	68.8 % (407/592)	non applicable	40.7 % (407/1 000)
M-RankSVM	71.5 % (423/592)	non applicable	42.3 % (423/1 000)
M-Proximite	53.5 % (317/592)	27.2 % (111/408)	42.8 % (428/1 000)
M-Saillance	**72.8 %** (431/592)	**60.0 %** (245/408)	**67.6 %** (676/1 000)

TABLE 6 – Résultats obtenus par les quatre méthodes sur le corpus de test Blogoscopie

Nous évaluons les 4 méthodes présentées précédemment sur les 40 documents du corpus test Blogoscopie. Ce corpus contient 1 000 couples cible-opinion : 592 de ces couples sont intraphrastiques, 408 sont interphrastiques.

Niveau intraphrastique Par rapport à une approche syntaxique, la combinaison d'indices morpho-syntaxiques améliore légèrement l'identification de la cible (71,5 % contre 68,8 %). Cette différence est néamoins beaucoup plus marquée dans les travaux de Kessler & Nicolov (2009). Par rapport à Kessler & Nicolov (2009) qui ont obtenu une précision de 74,8 % sur l'anglais, notre score de précision baisse d'environ 3 points sur le français (71,5 %). S'il est difficile de dire si cette baisse est significative, elle tend néanmoins à montrer que la méthode **M-RankSVM** est applicable pour traiter les cas intraphrastiques sur des corpus multithématiques en français. Elle est d'autant plus intéressante qu'elle fournit de bons indices à la méthode **M-Saillance** pour corriger certains couples intraphrastiques (72,8 % contre 71,5 %). L'hypothèse discursive consistant à reclasser les cibles d'opinions saillantes (celles qui sont fréquement liées à une opinion) semble être une hypothèse intéressante à développer.

Niveau interphrastique La méthode **M-Saillance** permet d'identifier correctement 60,0 % des cibles d'opinions en s'appuyant uniquement sur l'objet d'opinion le plus saillant dans une fenêtre de 4 phrases autour de l'opinion. Ce résultat est nettement supérieur à l'approche naïve consistant à prendre l'objet le plus proche (27,2 %). Ce score permet à la méthode **M-Saillance** d'être beaucoup plus performante que la méthode état de l'art **RankSVM** sur l'ensemble de la problématique d'identification de la cible d'une opinion (67.6 % contre 42,3 %). Ce score nous semble encore améliorable en introduisant d'autres types d'hypothèses sémantiques et discursives. Par exemple :

– si une cible A est évaluée fréquement positivement dans un document, si une cible B est évaluée fréquement négativement dans le même document, il est plus probable que A (respectivement B) soit reliée à une opinion positive (respectivement négative) dont la cible est inconnue ;
– si une ou plusieurs cibles appartenant au même graphe d'objets (modélisant les relations méronymiques entre objets) sont évaluées dans le même passage d'un document, il est probable qu'une opinion dont on ignore la cible porte sur un objet de ce graphe.

Ces hypothèses seraient intéressantes à comparer à des procédures de résolution d'anaphores nominales plus classiques pour le français. Néanmoins, ces procédures nous paraissent à ce stade encore complexes à mettre en place sur des corpus multithématiques et en particulier sur les blogs (dont le respect grammatical et orthographique est très variable). De ce point de vue, l'hypothèse de saillance d'un objet semble offrir une plus grande souplesse.

6 Conclusion

Nous avons abordé la problématique d'identification de la cible d'un passage d'opinion. Nos résultats montrent que la méthode état de l'art évaluée récemment sur des corpus anglophones monothématiques est adaptée pour un corpus français multithématique. Cette méthode, applicable au niveau intraphastrique uniquement, est utilisée comme amorce pour le développement de la nouvelle approche que nous proposons. Celle-ci est basée sur la saillance d'une cible d'opinion dans un segment textuel. Elle améliore l'état de l'art sur l'aspect intraphastique de la tâche. Elle permet également de couvrir davantage de problèmes en s'intéressant aux cas interphrastiques non traités dans le domaine jusqu'à présent. De ce point de vue nous améliorons significativement les résultats pour identifier la cible d'un passage d'opinion. Ces résultats nous semblent encore améliorables en prenant mieux en compte les relations sémantiques entre les objets d'un document.

Références

BACCIANELLA S., ESULI A. & SEBASTIANI F. (2010). Sentiwordnet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta : European Language Resources Association (ELRA).

BENVENISTE E. (1966). *Problèmes de linguistique générale II*. Gallimard edition.

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

BLOOM K., GARG N. & ARGAMON S. (2007). Extracting appraisal expressions. In *Human Language Technologies 2007 : The Conference of the North American Chapter of the Association for Computational Linguistics*, p. 308–315, Rochester, New York : Association for Computational Linguistics.

CHOI Y., CARDIE C., RILOFF E. & PATWARDHAN S. (2005). Identifying sources of opinions with conditional random fields and extraction patterns. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*.

DUBREIL E., VERNIER M., MONCEAUX L. & DAILLE B. (2008). Annotating opinion - evaluation of blogs. In *Workshop on LREC 2008 Conference, Sentiment Analysis Metaphor, Ontology and Terminology (EMOT-08)*.

GREFENSTETTE G., QU Y., SHANAHAN J. G. & EVANS D. A. (2004). Coupling niche browsers and affect analysis for an opinion mining application. In *Proceedings of Recherche d'Information Assistée par Ordinateur*.

HU M. & LIU B. (2004). Mining opinion features in customer reviews. In *Proceedings of AAAI*, p. 755–760.

JAKOB N. & GUREVYCH I. (2010). Using anaphora resolution to improve opinion target identification in movie reviews. In *Proceedings of the ACL 2010 Conference Short Papers*, ACLShort '10, p. 263–268, Stroudsburg, PA, USA : Association for Computational Linguistics.

JOACHIMS T. (2002). Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD on Knowledge discovery and data mining*, KDD '02, p. 133–142, New York, NY, USA : ACM.

KESSLER J. S. & NICOLOV N. (2009). Targeting sentiment expressions through supervised ranking of linguistic configurations. In *3rd Int'l AAAI Conference on Weblogs and Social Media (ICWSM 2009)*.

KIM S.-M. & HOVY E. (2006). Extracting opinions, opinion holders, and topics expressed in online news media text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, SST '06, p. 1–8, Stroudsburg, PA, USA : Association for Computational Linguistics.

MATHIEU Y. (2006). A computational semantic lexicon of french verbs of emotion. In W. B. CROFT, J. SHANAHAN, Y. QU & J. WIEBE, Eds., *Computing Attitude and Affect in Text : Theory and Applications*, volume 20 of *The Information Retrieval Series*, p. 109–124. Springer Netherlands.

MISHNE G. & GLANCE N. (2006). Predicting movie sales from blogger sentiment. In *AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW)*, p. 155–158.

PANG B. & LEE L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, **2**(1-2), 1–135.

RUPPENHOFER J., SOMASUNDARAN S. & WIEBE J. (2008). Finding the sources and targets of subjective expressions. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco : European Language Resources Association (ELRA). http//www.lrec-conf.org/proceedings/lrec2008/.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, p. 44–49.

TORRES-MORENO J. M., EL-BÈZE M., BÉCHET F. & CAMELIN N. (2007). Comment faire pour que l'opinion forgée à la sortie des urnes soit la bonne ? *Actes du 3ème Défi Fouille de Textes (AFIA 2007)*, p. 117–132.

TURNEY P. (2002). Thumbs up or thumbs down ? semantic orientation applied to unsupervised classification of reviews. *Proceedings 40th Annual Meeting of the ACL (ACL'02)*, p. 417–424.

VERGNE J. & GIGUET E. (1998). Regards théoriques sur le «tagging». In *Actes de Traitement Automatique des Langues Naturelles (TALN'98)*, p. 22–31.

VERNIER M. & MONCEAUX L. (2010). Enrichissement d'un lexique de termes subjectifs à partir de tests sémantiques. *Traitement Automatique des Langues*, **51**(1), 125–149.

WEHRLI E. (2007). Fips, a deep linguistic multilingual parser. In *Proceedings of the Workshop on Deep Linguistic Processing*, DeepLP '07, p. 120–127, Stroudsburg, PA, USA : Association for Computational Linguistics.

WIEBE J. M. & RILOFF E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. In *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, number 3406 in Lecture Notes in Computer Science, p. 486–497.

WILSON T. A. (2008). *Fine-grained Subjectivity and Sentiment Analysis : Recognizing the Intensity, Polarity, and Attitudes of Private States*. PhD thesis, University of Pittsburgh.