Analyse de discours évaluatif, modèle linguistique et applications

Stéphane Ferrari[*,***], Thierry Charnois[*,***]
Yann Mathet[*,***], François Rioult[*,***]
Dominique Legallois[**,***]

[*]GREYC - CNRS UMR 6072
[**]CRISCO EA 4255
[***]Université de Caen Basse-Normandie
{Prenom.Nom}@unicaen.fr

Résumé. Notre étude porte sur le discours évaluatif. L'approche que nous suivons est celle d'une analyse symbolique en vue d'un traitement sémantique de l'expression d'opinions.

Un travail préliminaire d'observation sur corpus a permis une première modélisation linguistique qui révèle la complexité du phénomène étudié. Nous en décrivons une mise en oeuvre destinée à constituer un outil d'aide à l'observation pour l'expert linguiste et permettre un retour sur le modèle.

Nous présentons et discutons ensuite deux approches informatiques testées pour la classification de documents d'opinions dans le cadre du défi DEFT07. La première s'appuie sur des n-grammes de mots, la deuxième reprend en partie le modèle linguistique, complété par un processus de classification fondé sur des règles d'association généralisées.

1 Introduction

Dans cet article, nous proposons un état de l'art synthétique pour mieux situer les différentes méthodes que nous avons développées pour l'analyse d'opinion. Nous présentons ensuite l'étude linguistique qui sous-tend notre approche, menée sur un corpus de 443 critiques de livres déposées par les internautes sur les sites fnac et amazon. Des formes récurrentes ainsi que des champs sémantiques privilégiés apparaissent qui caractérisent l'expression d'opinions, de jugements d'évaluation. Ces régularités en langue renseignent non seulement sur la polarité des sentiments exprimés, mais aussi sur d'autres propriétés comme leur intensité ou leur prise en charge par le locuteur. Une première mise en œuvre a partiellement validé ce modèle tout en mettant en évidence des limites dues à l'implémentation. Elle produit une qualification partielle de l'expression locale d'opinions au sein des textes en repérant différents motifs issus de l'étude linguistique, essentiellement à une échelle infra-phrastique.

La campagne d'évaluation DEFT07 (`http://deft07.limsi.fr`) a été l'occasion d'une exploitation de cette approche pour la classification de textes d'opinion. Cette édition 2007 du

DEfi Fouille de Texte proposait trois corpus de critiques sur lesquels nous avons pu tester l'ensemble de nos outils (films, livres, spectacles et bandes dessinées ; tests de jeux vidéo ; relectures d'articles). Nous présentons ici les grandes lignes de l'approche proposée, en commentant les résultats comparés des deux méthodes que nous avons pu y tester.

L'une avait pour objectif initial de préparer un *étalon*, un résultat de référence servant en quelque sorte à tester si notre approche méritait d'être présentée. Fondée sur des n-grammes discriminants, elle opère une classification de textes selon leur polarité (opinion positive, négative ou neutre), comme imposé par la tâche du défi.

L'autre, hybride, s'appuyait sur le modèle linguistique initial. L'outil informatique a été légèrement adapté au corpus de DEFT07 par la mise au point de ressources lexicales et de grammaires locales en adéquation avec les genres textuels qui y étaient proposés. Il a ensuite été complété par des outils de classification supervisée pour permettre d'exploiter les informations locales qu'il produit afin de déterminer la classe des textes analysés. Nous présentons l'ensemble de ces outils et commentons leurs résultats.

Dans une dernière partie, nous présentons des avancées plus récentes et quelques perspectives de poursuite de nos travaux. Des études complémentaires ont permis de dégager de nouvelles classes de régularités en rapport avec le discours évaluatif, comme des patrons syntaxiques privilégiés pour l'expression du jugement (au sens *Appraisal* (Martin et White, 2005), portant sur les individus ou les institutions) ou encore des classes de métaphores conceptuelles directement employées pour exprimer l'évaluation. Notre modèle s'est donc enrichi et continue de l'être dans le cadre d'une collaboration étendue pour s'orienter vers la mise en place d'outils d'analyse sémantique du discours évaluatif.

2 État de l'art

La quantité de travaux en rapport avec le discours évaluatif, l'expression d'opinions ou de sentiments, est telle que cet état de l'art ne peut être qu'à la fois partiel et partial. C'est donc dans cet esprit que nous proposons de dessiner les grandes orientations des travaux sur ce thème, en sélectionnant quelques approches particulières qui nous permettront ensuite de mieux situer celles que nous avons suivies dans notre propre étude.

Nous limitons cette présentation aux travaux en TAL et en fouille de texte. Ceux relatifs à la linguistique seront abordés dans la section suivante qui présente notre modèle.

Les travaux en informatique relatifs au phénomène de l'évaluation sont en majorité ceux traitant de la fouille d'opinion et de l'analyse de sentiments. Ils peuvent s'inscrire dans trois grandes catégories, pour reprendre une première classification proposée par Andrea Esuli[1] : (i) constitution de ressources lexicales pour la fouille d'opinion ; (ii) classification de textes d'opinions ; (iii) analyse d'opinion dans les textes. Nous compléterons cette présentation par un point de vue sur les propositions qui reposent sur des approches statistiques, ainsi que par quelques travaux faisant intervenir la notion de langage figuré.

[1] *cf.* http ://medialab.di.unipi.it/web/Language+Intelligence/OpinionMining06-06.pdf

2.1 Ressources lexicales

Dans la première catégorie, différentes approches sont proposées pour constituer des ressources de manière semi-automatique. On peut distinguer ces approches par le type de propriétés lexicales étudiées : le simple caractère subjectif ou objectif de termes comme dans (Baroni et Vegnaduzzo, 2004) (travail sur les adjectifs), la recherche de *patterns* subjectifs dans (Riloff et Wiebe, 2003), la polarité positive ou négative de termes subjectifs dans (Hatzivassiloglou et McKeown, 1997; Turney et Littman, 2003; Esuli et Sebastiani, 2005), ou encore préciser plus finement le sens de termes comme dans (Wiebe et Mihalcea, 2006) et dans les travaux relatifs à SentiWordNet (Esuli et Sebastiani, 2006), associant une combinaison des notions de subjectivité et de polarité à chaque SynSet de WordNet. L'atelier FODOP08 a été l'occasion de vérifier l'existence de travaux comparables sur le français, notamment dans (Harb et al., 2008).

Un peu en marge de ces travaux, on note une approche inspirée de la théorie de l'Appraisal (Martin et White, 2005) pour caractériser des expressions complexes dans (Whitelaw et al., 2005), notamment des groupes adjectivaux (p.e. : *not extremely brilliant*). Les caractéristiques annotées, plus nombreuses et plus complexes que dans les approches précédentes, sont l'attitude (affect, appréciation, jugement), l'orientation (positif, négatif), la graduation (force et focus) ainsi qu'une notion de polarité (prenant la valeur marquée ou non marquée, en fonction de la présence d'un marqueur de polarité comme notamment une négation). Dans une moindre mesure, l'approche proposée par Vernier et al. (2007b) associe un score à des expressions complexes en contexte qui rend compte de la graduation et de la polarité (p.e. dans *un livre vraiment très intéressant*, le groupe adjectival possède un score calculé sur la base de l'adjectif et de ses modificateurs, lui-même coefficienté si la tête du syntagme nominal est reconnue comme objet de la critique).

2.2 Classification de textes

De très nombreux travaux entrent dans cette catégorie, motivés en partie par les applications attendues dans différents domaines pour lesquels le suivi d'opinion est central. Notons en particulier DEFT07, campagne d'évaluation sur ce sujet pour la langue française. Les domaines d'applications sont très diversifiés : de l'analyse de critiques de film (Turney, 2002), (Pang et Lee, 2004) et une partie de DEFT07, à celle de textes politiques (un corpus de réactions à des propositions de lois dans DEFT07), en passant par les critiques de produits. Les techniques utilisées sont variées. Nombre d'entre elles sont issues des domaines de la fouille de données et de l'apprentissage automatique et utilisent des approches statistiques ou probabilistes. La dimension symbolique du TAL est au final peu présente. Des techniques telles que l'étiquetage grammatical et la recherche de syntagmes nominaux sont quelquefois exploitées à échelle locale, mais sont généralement absentes de la majorité de ces approches quantitatives. Certains travaux s'appuient sur les ressources lexicales, mais toujours dans un but de classification du texte dans sa globalité.

2.3 Analyse d'opinion au sein des textes

Une série de travaux s'attachent à déterminer le caractère objectif ou subjectif ou encore la polarité de mots, d'expressions complexes ou de phrases dans leur intégralité, en contexte. Les buts sont ici encore multiples, et les attentes nombreuses, à en juger par les usages des

industriels du domaine (Marcoul et Athayde, 2008). Les approches relatives s'attachent à la constitution de ressources plus fines que de simples mots comme dans (Riloff et Wiebe, 2003), à la classification de phrases et de propositions pour distinguer les opinions des faits dans un système de Question/Réponse (Yu et Hatzivassiloglou, 2003), ou pour proposer un résumé des points sur lesquels portent les critiques émises par les consommateurs dans les travaux de Hu et Liu (2004). Notons encore un travail sur la comparaison en tant que moyen d'évaluation (Jindal et Liu, 2006), qui devrait se poursuivre en une classification des comparaisons objectives et subjectives.

La pertinence des annotations proposées est liée à l'application attendue. On retrouve les classifications précédentes visant à déterminer le caractère subjectif ou la polarité. Une tendance se dégage cependant pour proposer une annotation plus fine que ces caractéristiques. Ainsi, dans (Wiebe et al., 2005), les auteurs utilisent des schémas pour annoter chaque expression d'une attitude (private state), à un niveau infra-phrastique, avec des renseignements sur la source (qui émet), la cible (sur laquelle porte l'opinion, et non pas la cible du discours), et des propriétés telles que l'intensité, la signification et le type d'attitude. Un corpus de 10 000 phrases annotées manuellement est disponible en téléchargement (au format GATE) . Parmi de nombreuses autres mesures, retenons que le taux d'agrément entre annotateurs sur le caractère subjectif des phrases est (κ) 0.77.

Des travaux comparables existent désormais pour la langue anglaise avec un schéma d'annotation directement inspiré de la théorie de l'Appraisal (Read et al., 2007). Une approche comparable existe pour le français (Maurel et al., 2007, 2008) ; les auteurs précisent avoir utilisé un format d'annotation inspiré de Wiebe et Mihalcea (2006) et Riloff et al. (2006) pour leur approche symbolique, incluant « les informations de cause, d'intensité et de l'émetteur du sentiment ».

2.4 Point de vue sur les approches statistiques

D'après Denoyer (2004), les approches statistiques considèrent généralement les textes comme des « sacs de mots », faisant fi de leur structure séquentielle : seule la présence, et éventuellement la fréquence, de chacun des mots est prise en compte. La représentation associée à un texte est alors un vecteur, qui peut être :

- binaire : on ne retient que la présence ou l'absence d'un mot, quel qu'en soit le nombre d'occurrences. C'est l'approche la plus ancienne, qui présente un bon compromis entre performance et complexité ;
- fréquentiel : il s'agit d'une extension du modèle binaire, dans laquelle les occurrences des termes sont comptées. Sa version normalisée permet de prendre en compte différentes tailles de textes sans induire de biais, chaque composante du vecteur étant pondérée par la taille du document ;
- TF-IDF : s'appuyant sur la loi de Zipf, qui indique que les termes les plus informatifs ne sont pas les plus fréquents, ces vecteurs privilégient les termes qui sont fréquents dans le texte qu'ils représentent, et peu fréquents dans les autres (Salton et Buckley, 1988; Joachims, 1998)

Un cadre d'application de cette approche vectorielle est SVM (machines à vecteurs supports), pour discriminer par apprentissage les échantillons d'entrée entre plusieurs classes (Trinh,

2007, 2008). À l'opposé des « sacs de mots », les représentations des documents s'envisagent ici de manière séquentielle. Elles sont notamment utilisées dans les chaînes de Markov et les réseaux Bayésiens. Ces approches issues de l'apprentissage ont l'avantage d'offrir des méthodes pouvant être mises en œuvre sur d'importants corpus, et donnent les meilleurs résultats dans des tâches telles que celle de DEFT07 où l'analyse d'une opinion porte sur un texte complet. Elles sont cependant pauvres d'un point de vue linguistique : une lemmatisation est parfois effectuée, ainsi que des traitements de surface préalables de « nettoyage » du texte.

2.5 Langage figuré

Différents travaux concernant l'expression d'opinion ou l'analyse de sentiments mentionnent l'utilisation récurrente de langage figuré, sans toutefois préciser comment prendre ce fait en considération. Les principales figures qui sont signalées sont la métaphore, l'ironie et le sarcasme (Wiebe et al., 2005). À l'inverse, quelques rares approches s'intéressent principalement au langage figuré et signalent cette fois son emploi possible notamment pour le traitement des opinions. Dans (Kreuz et Caucci, 2007), les auteurs s'intéressent au rôle du lexique dans l'expression du sarcasme, précisant que l'ironie est un mode privilégié pour l'expression d'opinion négative. Si l'étude ne permet pas de conclure en l'état, elle pointe un phénomène à surveiller pour l'analyse d'opinion (p.e., la polarité inversée de *Gee, I just love spending time waiting in line !*). L'approche de Vernier et al. (2007a); Vernier et Ferrari (2007) s'intéresse aux métaphores conceptuelles en étudiant l'opinion transmise par des expressions comme *laisser sur sa faim* ou *dur à avaler* dans un contexte de critiques d'objets culturels.

Dans la suite, nous commençons par présenter l'étude linguistique et le modèle sur lequel s'appuie nos travaux, ainsi qu'une première mise en œuvre.

3 Modèle linguistique

Dans cette partie, nous exposons quelques observations linguistiques dont les formalisations ont été implémentées dans la plate-forme de TAL LINGUASTREAM. Cette implémentation a pour objectif de constituer un premier outil d'aide à l'observation du phénomène évaluatif dans des genres textuels différents.

À partir de l'analyse d'un corpus constitué de 443 critiques d'internautes (51 092 mots), essentiellement de romans, mais aussi de BD, de poésie et d'essais – critiques extraites des sites amazon.fr et fnac.fr, nous avons considéré qu'il existe trois niveaux fonctionnels complémentaires et interactifs pertinents pour élaborer une « grammaire » de l'évaluation : ces trois niveaux sont :

1. Niveau des cadres expérientiels ;
2. Niveau des séquences lexico-grammaticales ;
3. Niveau des configurations énonciatives.

Ces niveaux correspondent aux méta-fonctions que distinguent Halliday (1994) : fonction idéationnelle (pour nous, cadre expérientiel), fonction textuelle (niveau lexico-grammatical), fonction interpersonnelle (niveau énonciatif).

3.1 Les cadres expérientiels

Le premier niveau identifie les aspects de l'objet évalué. Une analyse de l'évaluation d'un livre est vite confrontée à un problème inhérent à la constitution de l'objet même : on peut évaluer différents aspects ou qualia ; par exemple, le contenu, le style, la satisfaction ou la déception par rapport à des attentes, etc. L'évaluation peut porter également sur l'auteur du livre, sur l'histoire[2]. Autrement dit, la forme de l'expression d'un jugement est naturellement configurée par rapport à ce que nous avons nommé des cadres expérientiels. Quelques exemples de cadres :

L'emprise du livre sur le lecteur : *On ne peut plus le lâcher, jusqu'à la fin / Comme beaucoup d'entre vous, je suis tombée sous le charme de la douceur du récit de Philip Roth.*

Les attentes satisfaites ou non du lecteur : *Je reste de loin sur ma faim / Je m'attendais à mieux de K. DICK / J'ai été surprise par le style de ce livre / Vivement la suite !*

L'effort investi pour sa lecture : *Lisez le livre, il en vaut la peine / Le livre se lit facilement et rapidement / Il faut s'accrocher au début*

Son impact affectif sur le lecteur : *On pleure un peu, on rit, on s'émeut !...*

Sa valeur axiologique : *L'Aliéniste est avant tout un EXCELLENT roman.*

La prescription ou la proscription du livre (recommander un livre est une façon indirecte mais implacable de l'évaluer positivement) : *À conseiller pour ceux qui aiment les thrillers.*

Nous faisons l'hypothèse pour le défi DEFT'07 que ces cadres, même s'ils sont identifiés à partir d'un corpus précis, sont suffisamment généraux pour être appliqués à l'évaluation d'autres objets culturels ; en effet, l'observation d'avis portant sur des CD musicaux, des jeux vidéos ou des films permet de constater la présence de cadres identiques. Ce phénomène s'explique ainsi : l'évaluation porte rarement sur les propriétés intrinsèques de l'œuvre, mais sur les rapports que les sujets ont avec cette œuvre. De ce fait, les aspect jugés par la critique livresque sont facilement transposables à d'autres objets : efforts, impacts affectifs, prescriptions, attentes, mais aussi style, effets hédoniques (par ex. passer un agréable moment : *Voici le plus beau recueil de lettres au collège de pataphysique. Un réel moment de bonheur de découvrir ce monde inexploré* (à propos de Je voudrais pas crever de B. Vian)), *etc.* sont autant de cadres communs à l'expérience des objets culturels.

3.2 Séquences lexico-grammaticales

Le second niveau est celui des séquences lexico-grammaticales ; c'est ainsi que nous proposons une articulation du phénomène phraséologique à l'analyse de l'évaluation. À condition de ne pas voir dans la phraséologie un ensemble de formes radicalement figées, il est possible de concevoir des séquences lexico-grammaticales récurrentes, bien que polymorphes, dédiées ici à l'évaluation. Autrement dit, notre tâche a été de recenser les expressions « préfabriquées », de la simple collocation (par ex. *conseiller vivement*) aux configurations plus larges. Par ex.

on n'a jamais aussi bien rendu l'amour réciproque / Aucun livre de ma connaissance n'a jamais si bien démontré [...] les dégâts [...] que peuvent occasionner la vie

ce « motif » [ne jamais (aus)si bien + verbe de représentation / explication] est ici considéré comme une construction relativement ouverte, mais constituant malgré tout une unité

[2] *cf.* pour un développement (Legallois et Poudat, 2008).

prédonnée, directement disponible dans la compétence linguistique du locuteur. Les séquences lexico-grammaticales ont en partie été repérées grâce au logiciel « Collocates » qui permet d'identifier les n-grammes du corpus ; nous procédons à une vérification afin de nous assurer que les répétitions collocatives sont porteuses d'évaluation ou en sont des indices.

Parmi ces séquences, certaines sont entièrement dédiées à un cadre expérientiel, d'autres sont beaucoup plus indépendantes et peuvent s'actualiser dans plusieurs cadres. Nous donnons quelques exemples parmi les dizaines répertoriées (à noter que l'évalué renvoie à l'objet évalué, l'évaluatème à la valeur accordée à l'évalué, le siège à la personne qui « expérimente » l'évalué – le siège peut être ou non l'évaluateur) :

[à lire absolument] : cette séquence figée, employées 16 fois dans le corpus, s'actualise dans le cadre « prescription », comme la collocation [[Evaluateur [conseiller vivement] [Evalué]] [siège].

[ne pas pouvoir lâcher avant / jusque] : cette séquence (11 occurrences) s'actualise dans le cadre « emprise », et connaît plusieurs réalisations :

Pas question de lâcher le bouquin avant la fin.

Je n'ai pas pu le lâcher avant de l'avoir terminé.

On ne peut plus le lâcher, jusqu'à la fin.

On ne parvient à lâcher le roman qu'à la dernière page.

(Enfin / voilà / voici) un [évalué] qui [évaluatème] : il s'agit d'une construction à phrase averbale particulièrement récurrente dans le corpus (22 fois). Cette séquence s'actualise dans plusieurs cadres possibles : *un livre qui donne à rêver* (cadre « hédonique ») ; *un livre qui fait réfléchir* (cadre « valeur intellectuelle ») ; *un roman qui tiraille le lecteur entre notamment l'humour, l'amour, les rejets, les situations grotesques* (cadre « emprise »).

Det. ([enclosure]) [évaluatème] : cette séquence s'actualise principalement dans le cadre « valeur » : *Dix petits nègres est un vrai petit bijou* ; *un vrai petit Jules Vernes ou Barjavel* ; la présence de l'enclosure ici, est un indice imparable de la fonction évaluative du terme subséquent. Ainsi, *Jules Vernes / Barjavel* sont-ils étiquetés évaluatèmes.

Nous recensons ainsi près d'une trentaine de séquences évaluatives ou introductrices d'évaluation dont les rôles thématiques sont étiquetés, non pas à partir de catégories générales (par ex. agent, bénéficiaire, etc.), mais à partir de rôles propres à l'expression de l'évaluation. Ces séquences sont de dimensions et de natures hétérogènes : du syntagme récurrent à la phrase figée. Là encore, une projection sur d'autres textes (projection qui n'est pas encore systématisée à l'heure actuelle) permet de voir des constructions fort apparentées sémantiquement et grammaticalement ; par exemple, au sujet de l'audition du requiem de Mozart :

Cette interprétation du requiem K626 est un véritable feu d'artifice. J'en suis resté scotché sur mon fauteuil. Bravo ! (amazon.fr)

Ou à propos du jeu vidéo Morrowind :

Ce jeu est tout simplement magnifique : si vous avez une X-Box, Morrowind est incontournable. Les graphismes sont superbes et l'ambiance vous immerge totalement dans l'univers. Les quêtes sont très variées et le joueur ne s'ennuie jamais : il y a toujours quelque chose à faire ! ! ! Je suis resté scotché sur ce jeu pendant toute une semaine et je suis même pas au 1/4 du jeu ! Je le recommande même à ceux qui ne sont pas spécialement fan du genre : vous ne serez pas déçu ! (amazon.fr)

Ainsi, dans la perspective d'une implémentation rendant compte de l'évaluation de tout objet culturel, il est important d'assigner aux deux séquences *ne pas pouvoir lâcher / rester scotché* une catégorie subsumant les diverses réalisations. C'est par ce travail de généralisation que pourra être établie une systématicité valant pour l'ensemble des objets culturels.

3.3 Configurations énonciatives

Le niveau énonciatif est fondamental pour une analyse générale du discours évaluatif de l'objet culturel. Les évaluations, en tant qu'actes de discours, doivent être mesurées selon leur force illocutoire. C'est à ce niveau que s'articulent et se construisent les stratégies argumentatives. Il s'agit, pour le locuteur, de se mettre en scène pour faire partager son avis : premier plan, engagement, retrait, prise en charge faible, *etc.* Cette mise en scène, dans notre corpus, est relativement normée dans la mesure où le genre est lui-même partiellement stéréotypé ; mais là encore, la formalisation du niveau énonciatif devra permettre toute projection vers d'autres objets afin d'élaborer des points de comparaisons et de différences.

Ainsi, par exemple :

Les marqueurs restreignant au seul énonciateur la validation de l'énoncé : *À mon goût, à mon avis, selon moi*

Les marqueurs délimitant le public intéressé : *une mine d'informations pour tous ceux qui s'intéressent à la psychologie en général*

Les verbes d'attitude propositionnelle (impliquant la modalité épistémique) : *Je crois que Philip Roth a atteint le sommet avec Opération Shylock*

Les tournures concessives : *Ce bouquin est certes intéressant au début, mais il devient très vite rébarbatif.*

Les adverbes intensifs (marquant explicitement le degré d'engagement de l'énonciateur) : *Vraiment, véritablement, absolument, impérativement, totalement, etc.*

Pronoms personnels (l'évaluateur peut s'effacer devant l'expérimentateur, attribuer le jugement à une instance collective, projeter une évaluation du destinataire, *etc.*) : *Plus vous avancerez dans la lecture, plus vous serez dégoûtés par ce simili d'érudition prétentieux et bourré de fautes !*

Les interjections : *Vraiment, beurk…*

Ce niveau est le plus complexe des trois à formaliser dans la mesure où les formes sont extrêmement hétérogènes, de dimensions parfois larges, dépassant le simple énoncé. La « stratégie » consiste en fait à s'appuyer le plus possible sur les séquences lexico-grammaticales, qui constituent à notre avis, le niveau intermédiaire entre niveau des cadres expérientiels et niveau des configurations énonciatives.

3.4 Applications informatiques

La première application directe en informatique de ces observations a été la conception d'un outil de repérage des différentes structures, dont la visée première était plus de fournir un outil d'aide à l'observation sur de nouveaux corpus qu'une réelle analyse automatisée du phénomène. Nous en présentons les grandes lignes dans la suite, et nous renvoyons à Legallois et Ferrari (2006) pour une présentation détaillée de cette première application.

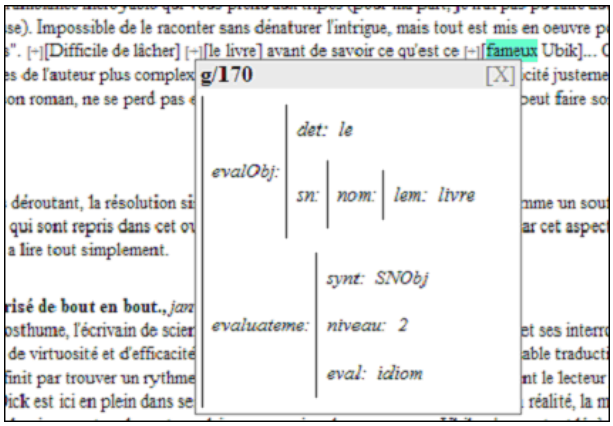

FIG. 1 – *Outil d'aide : exemple de résultats.*

Par la suite, la campagne d'évaluation DEFT07 a permis de proposer une adaptation de cette première mise en œuvre ayant cette fois une visée toute différente : la classification de documents d'opinion. Ce point fait l'objet de la section suivante.

Pour la mise en œuvre informatique issue de cette première étude, nous avons opté pour une utilisation de la plate-forme de TAL LINGUASTREAM [3], afin de réaliser une chaîne de traitements facilement réutilisable, et dont chaque ressource puisse être aisément modifiée, au gré des observations de l'expert linguiste sur de nouveaux textes. Ce sont essentiellement les deux premiers niveaux de la grammaire qui sont exploités, à savoir le niveau sémantique et le niveau lexico-grammatical. Ils fournissent des indices textuels qui permettent de déclencher d'autres analyses locales dont l'objectif final est de détecter les différents éléments de discours du niveau énonciatif. La chaîne d'analyses contient en particulier les composants suivants :

- une analyse de catégories grammaticales et une lemmatisation, à l'aide du TreeTagger (Schmid, 1994).
- une recherche d'indices textuels à l'aide d'expressions régulières. C'est dans cette phase que sont exploitées conjointement les ressources sémantiques du niveau 1 et les ressources lexico-grammaticales du niveau 2.
- une analyse syntaxique « locale » afin de déterminer certains éléments des *patterns* de niveau 2 et 3. Cette analyse se fait à l'aide d'une grammaire Prolog enrichie de l'extension GULP (Covington, 1994).

Chaque composant produit des annotations sous forme de structures de traits, que les composants ultérieurs exploitent à leur tour. L'exemple de la figure 1 illustre ce principe. Dans la phrase « difficile de lâcher le livre avant de savoir ce qu'est ce fameux Ubik. . . », une expression

[3] Nous renvoyons à (Ferrari et al., 2005; Widlöcher et Bilhaut, 2005, 2006, 2008) pour plus de détails sur cet environnement.

régulière permet de repérer la construction « difficile de lâcher », à laquelle est alors attachée une structure de trait contenant des informations sur le type d'analyses à mener ultérieurement. L'analyse syntaxique locale, qui s'appuie sur les résultats du TreeTagger, permet ici de déterminer un des actants de l'évaluation : « le livre » est considéré comme objet de l'évaluation introduite par l'expression idiomatique précédente.

L'objectif de cette première application, fournir un cadre d'observation pour permettre d'enrichir le modèle en le confrontant à de nouvelles données textuelles, ne nécessite pas que l'ensemble des informations repérées soient mises en relation pour construire une représentation plus élaborée des évaluations repérées. Nous avons donc préféré conserver une chaîne d'analyse relativement légère dont les ressources sont faciles à modifier. En effet, tant en ce qui concerne les lexiques que les grammaires exprimées sous forme de macro-expressions régulières, la plate-forme LINGUASTREAM que nous avons utilisée fournit des interfaces graphiques qui en permet la manipulation. Seule l'analyse syntaxique locale reste difficile à faire évoluer, ou du moins nécessite des compétences en TAL peu accessibles pour les experts linguistes, mais le caractère systématique des règles exprimées permet cependant une maintenance relativement aisée.

Dans la section suivante, nous présentons une adaptation du modèle et de la chaîne d'analyse en vue d'une application dans le cadre de la classification de documents d'opinion.

4 Classification de documents d'opinions

Cette section présente la mise en œuvre de notre traitement sémantique sur les textes du DÉfi Fouille de Texte 2007. Les participants de DEFT privilégient généralement l'application de modèles purement numériques. Cette campagne était l'occasion de montrer la faisabilité d'une approche symbolique pour la fouille de textes.

La tâche proposée pour DEFT07 consistait à déterminer l'opinion globale portée par des textes, sur quatre corpus de genres distincts, correspondant à quatre tâches indépendantes. Par « opinion globale », nous entendons l'avis porté par un texte dans sa globalité (sans par exemple s'attacher aux sous-rubriques éventuelles sur lesquelles se fonde l'opinion), fourni par une note venant en parallèle au texte. Pour trois des corpus, trois classes sont proposées (avis positif, neutre ou négatif), tandis que pour un autre, seules deux classes sont présentes (pour ou contre). Un corpus d'apprentissage (60% du total) était fourni, chaque texte catégorisé, et la tâche consistait à proposer une catégorisation automatique des corpus de test (les 40% restants).

Pour évaluer notre approche symbolique, nous la comparons à une méthode numérique *étalon* à base de n-grammes émergents. L'évaluation des méthodes linguistiques est coûteuse, cependant les méthodes numériques fournissent de bonnes performances qu'il est pertinent d'utiliser comme référence.

Pour permettre le passage d'une analyse locale des expressions d'évaluation à la détermination de l'orientation générale d'un énoncé plus conséquent, les informations fournies par notre approche symbolique nécessitent une interprétation. Aucune étude sur le discours évaluatif ne nous permettait de proposer des règles pour effectuer ce changement de dimension. Nous

utilisons donc un modèle à base de règles d'association généralisées pour effectuer la classification. L'observation des résultats nous a permis de dégager quelques régularités intéressantes sur lesquelles nous revenons dans la section suivante.

Les résultats obtenus avec l'une et l'autre des deux approches testées, comparables à ceux des autres participants à DEFT07, justifient au final que nous les présentions ici toutes les deux. Nous espérons ainsi apporter un éclairage intéressant sur deux façons bien distinctes d'analyser automatiquement les opinions portées par des textes. Nous renvoyons à (Vernier et al., 2007b; Ferrari et al., 2008) pour d'autres présentations de ces approches.

4.1 Une approche étalon, par n-grammes émergents de lemmes

La méthode *étalon* fondée sur les n-grammes a vocation à faire apparaître les suites de mots qui sont très caractéristiques d'une catégorie d'opinion donnée. À l'origine, il s'agissait pour notre équipe de créer, en amont de nos pistes de réflexion, un outil facilitant l'observation et l'analyse linguistiques, en faisant émerger des combinaisons syntagmatiques particulièrement révélatrices. Cet outil, enrichi de certains traitements, s'est finalement révélé suffisamment puissant pour constituer à lui seul un classifieur relativement performant.

Il est important d'insister sur le fait que notre équipe est loin d'être spécialiste de la question, et que notre traitement à base de n-grammes s'est mis en place de façon progressive suite à la pertinence constatée des indices que fournissent ces derniers pour la tâche DEFT, et sans doute de façon très naïve par rapport aux nombreuses études existant dans le domaine suite à (Rocchio, 1971). Aussi, la méthode présentée ici ne correspond pas aux standards du domaine, mais comporte des aspects originaux. Il serait donc intéressant dans le futur d'étudier tant d'un point de vue théorique que du point de vue des résultats les différences qu'induisent nos choix par rapport à ces derniers. En particulier, il sera important de mettre en parallèle le système de pondération et de discrimination expliqués ci-après avec les calculs vectoriels proposés par Rocchio.

4.1.1 Classification par n-grammes émergents

La technique des n-grammes consiste à observer les collocations contiguës sur une fenêtre de n tokens consécutifs d'un flux, et à essayer de tirer de ces observations des régularités relatives à un aspect particulier de ce flux (Stubbs et Barth, 2003). Dans le cas présent, l'analyse d'opinions, les tokens pris en compte sont les mots et ponctuations du texte. Par exemple, certains n-grammes seront caractéristiques de tel type de corpus car très récurrents dans ce dernier, et beaucoup plus rares ailleurs. Dans le cadre de l'analyse d'opinions, le flux d'entrée est un matériau linguistique (textes écrits en français), et nous essayons de catégoriser les différents textes de ce flux selon le jugement porté par leur auteur. Pour illustrer de façon très simplifiée l'hypothèse de cette approche, nous espérons trouver des n-grammes caractéristiques d'un jugement favorable, défavorable ou neutre.

Par exemple, pour des articles relatifs à des critiques de livres, et après analyse automatique des corpus d'apprentissage, nous trouvons des tri-grammes caractéristiques tels que « une vraie catastrophe » (catégorie 0 : avis négatif), « roman assez moyen » (catégorie 1 : avis neutre) et « très belle œuvre » (catégorie 2 : avis positif). Lors de l'analyse d'un texte du corpus de test, le tri-gramme « très belle œuvre » aura tendance à ranger ce texte en catégorie 2. Bien sûr, il y a un risque qu'un même texte contienne des n-grammes de différentes catégories, rendant le

choix plus difficile. L'idée que nous mettons en œuvre pour pallier cette difficulté est de deux ordres :

1. ne retenir pour chaque catégorie que les n-grammes les plus émergents ou discriminants, *i.e.* moins susceptibles d'apparaître dans des textes d'autres catégories ;
2. pondérer les n-grammes, *i.e.* associer à chacun un poids d'autant plus important qu'il apparaît fréquemment dans sa catégorie relativement aux autres.

4.1.2 Apprentissage

Conformément aux souhaits que nous avons formulés précédemment, un traitement ultérieur a vocation à déterminer quels sont les n-grammes émergents pour une catégorie, par rapport **aux autres**. Cette notion de vis-à-vis est très importante pour la tâche que nous avons à réaliser. En effet, trouver des n-grammes représentant une catégorie en toute généralité serait bien moins performant que de trouver des n-grammes opposant une catégorie aux autres.

Ce traitement fait ressortir les n-grammes émergents selon le principe suivant :

- pour chaque n-gramme d'une catégorie, regarder s'il est présent dans les autres ;
- s'il est absent des autres catégories, et que son nombre d'occurrences dans le premier corpus est supérieur à un certain seuil paramétrable (par exemple réglé sur 1 pour éviter les orphelins), lui attribuer le poids INFINITY ;
- s'il est présent, lui associer un poids égal au taux d'émergence, *i.e.* le rapport entre ses fréquences relatives dans la catégorie caractérisée et les autres. Ne le garder que si le poids ainsi calculé est supérieur à un certain seuil paramétrable.

Prenons un exemple : le trigramme « une vraie catastrophe » apparaît 12 fois dans la première catégorie, donnant lieu à une fréquence relative de 12/13247 (cette catégorie comportant 13247 trigrammes), et seulement 2 fois dans les autres, donnant lieu à une fréquence relative de 2/17523. Ce trigramme se verra ainsi attribuer un poids égal au rapport de ces deux fréquences relatives, soit (12/13247) / (2/17523), c'est-à-dire $7,93$. Cela signifie que l'on a pratiquement 8 fois plus de chances de trouver ce trigramme dans un texte du premier corpus que du second. Si cette valeur est supérieure au seuil que nous avons fixé, ce trigramme sera donc conservé comme trigramme discriminant, et son poids de $7,93$ lui sera associé.

Prenons un autre exemple : le trigramme « très belle œuvre » apparaît 4 fois dans le premier corpus, et jamais dans le second. Si 4 est supérieur au seuil paramétrable, nous conservons ce trigramme et lui associons le poids INFINITY (on a une infinité de chances supplémentaires de trouver ce trigramme dans le premier corpus que dans le second), valeur fixée dans la pratique non pas à l'infini, ce qui interdirait la prise en compte d'autres n-grammes, mais à 15, après une série de tests.

4.1.3 Classification : choisir une catégorie

Lors de l'analyse d'un texte, il est ensuite tenu autant de comptes qu'il y a de catégories. Pour chaque catégorie, nous établissons la somme des poids de tous les n-grammes de cette catégorie trouvés dans le texte. Il vient souvent à l'esprit, lorsque l'on pense en termes probabilistes, d'effectuer le produit des poids plutôt que la somme. Cependant, nous n'établissons

pas ici une probabilité (qui résulterait effectivement en un produit), mais nous basons sur une ensemble d'indices qui sont, pour chaque classe, en nombre variable.

Imaginons par exemple que pour la catégorie 0 (négatif), nous ayons 5 n-grammes discriminants, tous de poids 3 (leur produit donnerait 243, et leur somme 15), et pour la catégorie 1 (neutre), nous ayons seulement 2 n-grammes discriminants, mais de poids très marqués, 10 (leur produit donnerait 100, leur somme 20). Notre exemple montre que le choix de la somme peut donner plus de force à un nombre de n-grammes discriminants limité, mais de poids importants, qu'à de plus nombreux n-grammes de poids plus faible. Les tests ont confirmé la pertinence de ce choix.

De la sorte, nous obtenons une note globale pour chacune des catégories, que nous pouvons mettre en balance avec les notes globales obtenues pour les autres catégories.

4.2 Adaptation du modèle linguistique

Nous présentons ici l'adaptation du modèle linguistique (*cf.* Section 3) que nous avons réalisée pour participer à DEFT07. Après une discussion sur l'adapation d'un tel modèle, nous détaillons les caractéristiques de notre classifieur à base de règles d'association généralisées.

4.2.1 Principes généraux

Dans la même lignée que certains travaux sur l'analyse de sentiments, l'adaptation du modèle linguistique consiste en une approche symbolique fondée sur des indices textuels pour l'expression de l'opinion : adjectifs, noms, adverbes, verbes ou expressions porteurs d'une opinion. Ces indices constituent les ressources lexicales propres à l'évaluation, et permettent de préciser notamment des propriétés de subjectivité, d'orientation et éventuellement d'intensité. Lors d'une analyse locale, ils peuvent être renforcés par d'autres éléments, de deux types : des modificateurs d'intensité (essentiellement adjectif, adverbe ou locution adverbiale) et des termes du domaine cible des textes considérés.

Ainsi, le nom « bonheur » peut, hors contexte, être considéré comme indice potentiel d'évaluation positive, mais son usage dans une phrase telle que « l'héroïne est tout simplement en quête du bonheur » ne sera renforcé par aucun autre. En revanche, dans l'expression « un pur bonheur », l'adjectif vient renforcer son intensité. De la même manière, « mauvais » et « probant » sont renforcés dans les contextes suivants extraits des corpus DEFT07 : « un papier véritablement mauvais » et « approche ne me semble guère probante », où « papier » et « approche » peuvent être considérés comme des termes du domaine.

L'étude linguistique initiale permet aussi théoriquement de fournir des informations d'une autre nature (évaluateur, évaluatème et configurations énonciatives), en précisant certaines propriétés relatives à l'émetteur et à l'objet d'une évaluation : engagement ou prise en charge par l'énonciateur, focus ou facette particulière de l'objet concernée par l'évaluation...

4.2.2 Adaptation pour la classification de textes

Dans le cadre de DEFT07, seules les propriétés d'orientation et d'intensité ont pu être retenues, la diversité des corpus nécessitant un long travail d'analyse linguistique pour mettre en place des ressources telles que suggérées précédemment. Le coût de l'adaptation et de la mise

en œuvre informatique présentées ici reste relativement léger, estimé à environ deux hommes mois.

Nous avons donc limité nos analyses locales à une détection d'indices et à leur renforcement, sans qualifier les cibles ni les champs sémantiques. N'ayant pas de stratégie pour déterminer le caractère global de l'opinion à partir des indices locaux, nous avons décidé de déléguer la tâche de classification proprement dite à un processus automatique de classification supervisée. Afin de permettre une telle utilisation des indices, des scores numériques leur ont été attribués de manière heuristique, avec comme objectif de rendre compte de leur intensité en contexte mais aussi d'aténuer le poids des éventuels faux indices, comme la présence d'adjectifs ne qualifiant pas un objet du domaine. Le score d'un indice est positif ou négatif selon l'orientation du jugement (signe éventuellement modifié par la présence d'une négation), multiplié par d'éventuels premiers facteurs rendant compte de son intensité (présence de modificateurs ou d'enclosures), puis par un éventuel dernier facteur lorsque la cible de l'opinion est reconnue comme un objet du domaine (voir tableau 1). Ainsi, pour reprendre les exemples précédents, « bonheur » voit son score initial de +1 multiplié par +2 dans « un pur bonheur » (enclosure), tandis que l'expression « un papier véritablement mauvais » permet de dégager un score local de -8 (modificateur d'intensité et objet du domaine reconnu).

« un pur bonheur » :			
pur	→	*coefficient (intensité)*	2
bonheur	→	*évaluation intrinsèque*	1
		score de l'indice →	**2**
« un papier véritablement mauvais » :			
papier	→	*coefficient (terme général du domaine)*	4
véritablement	→	*coefficient (intensité)*	2
mauvais	→	*évaluation intrinsèque*	-1
		score de l'indice →	**-8**
« approche ne me semble guère probante » :			
approche	→	*coefficient (terme partiel du domaine)*	2
ne … guère	→	*coefficient (négation)*	-1
probant	→	*évaluation intrinsèque*	1
		score de l'indice →	**-2**

TAB. 1 – *Exemples d'indices détectés et scores localement attribués.*

Cette analyse produit, pour tous les textes d'un corpus, un score positif et un score négatif sur l'ensemble de l'énoncé, obtenu en faisant la somme des scores des indices trouvés. Les scores faibles des indices incertains sont ainsi aténués face à ceux plus élevés d'indices portant sur une cible reconnue, mais pas totalement ignorés, la reconnaissance des cibles n'étant pas systématique. Du point de vue de l'analyse du discours, il nous a semblé cohérent de préciser également des scores propres à certaines parties du discours qui peuvent marquer plus fortement l'évaluation ou ayant des chances de refléter au mieux l'opinion associée à l'énoncé. En général, le premier et le dernier paragraphe (« introduction » et « conclusion ») ont ainsi un score qu'il peut être intéressant de préciser indépendamment du score général. L'hypothèse émise est que l'auteur aura tendance à annoncer la couleur de son opinion dès les premiers

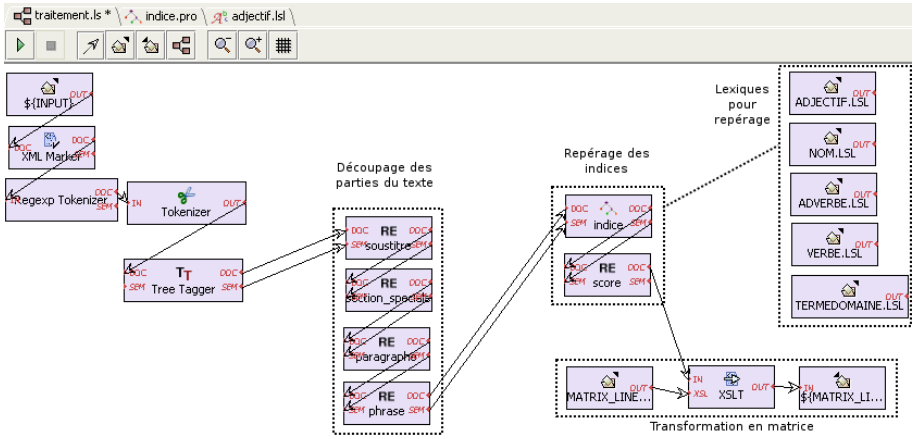

FIG. 2 – *Chaîne d'analyse dans la plate-forme* LINGUASTREAM. *Cette figure montre les différents composants qui s'enchaînent séquentiellement (de la gauche vers la droite) : textes en entrée, puis tokenisation, analyse lexico-grammaticale (TreeTagger) suivie du découpage du texte (expressions régulières), repérage des indices et calcul du score (grammaire DCG et expressions régulières), et en sortie les matrices utilisées pour la classification.*

instants de l'énoncé, et qu'il pourra éventuellement synthétiser ses arguments en fin de texte. L'ensemble de ces traitements donnent lieu à la chaîne présentée dans la figure 2.

Au final, ce sont 6 mesures rendant compte des opinions positives et négatives dans 3 zones textuelles différentes (dont le texte intégral) qui sont calculées pour chaque texte de chaque corpus d'apprentissage et qui constituent l'entrée d'un extracteur de règles généralisées (Rioult et al., 2008) pour produire un classifieur par corpus (Vernier et al., 2007b).

4.2.3 Classification supervisée à base d'associations généralisées

Bien que les méthodes à base de *motifs ensemblistes* soient peu utilisées dans le domaine du texte – les participants de DEFT plébiscitent plutôt les méthodes à noyau – elles n'en sont pas moins performantes et offrent le net avantage de fournir un modèle interprétable à l'expert. La méthode de décision n'est plus une boîte noire et les interactions entre expert et fouilleur sont bien plus vivantes. Nous décrivons donc ci-dessous la méthode de classification supervisée à base de règles d'association généralisées que nous avons utilisée.

Les techniques de motifs et de règles s'appliquent sur des contextes booléens, qui nécessitent une discrétisation des attributs quantitatifs. Les bases de données obtenues recensent des objets décrits par des attributs booléens ; un *motif* est une conjonction d'attributs. La communauté fouille de données dispose depuis une douzaine d'années d'algorithmes performants pour extraire les motifs fréquents (les motifs présents dans un nombre minimum d'objets) et construire les règles d'association. De la forme $X \rightarrow Y$, ces règles sont mesurées par une fréquence, indiquant le nombre d'objets contenant $X \cup Y$, et une confiance, probabilité condi-

tionnelle d'apparition de Y connaissant celle de X. Lorsque ces règles concluent sur un attribut de classe, elles peuvent être utilisées pour construire un classifieur automatique.

Plusieurs méthodes existent pour classer à partir de règles associations. Historiquement, la première et la plus simple est CBA ((Liu et al., 1998)) (Classification Based on Association). Cette méthode extrait les règles d'association de fréquence et confiance minimales indiquées par l'utilisateur, et ordonne ces règles suivant leur confiance. Lorsqu'un nouvel exemple se présente, la première règle qui peut s'appliquer propose une valeur de classe.

Ce procédé a été raffiné par la méthode CMAR ((Li et al., 2001)) (Classification based on *Multiple* class-Association Rules) qui ne se contente plus d'une seule règle pour prendre la décision de classification. Les règles sont cette fois-ci pondérées par un indice de corrélation fourni par un χ^2 normalisé. On évite également la redondance entre les règles en ne conservant que celles qui sont à prémisse minimale. Un procédé de *couverture* sélectionne les règles essentielles pour la tâche et fournit un nombre homogène de règles pour chaque classe même si la distribution est légèrement deséquilibrée. Un nouvel exemple sera classé à l'issue d'un vote réalisé par toutes les règles qui s'appliquent, selon leur pondération.

Pour nos expériences, nous avons implémenté une méthode proche de CMAR, mais qui utilise des règles d'association généralisées (ou disjonctives). Contrairement aux règles d'association classiques, les règles généralisées sont de la forme $X \rightarrow \vee Y$ et concluent sur une disjonction d'attributs plutôt que sur une conjonction.

Ces règles présentent le modèle sous une forme normale conjonctive car $X \rightarrow \vee Y$ équivaut à $\overline{X} \vee Y$ soit $\overline{x_1} \vee \ldots \vee \overline{x_m} \vee y_1 \ldots y_n$. Cette généralisation couvre les modèles utilisant des règles de forme quelconque.

Considérons la règle généralisée suivante :

$$longIntro <= 70 \longrightarrow negatifIntro = 0 \vee negatifIntro = 1 \vee categorie = negatif.$$

Elle exprime, avec une mesure $\chi^2 = 7,16971$, qu'un texte dont l'introduction est courte ($<=$ 70 mots) est : soit de catégorie négative, soit le nombre de termes négatifs dans l'introduction vaut 0 ou 1. Elle est reformulée en règle de classification sous la forme :

$$longIntro <= 70 \wedge \overline{negatifIntro = 0} \wedge \overline{negatifIntro = 1} \longrightarrow categorie = negatif.$$

La forme très générale des règles peut nuire à leur intelligibilité par l'expert. En effet, une règle équivalente et plus lisible serait :

$$longIntro <= 70 \wedge negatifIntro \geq 2 \longrightarrow categorie = negatif.$$

Cependant, la généralité de ce modèle est précieuse lors de la phase d'interaction avec l'expert. Non seulement il indique ici qu'un texte avec quelques mots négatifs dans une introduction courte est négatif, mais il suggère aussi des transformations de la nature des attributs. Ici, la différence entre 0 et 1 mot négatif n'est pas pertinente.

L'utilisation de règles généralisées pour calculer un modèle offre ainsi de nombreux avantages (Rioult et al., 2008) :

- des règles excluant des classes sont disponibles. Cependant, la sémantique introduite produit des améliorations mineures en classification supervisée, ce qui laisse à penser que cette sémantique est équivalente à celle du problème de classification en inversant les valeurs de classe ;
- la sémantique des règles est étendue par la présence de négations en prémisse et améliore significativement les performances, notamment sur les jeux de données possédant peu d'objets ou peu d'attributs. En effet, il y a beaucoup plus de règles contenant des attributs négatifs que de règles ne contenant que des attributs positifs.

Ce dernier argument justifie l'utilisation de règles généralisées pour la fouille de texte. D'une part certains corpus peuvent contenir peu d'objets, ce qui rend difficile l'obtention d'un modèle performant ; d'autre part les indices sémantiques sont moins nombreux que les indices plus numériques comme les n-grammes. Certains corpus, dont l'une des dimensions est faible, profiteront avantageusement d'un modèle calculé par des règles de sémantique généralisée.

4.3 Résultats sur les données de DEFT07

Le tableau 2 compare les F-scores obtenus, sur les trois premiers corpus[4], entre des juges humains, la méthode étalon à base de n-grammes émergents, et l'approche symbolique. L'approche symbolique n'a été utilisée que pour le corpus 3 dans DEFT07, une approche hybride entre n-grammes et symbolique a été exploitée pour les tests, donnant lieu aux scores de la dernière colonne du tableau. Ces scores ont permis à notre équipe d'obtenir la quatrième place parmi dix concurrents. Les résultats de l'approche symbolique sur les corpus 1 et 2 ont été établis par nos soins.

	corpus	humain	n-grammes	symbolique	mixte
1	objets culturels	0,52 - 0,79	0,577	*0,457*	0,532
2	jeux vidéos	0,73 - 0,90	0,761	*0,506*	0,715
3	relectures d'articles scientifiques	0,41 - 0,58	0,414	0,474	NA

TAB. 2 – *Comparaison des F-scores entre le juge humain, l'étalon n-grammes et l'approche symbolique.*

Sur les deux premiers corpus, l'approche n-grammes montre clairement sa supériorité et tient le rôle d'étalon. Certes, les résultats de l'approche symbolique doivent être améliorés, mais ils montrent que la démarche est pertinente et les indices sémantiques produisent des scores significatifs.

De manière assez inattendue, c'est sur le corpus 3 (relectures d'articles scientifiques) que l'approche symbolique obtient les meilleurs résultats relatifs, alors que les études ont été menées plus particulièrement sur le corpus 1 (critiques d'objets culturels), dont le thème est aussi assez proche du corpus 2. Après observation des modes d'expression utilisés dans ce corpus 3, nous pensons que les relecteurs y font preuve d'une certaine retenue dans leurs propos, les tournures très négatives semblent assez peu employées. Aussi, le traitement local que nous proposons pour le calcul d'un poids à attribuer à chaque expression d'évaluation a probablement

[4]le temps de calcul des associations sur le corpus 4 était incompatible avec le court délai du défi.

permis une meilleure prise en considération de ce style retenu, tant en y repérant les évaluations faisant intervenir des négations explicites qu'en pondérant certaines évaluations positives faites sous forme de concessions.

Nous expliquons partiellement les scores peu élevés par la pauvreté des ressources développées pour cette tâche spécifique (les thèmes multiples du corpus 1, spectacles, films, livres, disques, n'ont pas tous été étudiés, et les tournures de langage très particulières du corpus 2, s'adressant plutôt à un public jeune, n'ont pas non plus fait l'objet d'études spécifiques). Dans cette perspective, les résultats nous paraissent honorables et confortent l'idée que les indices étudiés sont au moins pertinents pour la tâche.

Comparativement avec l'approche par n-grammes, c'est aussi sur le corpus 3 uniquement que l'approche symbolique permet d'avoir de meilleurs résultats. Ce corpus étant plus petit que les autres, il est difficile de mettre en œuvre une approche statistique sur un petit nombre de données.

Une étude des mauvais résultats montre qu'ils concernent plus particulièrement les textes de la classe neutre. Plutôt que neutres, il s'agit en réalité de textes mitigés, contenant des indices des deux orientations. C'est pourquoi nous envisageons une poursuite de nos travaux dans le sens d'une analyse textuelle permettant de dégager les zones de textes *cohérentes en terme d'opinion*.

L'approche mixte testée pour compléter les soumissions proposées lors de DEFT07 a globalement dégradé les résultats. Sur le corpus 3, de petite taille, il ne nous a pas été possible de l'évaluer sur un échantillon significatif de textes, la majeure partie du corpus initial ayant été consommée pour l'apprentissage ; c'est pourquoi elle n'a pas figurée dans les soumissions à DEFT07. Cette piste peu encourageante a été abandonnée depuis.

5 Conclusion

À partir d'un état de l'art sur le thème de la fouille d'opinions, nous avons montré que la tendance actuelle est celle des approches probabilistes. La démarche que nous avons présentée est donc fondamentalement différente dans la mesure où nous nous appuyons sur une modélisation linguistique pour mettre en œuvre une analyse informatique de l'opinion, et plus généralement du discours évaluatif. Nous avons décrit différentes réalisations allant de l'outil d'aide à l'observation destiné à l'expert linguistique, à la classification de documents d'opinions dans le cadre de la campagne DEFT07. Pour cette tâche étalonnée par une méthode de n-grammes émergents, nous avons proposé une combinaison de traitement automatique de la langue et de classification à base de règles d'association généralisées.

Nos travaux se poursuivent actuellement dans deux directions. La première porte sur l'étude du langage figuré dans l'expression de l'opinion. En effet, sur l'un des quatre corpus de DEFT07, constitué de critiques de films, de livres et d'autres objets culturels, notre analyse a mis en évidence de nombreux emplois métaphoriques, comparables à ceux déjà observés dans l'étude initiale sur les corpus fnac.fr et amazon.fr. Nous avons répertorié le lexique relatif en fonction des domaines sources des métaphores conceptuelles utilisées, selon la terminologie empruntée à Lakoff et Johnson (1980) : « Un objet culturel C'EST de la nourriture pour l'esprit », « Lire un livre, regarder un film… C'EST un voyage, une aventure », *etc.*

L'étude de ces emplois métaphoriques permet de dégager un modèle des cibles de l'opinion exprimée dans le genre textuel étudié : la réception de l'objet culturel dans des expressions telles que *laisser sur sa faim*, *dur à avaler*, *délicieux*, sa création dans *concocter*, *mijoter*, *mitonner*, l'objet en soi dans *sucrerie*, *délice*, *festin*, *etc.* Combinés avec les précédents, ces nouveaux indices apportent donc deux informations supplémentaires sur l'opinion exprimée : sa cible potentielle, et le champ sémantique utilisé par l'auteur pour exprimer son point de vue. Ce travail a déjà fait l'objet d'une première validation (Vernier et al., 2007a; Vernier et Ferrari, 2007).

Parallèlement, nous avons amorcé un travail qui s'intéresse aux jugements d'évaluation portés par des individus ou des institutions, et dénotés par des constituants détachés extérieurs à la prédication principale : *En véritable requin de studio, il…* *Courageusement, les médecins…* Jackiewicz et al. (2009)

Le modèle linguistique élaboré dans le cadre d'une collaboration en cours avec Agata Jackiewicz croise les travaux sur les constituants périphériques (CP) de Combettes (1998) et Charolles (1997) avec ceux de l'Appraisal (Martin et White, 2005). Ce modèle est particulièrement intéressant pour envisager une analyse sémantique discursive du jugement à travers les constituants périphériques qui constituent un cadre d'observation privilégié. En effet, en tant que configuration récurrente particulière, ce type de constituants permet le repérage systématique non seulement d'un jugement d'évaluation mais surtout de différentes propriétés associées, comme la cible du jugement (référent du CP : personne, institution...), la facette de la personnalité évoquée (le focus), le rapport au contexte (évaluation circonstanciée...). Le repérage et l'analyse des CP, dont la mise en œuvre informatique est en cours, permettra des retours sur le modèle et sa validation. À terme, les expérimentations sur un corpus volumineux viseront à évaluer la faisabilité d'une véritable analyse automatique de l'opinion.

D'une manière générale, nous poursuivons l'étude de différents modèles linguistiques afin de mettre en place des analyses symboliques permettant le traitement sémantique du discours évaluatif dans la diversité de ses formes.

Références

Baroni, M. et S. Vegnaduzzo (2004). Identifying subjective adjectives through web-based mutual information. In *Proceedings of KONVENS-04*, pp. 17–24. Vienna, Austria.

Charolles, M. (1997). L'encadrement du discours : univers, champs, domaines et espaces. *Cahier de recherche linguistique 6*, 1–73.

Combettes, B. (1998). *Expressions détachées en français*. Collection L'essentiel Français. Ophrys.

Covington, M. (1994). Gulp 3.1 : An extension of prolog for unification-based grammar. Technical Report Report AI-1994-06, Artificial Intelligence Center, University of Georgia.

Denoyer, L. (2004). *Sequence Labeling with Reinforcement Learning and Ranking Algorithms*. Ph. D. thesis, LIP6 - University of Paris 6.

Esuli, A. et F. Sebastiani (2005). Determining the semantic orientation of terms through gloss analysis. In *Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management*. Bremen, Germany : ACM Press.

Esuli, A. et F. Sebastiani (2006). SentiWordNet : A Publicly Available Lexical Resource for Opinion Mining. In *Proceedings of LREC-06, the 5th Conference on Language Resources and Evaluation*. Genova, Italy.

Ferrari, S., F. Bilhaut, A. Widlöcher, et M. Laignelet (2005). Une plate-forme logicielle et une démarche pour la validation de ressources linguistiques sur corpus : application à l'évaluation de la détection automatique de cadres temporels. In G. Williams (Ed.), *Actes des 4èmes Journées de la Linguistique de Corpus*, Lorient, France. 15-17 septembre 2005 : Université de Bretagne-Sud.

Ferrari, S., Y. Mathet, T. Charnois, et D. Legallois (2008). Analyse d'opinion : discours évaluatif et classification de documents – Retour d'expérience sur deux approches. In M. Roche et P. Poncelet (Eds.), *INFORSID'08 : Informatique des Organisations et Systèmes d'Information et de Décision - Atelier FODOP'08 (FOuille des Données d'OPinions)*, pp. 23–36. Fontainebleau, France : INFORSID.

Halliday, M. (1994). *An Introduction to Functional Grammar* (2 ed.). London : Arnold.

Harb, A., G. Dray, M. Plantié, P. Poncelet, M. Roche, et F. Trousset (2008). Détection d'Opinion : Apprenons les bons Adjectifs ! In M. Roche et P. Poncelet (Eds.), *INFORSID'08 : Informatique des Organisations et Systèmes d'Information et de Décision - Atelier FODOP'08 (FOuille des Données d'OPinions)*, pp. 59–66. Fontainebleau, France : INFORSID.

Hatzivassiloglou, V. et K. McKeown (1997). Predicting the semantic orientation of adjectives. In *Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics*, pp. 174–181. Madrid, Spain : ACL.

Hu, M. et B. Liu (2004). Mining opinion features in customer reviews. In D. L. McGuinness et G. Ferguson (Eds.), *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence AAAI 2004*, pp. 755–760. San Jose, California, USA : AAAI Press / The MIT Press.

Jackiewicz, A., T. Charnois, et S. Ferrari (2009). Jugements d'évaluation et constituants périphériques. In *Actes de TALN'09*. 24-26 juin 2009. Senlis, France. À paraître.

Jindal, N. et B. Liu (2006). Identifying Comparative Sentences in Text Documents. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR-06)*, pp. 244–251. Seattle, Washington, USA : ACM Press.

Joachims, T. (1998). Text categorization with support vector machines : Learning with many relevant features. In C. Nédellec et C. Rouveirol (Eds.), *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pp. 137–142. Chemnitz, Germany : Springer Verlag.

Kreuz, R. et G. Caucci (2007). Lexical influences on the perception of sarcasm. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, pp. 1–4. Rochester, New York : ACL.

Lakoff, G. et M. Johnson (1980). *Metaphors We Live By*. Chicago, U.S.A. : University of Chicago Press.

Legallois, D. et S. Ferrari (2006). Vers une grammaire de l'évaluation des objets culturels. In *Actes d'ISDD06, colloque international Discours et Document, Schedae, 2006, fascicule n^o 1*, pp. 57–68. Caen, 15 et 16 juin 2006 : Presses universitaires de Caen. prépublication n^o 8.

Legallois, D. et C. Poudat (2008). Comment parler des livres que l'on a lus ? Discours et axiologie des avis des internautes. *Semen 26*, 49–80.

Li, W., J. Han, et J. Pei (2001). CMAR : Accurate and efficient classification based on multiple class-association rules. In *IEEE International Conference on Data Mining (ICDM'01), San Jose, USA*.

Liu, B., W. Hsu, et Y. Ma (1998). Integrating classification and association rules mining. In *International Conference on Knowledge Discovery and Data Mining (KDD'98), New York, USA*, pp. 80–86.

Marcoul, F. et A. Athayde (2008). La détection automatique de l'opinion : contraintes et enjeux. In M. Roche et P. Poncelet (Eds.), *INFORSID'08 : Informatique des Organisations et Systèmes d'Information et de Décision - Atelier FODOP'08 (FOuille des Données d'OPinions)*, pp. 1–8. Fontainebleau, France : INFORSID.

Martin, J. R. et P. R. White (2005). *The Language of Evaluation : Appraisal in English*. Palgrave Macmillan Hardcover.

Maurel, S., P. Curtoni, et L. Dini (2007). Classification d'opinions par méthodes symbolique, statistique et hybride. In *Troisième DÉfi de Fouille de Textes (DEFT'07), plate-forme AFIA 2007*, pp. 111–117. Grenoble, France : AFIA, Association Française d'Intelligence Artificielle.

Maurel, S., P. Curtoni, et L. Dini (2008). L'analyse des sentiments dans les forums. In M. Roche et P. Poncelet (Eds.), *INFORSID'08 : Informatique des Organisations et Systèmes d'Information et de Décision - Atelier FODOP'08 (FOuille des Données d'OPinions)*, pp. 9–22. Fontainebleau, France : INFORSID.

Pang, B. et L. Lee (2004). A sentimental education : Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the ACL*, pp. 271–278. Barcelona, Spain : ACL.

Read, J., D. Hope, et J. Carroll (2007). Annotating expressions of appraisal in english. In *Proceedings of the Linguistic Annotation Workshop*, pp. 93–100. Prague, Czech Republic : ACL.

Riloff, E., S. Patwardhan, et J. Wiebe (2006). Feature Subsumption for Opinion Analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP06)*, pp. 440–448. Sydney, Australia : ACL.

Riloff, E. et J. Wiebe (2003). Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*, pp. 105–112. Sapporo, Japan : ACL.

Rioult, F., B. Zanuttini, et B. Crémilleux (2008). Apport de la négation pour la classification supervisée à l'aide d'associations. In F. d'Alché Buc (Ed.), *Actes de la 10e Conférence d'Apprentissage (CAp 2008)*, pp. 183–196. Cépaduès éditions.

Rocchio, J. J. (1971). Relevance feedback in information retrieval. In G. Salton (Ed.), *The*

SMART Retrieval System : Experiments in Automatic Document Processing, pp. 313–323. Englewood Cliffs, NJ, USA : Prentice-Hall.

Salton, G. et C. Buckley (1988). On the Use of Spreading Activation Methods in Automatic Information Retrieval. In *Proceedings of the eleventh Annual International Conference on Research and Development in Information Retrieval*, pp. 147–160. ACM.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing ((NeMLaP))*, pp. 44–49. Manchester, U.K.

Stubbs, M. et I. Barth (2003). Using recurrent phrases as text-type discriminators : a quantitative method and some findings. *Functions of Language 10*(1), 65–108.

Trinh, A.-P. (2007). Classification de texte et estimation probabiliste par Machine à Vecteurs de Support. In *Troisième DÉfi de Fouille de Textes (DEFT'07), plate-forme AFIA 2007*, pp. 69–83. Grenoble, France : AFIA, Assiciation Française d'Intelligence Artificielle.

Trinh, A.-P. (2008). La classification des textes d'opinion par les Séparateurs à Vaste Marge (SVM) avec sorties probabilistes. In M. Roche et P. Poncelet (Eds.), *INFORSID'08 : Informatique des Organisations et Systèmes d'Information et de Décision - Atelier FODOP'08 (FOuille des Données d'OPinions)*, pp. 67–74. Fontainebleau, France : INFORSID.

Turney, P. (2002). Thumbs Up or Thumbs Down ? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting of the ACL (ACL'02)*, pp. 417–424. Philadelphia, Pennsylvania, USA : ACL.

Turney, P. et M. L. Littman (2003). Measuring praise and criticism : Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS) 21*(4), 315–346.

Vernier, M. et S. Ferrari (2007). Tracking evaluation in discourse. In *ASOS07, Workshop on Applications of Semantics, Opinions and Sentiments*. Eurolan summer school, July 23 - August 3, 2007, University of Iaşi, Romania.

Vernier, M., S. Ferrari, et D. Legallois (2007a). Discours évaluatif et suivi d'opinion. In G. Williams (Ed.), *Actes des 5èmes Journées de la Linguistique de Corpus*. 13, 14 et 15 septembre, Lorient, France, 2007 : Université de Bretagne-Sud.

Vernier, M., Y. Mathet, F. Rioult, T. Charnois, S. Ferrari, et D. Legallois (2007b). Classification de textes d'opinions : une approche mixte n-grammes et sémantique. In *Troisième DÉfi de Fouille de Textes (DEFT'07), plate-forme AFIA 2007*, pp. 95–109. Grenoble, France : AFIA, Assiciation Française d'Intelligence Artificielle.

Whitelaw, C., N. Garg, et S. Argamon (2005). Using appraisal taxonomies for sentiment analysis. In *Proceedings of CIKM-05, the ACM SIGIR Conference on Information and Knowledge Management*.

Widlöcher, A. et F. Bilhaut (2005). La plate-forme LinguaStream : un outil d'exploration linguistique sur corpus. In M. Jardino (Ed.), *Actes de la 12e Conférence Traitement Automatique du Langage Naturel (TALN)*, pp. 517–522. Dourdan, France.

Widlöcher, A. et F. Bilhaut (2006). LinguaStream : An Integrated Environment for Computational Linguistics Experimentation. In *Proceedings of EACL 2006, the 11th Conference of the European Chapter of the Association of Computational Linguistics (Companion Vo-*

lume), pp. 95–98. Trento, Italy, April 15-16 2006.

Widlöcher, A. et F. Bilhaut (2008). Articulation de traitements en TAL : Principes méthodologiques et mise en oeuvre dans la plate-forme LinguaStream. *Traitement Automatique des Langues 49*(2), 73–101.

Wiebe, J. et R. Mihalcea (2006). Word sense and subjectivity. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Sidney, Australia : ACL.

Wiebe, J., T. Wilson, et C. Cardie (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation 39*(2-3), 165–210.

Yu, H. et V. Hatzivassiloglou (2003). Towards Answering Opinion Questions : Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)*, pp. 129–136. Sapporo, Japan : ACL.

Summary

This study concerns evaluative discourse. We follow a symbolic approach in order to provide semantic analysis of opinion texts.

A preliminary corpus study led to a linguistic model which shows the complexity of the studied phenomenon. We describe a first implementation designed to provide the linguist expert with an interface for observing regularities on new corpora and getting a feedback on the model.

We then present and discuss two computing approaches for classifying opinion texts which were tested during the DEFT07 challenge. The first one is based on n-grams of words, the second partially implements our linguistic model, enhanced by a classification process using generalized association rules.