Détection fine d'opinion et sentiments : attribution fine de la polarité et calcul incrémental de l'intensité

Claude Martineau, Stavroula Voyatzi, Lidia Varga, Stéphanie Brizard, Aurélie Migeotte

DÉTECTION FINE D'OPINIONS ET SENTIMENTS : ATTRIBUTION DE POLARITÉ ET CALCUL INCRÉMENTAL DE L'INTENSITÉ

Claude MARTINEAU(1), Stavroula VOYATZI(1), Lidia VARGA(1),
Stéphanie BRIZARD(2), Aurélie MIGEOTTE(2)

(1) LIGM, Université Paris-Est – 77454 Marne-la-Vallée Cedex 2
{claude.martineau, stavroula.voyatzi, lidia.varga}@univ-mlv.fr
(2) ARISEM – 1-5 rue Carnot- 91883 Massy cedex
{stephanie.brizard, aurelie.migeotte}@arisem.com

Résumé. Cet article décrit notre contribution sur la détection fine d'opinions dans les blogs et les enquêtes de satisfaction client, et porte plus spécifiquement sur l'étude et la construction du vocabulaire permettant de caractériser une opinion positive ou négative dans les documents. L'approche adoptée ici pour l'analyse et détection d'opinions s'appuie sur la fusion d'un modèle sémantique et d'un modèle numérico-symbolique. Une méthode incrémentale est mise en œuvre permettant de calculer l'intensité des segments évaluatifs en tenant compte de phénomènes linguistiques complexes tels que la négation, la comparaison, la coordination ou l'opposition.

Mots-clés : détection d'opinions et sentiments, segment évaluatif, polarité, intensité.

1. Introduction

Avec l'émergence du Web, et surtout du Web 2.0, le nombre de documents contenant des informations exprimant des opinions, des sentiments ou des jugements d'évaluation devient de plus en plus important. Récemment, les chercheurs de différentes communautés, *i.e.* Fouille de données, Linguistique, Traitement Automatique des Langues, se sont intéressés à l'extraction automatique de ces données d'opinions sur le Web. La détection ou l'extraction automatique d'opinions ou encore d'assertions objectives ou subjectives dans un texte est alors un domaine de recherche en pleine expansion (Wiebe et *al.*, 2005 ; Yang et *al.*, 2007).

Du point de vue des utilisateurs, les deux principales applications de ce type de détection concernent, d'une part, l'analyse automatique d'opinions dans des messages contenant par exemple l'avis de consommateurs sur un produit ou un phénomène particulier (Popescu & Etzioni, 2005), et visent plus particulièrement le développement de tâches de veille (technologique, concurrentielle, sociétale), l'évaluation d'un produit par la communauté avant un achat, la détection de rumeurs (buzz) sur le web ou encore la détection d'opinions émergents et/ou significatives dans les forums. D'autre part, l'analyse de la subjectivité d'une phrase est essentielle notamment pour les systèmes de résumé automatique ou de question/réponse (Riloff & Wiebe, 2003). D'un point de vue scientifique, la problématique posée par la détection d'opinions se situe dans le cadre de la compréhension automatique de messages. Ce problème constitue une possibilité d'aborder un niveau intermédiaire entre la simple détection des entités présentes et l'analyse sémantique complète du message.

Nombreuses sont les questions[1] qui sont liées à la tâche de détection d'opinions et qui sont au cœur des principaux axes de recherche. Dans cet article, nous nous intéressons plus particulièrement à l'étape de construction et structuration du vocabulaire permettant de caractériser une opinion positive ou négative d'un document. L'article est organisé de la manière suivante : la section 2 présente brièvement un état de l'art des principales approches pour la détection d'opinion et de la polarité. La section 3 décrit les expériences réalisées à partir de données réelles issues de blogs et d'enquêtes de satisfaction client. Le calcul incrémental de l'intensité et son implémentation sont décrits dans les sections 4 et 5. La section 6 donne un aperçu global des ressources lexicales développées.

[1]A savoir : (i) la modélisation linguistique et informatique ainsi que la gestion des données d'opinion (qu'est-ce qu'une « opinion », comment la représenter informatiquement ?) ; (ii) l'expression en langue et en discours (comment les opinions, sous leurs différentes facettes, sont-elles formulées ?) ; (iii) la construction, l'acquisition et la validation des ressources linguistiques ; (iv) les méthodes pour identifier, annoter et extraire automatiquement opinions et sentiments dans des documents textuels ou audiovisuels ; etc.

2. La détection d'opinions : état de l'art

Plusieurs travaux se sont intéressés à la détection d'opinions et à la détection de la polarité. La détection d'opinions est une tâche qui permet d'extraire les opinions d'un ensemble de documents pertinents pour un sujet donné. Elle est confrontée à des problèmes qui la distinguent de la recherche traditionnelle thématique dont les sujets sont souvent identifiés par des mots-clés seulement. L'opinion peut être exprimée de manières très variées et subtiles, et donc il est souvent difficile de la déterminer exactement. La classification des sentiments selon la polarité est une sous-tâche de la détection d'opinions. Elle consiste de façon générale à déterminer si l'opinion du document sur le sujet est positive ou négative. De ce fait, plusieurs travaux de recherche se sont intéressées à ce problème, par exemple, (Pang & Lee, 2008) essaient de quantifier le sentiment, (Mishne & de Rijke, 2006a) capturent les niveaux d'humeur dans des notes de blogs, ou encore (Mishne & Glance, 2006) président les ventes de film en fonction des notes des Bloggers.

Afin d'évaluer les résultats des chercheurs dans le domaine, plusieurs campagnes d'évaluations ont vu le jour. Sur le plan international, citons tout d'abord TREC qui signifie « Text Retrieval Conference » et désigne l'ensemble des conférences organisées par le NIST (National Institute of Standard and Technology)[2] sur la recherche d'information. Plusieurs tâches ont fait l'objet de recherches dans ces conférences, dont le Blog Track qui a été introduit en 2006. Chaque année, de nouvelles tâches sont définies dans la détection d'opinions et la détection de la polarité[3]. Signalons encore la campagne d'évaluation internationale SemEval 2007 qui intègre en complément de la tâche d'annotation des textes en fonction de la polarité, une tâche d'annotation des textes à partir d'une liste d'émotions prédéfinies (e.g. peur, colère, joie, surprise, etc.).

Sur le plan francophone, plusieurs sont les ateliers et les campagnes d'évaluation en fouille de données d'opinion qui témoignent d'un intérêt croissant pour leur traitement informatisé. En 2007, le défi DEFT (Défi Fouille de Textes) organisé par le LIMSI a porté sur la classification de textes en français selon le jugement favorable ou défavorable qu'ils expriment. En mai 2008, l'atelier FODOP'08 (Fouille de Données d'Opinions) organisé conjointement à la Conférence INFORSID avait pour objectif de promouvoir des échanges entre chercheurs issus de différentes communautés.

Dans la littérature, il existe généralement deux types d'approches pour la détection d'opinion et de la polarité. Certaines sont basées sur le lexique, d'autres sur l'apprentissage. Le premier type d'approche utilise un lexique de mots qui désignent un sentiment. Ce lexique est soit externe c'est-à-dire construit indépendamment de tout corpus, et dans ce cas, il peut être général (SentiWordNet[4], lexique SUBJ, General Inquiry, Wilson lexicon) ou construit manuellement, soit généré automatiquement à partir du corpus (les mots qui contiennent une opinion sont extraits directement du corpus). À chaque mot du lexique est associé un ensemble de scores d'opinions et de score de polarité. Ce score est traité différemment par les différentes approches pour le calcul du score d'opinion d'un document. La méthode la plus simple est de donner à un document un score égal au nombre total de mots qui contiennent une opinion présents dans le document (e.g. Zhou et *al.*, 2007 ; Fautsch & Savoy, 2008).

Le deuxième type d'approche basée sur l'apprentissage automatique consiste à attribuer des données à un *classifieur* pour l'apprentissage. Ce dernier génère un modèle qui est utilisé pour la partie test de l'apprentissage. Ce type d'approche comprend deux aspects : extraction de *features* et apprentissage du classifieur. Les principales features utilisées sont les suivants : mots seuls, bigrammes, tri-grammes, parties du discours (POS, analyse de l'arbre syntaxique) et polarité. Les principaux classifieurs sont les SVM, Naive Bayes, Maximum Entropy et la régression logistique (Song et *al.*, 2007 ; Mishne & de Rijke, 2006b ; Lee et *al.*, 2008).

Notre expérimentation utilise un modèle de représentation et d'analyse des opinions et sentiments élaboré conjointement avec nos partenaires, THALES, ARISEM et le LIP6[5] qui s'appuie sur la fusion d'un modèle sémantique et d'un modèle numérico-symbolique combinant une expertise

[2] http://www.nist.gov/index.html.
[3] http://trec.nist.gov/.
[4] http://sentiwordnet.isti.cnr.it/.
[5] le LIP6 est le Laboratoire d'Informatique de Paris6, la société ARISEM est une filiale de THALES.

linguistique avec des outils d'intelligence artificielle. Nous présentons notre approche plus en détail en section 3.2.

3. Étude expérimentale

Notre étude s'inscrit dans le cadre du projet de Recherche et Développement DoXa, labellisé par le pôle de compétitivité francilien CAP DIGITAL, et qui concerne le domaine de l'Ingénierie des Connaissances. Le projet vise à mettre en place une plateforme de technologies et méthodologies d'analyse automatique des opinions et sentiments (abrégés en O&S) au sein de grands volumes de textes rédigés en langue française. Le présent travail porte sur la construction et structuration du vocabulaire permettant l'extraction des données d'opinion positives ou négatives.

Dans le cadre de nos recherches, nous prenons l'opinion au sens de jugement de valeur (par opposition au jugement de réalité) sur une entité concrète ou abstraite laquelle peut être un objet, une idée, un projet, un fait, un événement, une situation, ou une personne. Cette entité est le thème sur lequel porte l'opinion. Comme l'indique (Kerbrat-Orecchioni, 1980), « le jugement de valeur peut-être exprimé de manière affective –engagement affectif de l'énonciateur vis-à-vis de l'objet qualifié– ou de manière évaluative ou appréciative –engagement intellectuel de l'énonciateur vis-à-vis de l'objet qualifié. Le jugement peut être exprimé à la fois de manière affective et de manière évaluative ».

3.1. Corpus d'étude et environnements logiciels

Le langage des opinions et sentiments dépend fortement du domaine concerné, ce qui implique que, malgré notre ambition de pouvoir couvrir à l'aide de nos ressources de grands corpus avec des domaines et sous domaines variés, plus nous diversifions le domaine moins les résultats d'extractions seront précis. Pour les besoins du projet, nous avons utilisé deux corpus. D'une part, un corpus portant sur les jeux vidéo, et composé de critiques, de blogs, de reportages sur des salons ou événements ayant traits au domaine des jeux vidéo et touchant parfois celui du cinéma. Ce corpus se présente sous la forme de 7.665 articles et contient 13.601.826 mots. D'autre part, un corpus rassemblant des conversations téléphoniques issues d'une enquête de satisfaction client qui contient 7.256.055 mots. Les textes analysés dans le cadre de nos travaux sont principalement des textes de types *posts* dont la longueur est comprise en moyenne entre 200 et 2000 mots.

Nous utilisons conjointement deux environnements logiciels. D'une part, Unitex 3.0. beta (Paumier, 2008), développé à l'Université Paris-Est, est un environnement logiciel open source multi-plateforme et multilingue. Il permet d'analyser des textes en langue naturelle en utilisant des ressources linguistiques telles que des dictionnaires électroniques, des grammaires locales ou des tables de lexique-grammaire qui sont représentées sous forme d'automates, de transducteurs ou (pour les grammaires locales) de réseaux de transitions récursifs RTN. D'autre part, le moteur d'analyse HST (High Speed Transducer) développé par la compagnie Arisem, utilise des formats semblables à ceux d'Unitex, et gère également des ressources de type ontologique.

3.2. Méthode d'analyse et de détection des opinions et sentiments

La méthode d'analyse et de détection des opinions et sentiments proposé ici s'appuie sur la fusion d'un modèle sémantique et d'un modèle numérico-symbolique. Elle vise à aller au-delà d'une classification binaire permettant de catégoriser les textes selon l'axe de la polarité ou d'une classification quaternaire croisant l'axe de la polarité et l'axe de l'intensité. Elle vise également à mettre en œuvre pour un texte donné, une analyse locale des opinions ou sentiments exprimés au niveau phrastique, et une analyse globale des opinions ou sentiments exprimés au niveau des portions de texte et du texte entier. L'objectif est de permettre la mise en œuvre de parcours d'analyse allant d'une vision macro et quantitative à une vision micro et qualitative.

En schématisant, la représentation des O&S du modèle DoXa s'articule sur trois niveaux :

i. MICRO : l'analyse est faite au niveau de la phrase ou portion de phrase.
ii. MESO : l'analyse concerne le paragraphe ou la portion de texte.
iii. MACRO : l'analyse porte sur l'ensemble du texte.

Au niveau MICRO, l'analyse est effectuée grâce à une approche symbolique qui, malgré un coût parfois élevé, permet d'annoter le plus finement possible des segments de texte sensiblement

longs (≤7 mots), appelés *segments évaluatifs*, et de leur attribuer des traits tels que la polarité et l'intensité. Cette annotation s'appuie sur un ensemble de catégories sémantiques d'O&S que nous décrivons en détail en section 3.3.

L'application du modèle numérico-symbolique permet de synthétiser l'ensemble des annotations posées au niveau MICRO afin de caractériser premièrement le contenu évaluatif de chaque paragraphe (niveau MESO) et, ensuite, dans un second temps, celui du texte dans son intégralité (niveau MACRO). Elle permet également, notamment lorsque les informations sont ambigües, imprécises, contrastées voire contradictoires, de prendre des décisions sur des annotations isolées, par exemple « je suis ni content ni mécontent » ou « je suis à la fois en colère et déçu ». La composante numérico-symbolique est fondée sur des opérateurs et des heuristiques d'agrégation issus de l'apprentissage automatique et la théorie des ensembles flous. Nos travaux de recherche sont consacrés à l'analyse et annotation fine des O&S au niveau MICRO[6].

3.3. Modèle de représentation sémantique des opinions et sentiments

Les annotations produites reposent sur le modèle O&S du projet DoXa, qui est inspiré des travaux de (Mathieu, 2006) sur la classification des verbes de sentiment, et la théorie de l'évaluation (Martin et *al.*, 2005). Un premier jeu de catégories sémantiques a été soumis à des annotateurs humains pour évaluation sur un corpus de *posts* issus de blogs portant sur les jeux vidéos. Les retours des annotateurs ont permis de simplifier le modèle, en réduisant le nombre de catégories initialement définies sur la base de regroupements des catégories. Le tableau 1 présente les vingt catégories sémantiques retenues, munies de leur polarité intrinsèque, de leur étiquette en anglais (utilisée dans les ressources avec le préfixe *cat_*), de celle de la catégorie antonyme si elle existe et, enfin, illustrées d'un exemple.

Ces catégories sémantiques s'appliquent à tout type de catégorie grammaticale appelées ici *constituants de base* : adjectif, nom, verbe, adverbe et expression (semi-)figée. La présence d'une négation dans la phrase peut donner lieu à une inversion de polarité qui se traduit dans l'annotation du segment évaluatif traité, soit par un passage à la catégorie antonyme (cf. *Etiquette Cat. Antonyme*) soit par l'ajout de l'attribut *neg*. En voici quelques exemples :

(1) intéressant, cat_Satisfaction|int3 pas intéressant, cat_Dissatisfaction|int3
(2) inquiet, cat_Fear|int3 pas inquiet, cat_Fear|int3|neg

Catégorie Sémantique	Pol. Intrinsèque	Etiquette	Etiq. Cat. Ant.	Exemple
Accord	positive	Agreement	Disagreement	approbation
Colère	négative	Anger		exaspération
Apaisement	positive	Appeasement		rassurée
Valorisation	positive	Appraisal	Depreciation	bienveillant
Ennui	négative	Boredom		rébarbatif
Mépris	négative	Contempt		<prendre> en grippe
Dévalorisation	négative	Depreciation	Appraisal	dénigrer
Mésentente	négative	Disagreement	Agreement	<mettre> en doute
Gêne	négative	Discomfort		perturber
Déplaisir	négative	Displeasure		répugnant
Insatisfaction	négative	Dissatisfaction	Satisfaction	incompétent
Crainte	négative	Fear		effroi
Surprise Négative	négative	NegSurprise	PosSurprise	sidéré
Plaisir	positive	Pleasure		divertir
Surprise Positive	positive	PosSurprise	NegSurprise	<couper> le souffle
Tristesse	négative	Sadness		découragement
Satisfaction	positive	Satisfaction	Dissatisfaction	adorable
Connotation méliorative	positive	MelConnot		bravo
Connotation péjorative	négative	PejConnot		problématique
Attente	neutre	Expectation		souhaiterais

Tableau 1. *Catégories sémantiques des opinions et sentiments*

[6]Au sein du projet DoXa, la tâche d'agrégation d'annotation est confiée au LIP6, quant à celle d'annotation MICRO, elle est le fruit de la collaboration du LIGM et de la société ARISEM.

4. Annotation des segments évaluatifs et calcul de l'intensité

L'annotation d'un segment évaluatif indique son appartenance à une ou plusieurs catégories sémantiques (tableau 1), chacune munie d'une valeur d'intensité prise sur une échelle en comportant dix (1-10). Cette intensité résulte de la prise en compte de l'intensité intrinsèque[7] associée à tout constituant de base prenant ses valeurs entre 3 et 7, et éventuellement d'un ou plusieurs modifieurs spécifiques qui possèdent trois niveaux en intensification comme en atténuation. Les valeurs inférieures (<3) et supérieures (>7) sont respectivement atteintes par l'application de ces modifieurs. L'exemple suivant présente un adjectif isolé, puis combiné avec deux modifieurs différents :

(3) intéressant, cat_Satisfaction|int3
très, AdvInt2 — très intéressant, cat_Satisfaction|int5
peu, AdvAtt1 — peu intéressant, cat_Satisfaction|int2

La modification d'intensité peut également être produite par la présence de préfixes (e.g. *ultra* intéressant, *mega* jeu), de superlatifs (e.g. le jeu *le plus* marrant *du monde*), ou encore de modifieurs adverbiaux divers (e.g. *très, extrêmement, à peu près* satisfait). Nous avons divisé ces derniers en huit classes[8]. Pour expliquer le processus du calcul de l'intensité, nous avons construit une phrase d'exemple qui intègre l'ensemble des niveaux de modification d'intensité traités par les ressources développés :

(4) Ce jeu est unanimement vraiment le plus hyper intéressant qu'on connaisse

L'adjectif *intéressant* est précédé de plusieurs mots qui contribuent chacun à leur tour à la modification de son intensité de base (intensité intrinsèque = 3). L'intensité de base pouvant prendre dix valeurs, celle des modifieurs pouvant en prendre trois en intensification comme en atténuation, les combinaisons s'avèrent fort nombreuses. Cette explosion combinatoire rend quasiment impossible le calcul de l'intensité résultante par un simple transducteur. C'est pourquoi nous avons dû opter pour une approche incrémentale qui calcule l'intensité résultante[9] de proche en proche. Cette méthode est explicitée par la ligne ci-dessous dans laquelle les crochets symbolisent l'intensité intrinsèque ou la modification d'intensité apportée par un constituant de base ; et les parenthèses, la manière dont ces intensités sont deux à deux combinées :

(5) Intensité résultante = (([*unanimement*] [*vraiment*]) ([*le plus*] ([*hyper*] [*intéressant*])))
10 = +1 +2 +2 +3 3

4.1. Annotation des segments évaluatifs consécutifs

Les ressources de chaque catégorie sémantique représentées sous forme de graphe dictionnaire et les données indiquant l'ordre dans lequel les appliquer constituent un module. L'analyse des segments évaluatifs consiste d'abord à traiter le texte par un module appelé *transverse* qui reconnaît les modifieurs de toutes sortes ainsi que les négations. Ensuite, les modules des catégories sémantiques (cf. tableau 1, section 3.3) sont successivement appliqués au texte afin de reconnaître chacun les données lexicales qui leur sont propres. En les combinant avec les négations et modifieurs précédemment identifiés, on produit les annotations des segments complexes. Un ultime traitement, s'appuyant sur la présence des connecteurs, permet de repérer parmi les segments reconnus ceux qui seraient diversement reliés entre eux : comparatifs (supériorité, égalité, infériorité), conjonctifs (coordination, disjonction, énumération, opposition). En voici des exemples extraits de nos corpus d'étude :

---

[7]Nous situant dans une perspective de TAL et de linguistique de corpus, nous avons adopté une démarche empirique et itérative pour l'attribution des valeurs d'intensité intrinsèque aux constituants de base. Faute de données appropriées pour le français, nous avons fait appel à des linguistes de l'équipe du LIGM qui ont attribué des intensités sur un certain nombre représentatif des unités lexicales (constituants de base). Puis, les retours des annotateurs ont permis, d'une part, de résoudre les conflits de valeurs attribuées et, d'autre part, de définir une échelle opérationnelle pour le calcul de l'intensité.

[8]Bien qu'ils ne soient pas tous des quantifieurs *stricto sensu*, nous avons tenté de traduire au niveau de l'intensité (seule variable de notre modèle) les variations aspectuelles ou modales qu'ils peuvent apporter, et qui incluent des notions comme par exemple, la **source de l'information émise** ou le **positionnement du locuteur vis-à-vis de son énoncé**.

[9]Toute valeur de l'intensité résultante qui dépasse l'intensité maximale de 10 est remplacée par 10.

(6) *plus* de frustration *que* de plaisir,.ComparSup+Annotation1=cat_Dissatisfaction|int5+ DissatisfactionNoun;Annotation2=cat_Pleasure|int3}+PleasureNoun
(7) charmante *mais* pas forcément compétente,.Opposition+Annotation1=cat_Satisfaction|int4+ SatisfactionAdj;Annotation2=cat_Depreciation|int3+DepreciationMais+MaisComp

5. Implémentation et importation sous Unitex

En termes d'implémentation, sous HST, cette approche s'exprime à l'aide d'un format de ressources intermédiaire entre dictionnaire et grammaire que nous appellerons *dictionnaires de motifs*. Ils sont composés de lignes dont la partie gauche est semblable à une expression régulière simplifiée et la partie droite à une entrée de dictionnaire Dela. Chaque ligne est comparable à une grammaire à plat représentable par un graphe ne comportant qu'un seul chemin comme par exemple : <*faire*> *d'une pierre deux coups* −> Expression+Verbe.

Les dictionnaires de motifs sont utilisés pour représenter, d'une part, des constituants de base avec leur intensité intrinsèque ou la modification d'intensité qu'ils opèrent :

(8) <*avoir*> *le bourdon* −> cat_Sadness|int4+SadnessSemiFrozen

D'autre part, ils représentent des règles de modification de l'intensité comme :

(9) {AdvInt2} {cat_Displeasure|int1} −> cat_Displeasure|int3+DispleasureComp

En appliquant, dans l'ordre adéquat, de tels dictionnaires, on peut reconnaître chaque composant d'un segment évaluatif (simple ou complexe), et calculer de manière incrémentale son intensité. Lors de l'analyse d'une phrase, HST utilise, d'une part, des ressources de type ontologique pour capter les thèmes sur lesquels portent les opinions exprimées dans les segments évaluatifs ; et, d'autre part, des ressources représentées par des grammaires locales ou des dictionnaires pour traiter les segments évaluatifs.

Afin de profiter des possibilités des deux environnements HST et Unitex, et d'améliorer ainsi potentiellement les ressources produites, nous avons développé un programme qui permet d'importer dans l'environnement Unitex des données issues de HST. Chaque dictionnaire de motif est importé sous la forme d'un *graphe dictionnaire* qui s'applique comme un dictionnaire Dela et construit dynamiquement des entrées dans le dictionnaire du texte. À titre d'exemple, considérons le mini dictionnaire ci-dessous qui comprend divers types d'entrées[10] impliquées dans le traitement de notre exemple :

hyper −> PrefInt3+ModInt3+PrefModifier
<intéressant> −> cat_Satisfaction|int3+SatisfactionAdj+SatisfactionAdjInt3
{PrefInt3}={SatisfactionAdjInt3} −> cat_Satisfaction|int6+SatisfactionAdjInt6+SatisfactionPref

Voici (cf. figure 1) le graphe dictionnaire équivalent généré par le programme d'importation :

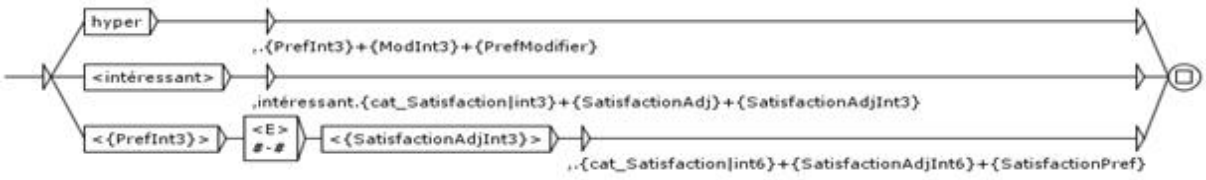

Figure 1. *Graphe dictionnaire issu d'un dictionnaire de motif*

L'application d'un ensemble de graphes dictionnaires à notre exemple de référence permet de visualiser sous Unitex (cf. figure 2), dans le dictionnaire du texte, les analyses et intensités partielles consécutivement produites, i.e. *hyper intéressant*, *le plus hyper intéressant*, ainsi que le segment évaluatif intégralement reconnu avec l'intensité correcte.

[10] *Hyper* est un préfixe intensifieur entrainant une incrémentation d'intensité *+3*, <*intéressant* > permet de reconnaitre les formes fléchies de cet adjectif auxquelles une intensité intrinsèque de *3* est attribuée. La dernière ligne est une règle qui calcule l'intensité résultante d'un préfixe intensifieur d'intensité *+3* appliqué à un adjectif de catégorie *Satisfaction* d'intensité *3*. Le signe « = » permet d'accepter les formes avec ou sans trait d'union.

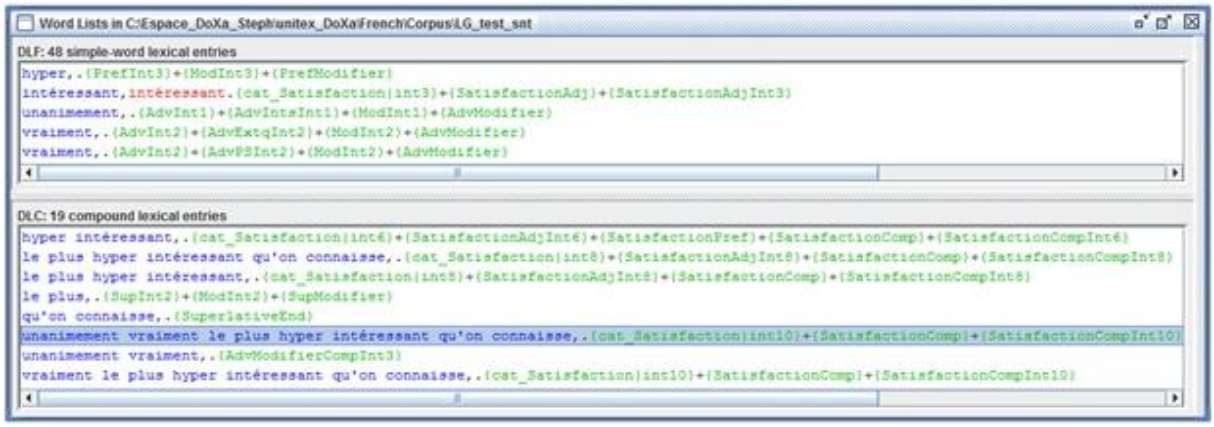

Figure 2. *Dictionnaire du texte : segments évaluatifs reconnus*

6. Dictionnaires d'opinions et sentiments

Les dictionnaires contiennent, à ce jour, 6.703 entrées de type lexical et 23.188 entrées de type grammatical (règles de calcul d'intensité résultante et de négation). Les tableaux 2 et 3 donnent, pour les vingt catégories sémantiques O&S (cf. tableau 1, section 3.3), le nombre d'entrées lexicales respectivement par catégorie sémantique et par catégorie syntaxique :

Catégorie sémantique	Entrées	Catégorie sémantique	Entrées
Agreement	189	Dissatisfaction	169
Anger	283	Expectation	565
Appeasement	107	Fear	195
Appraisal	485	MelConnot	83
Boredom	61	NegSurprise	141
Contempt	245	PejConnot	264
Depreciation	653	Pleasure	339
Disagreement	223	PosSurprise	96
Discomfort	92	Sadness	288
Displeasure	126	Satisfaction	202

Catégorie syntaxique	Entrées
Adjectifs	2279
Adverbes	169
Noms	826
Verbes	832
Expressions Figées	261
Expressions Semi-Figées	558
Adjectifs Modifieurs	51
Adverbes Modifieurs	535

Tableaux 2 et 3. *Catégories sémantiques et syntaxiques : nombre d'entrées*

7. Conclusion et perspectives

Dans cet article, nous avons décrit notre contribution sur la détection d'opinions et de la polarité dans les blogs et les enquêtes de satisfaction client, qui porte plus spécifiquement sur le développement des ressources linguistiques permettant de caractériser une opinion positive ou négative dans les documents. Ces ressources ont été développées selon le modèle des opinions et sentiments (O&S) du projet DoXa. Nous avons proposé une méthode incrémentale permettant de calculer l'intensité des segments de texte en tenant compte de phénomènes linguistiques complexes tels que la négation, la comparaison, la coordination ou l'opposition. Dans la phase suivante du projet, nous envisageons une évaluation des ressources produites afin de pouvoir, d'une part, procéder à des levées d'ambigüité et, d'autre part, compléter et raffiner les dictionnaires et grammaires existants. Une évaluation globale de la tâche[11] de détection d'opinions et sentiments est également envisagée à la fin du projet[12].

Remerciements

Ce travail a été financé conjointement par la Direction Générale de la Compétitivité, de l'Industrie et des Services (DGCIS) et le Fonds unique interministériel dans le cadre du projet de Recherche et Développement collaboratif, DoXa, labellisé par le pôle de compétitivité CAP DIGITAL.

[11]Les développements déjà réalisés par l'ensemble des partenaires du projet DoXa fait l'objet d'un chapitre du livre *Next Generation Search Engines: Advanced Models for Information Retrieval* à paraitre début 2012.
[12]A cette période, une version publique des ressources développées sera mise à la disposition de la communauté.

Bibliographie

FAUTSCH, C. & SAVOY, J., UniNE at TREC 2008: Fact and Opinion Retrieval in the Blogsphere, In *Proceedings of the 17th Text REtrieval Conference (TREC 2008)*, 2008.

KERBRAT-ORECCHIONI, C., *L'énonciation. De la subjectivité dans le langage*. Paris, Armand Colin, 1980.

LEE, Y., Na, S.-H., KIM, J., NAM, S.-H., JUNG, H.-Y. & LEE, J.-H., « KLE at TREC 2008 Blog Track: Blog Post and Feed Retrieval », In *Proceedings of the 17th Text REtrieval Conference (TREC 2008)*, 2008.

MARTIN, J. R. & WHITE, P. R. R., *The Language of Evaluation: Appraisal in English*, London & New York, Palgrave MacMillan, 2005.

MATHIEU, Y. Y., « A Computational Lexicon of French Verbs of Emotion », *Computing Attitude and Affect in Text: Theory and Applications*, Springer Dordrecht, The Netherlands, 2006, p. 109–123.

MISHNE, G. & de RIJKE, M., « Capturing global mood levels using blog posts », In *Proceedings of the AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, Stanford, California, USA, 2006a, p. 145–152.

MISHNE, G. & de RIJKE, M., « A study of blog search », In *Proceedings of the 28th European Conference on Information Retrieval (ECIR 2006)*, vol. 3936, London, UK, 2006b, p. 289–301.

MISHNE, G. & GLANCE, N., « Predicting movie sales from blogger sentiment », In *Proceedings of the AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, Stanford, California, USA, 2006, p. 155–158.

PANG, B. & LEE L., « Opinion Mining and Sentiment Analysis », *Foundations and Trends in Information Retrieval*, vol. 2 (1-2), 2008, p. 1–135.

PAUMIER, Sébastien, Unitex 2.0 user Manual, http://www-igm.univ-mlv.fr/~paumier/recherche.php, 2008.

POPESCU, A.-M. & ETZIONI, O., « Extracting product features and opinions from reviews », In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT/EMNLP'05)*, Vancouver, B.C., Canada, 2005, p. 339–346.

RILOFF, E. & WIEBE, J., 2003. « Learning extraction patterns for subjective expressions », In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03)*, Sapporo, Japan, 2003, p. 105–112.

SONG, R., TANG, Q., SHI, D., LIN, H. & Yang, Z., « DUTIR at TREC 2007 Blog Track », In *Proceedings of the 16th Text REtrieval Conference (TREC 2007)*, 2007.

WIEBE, J., WILSON, T, & CARDIE, C., « Annotating expressions of opinions and emotions in language », *Language Resources and Evaluation*, vol. 39 (2-3), 2005, p. 165–210.

YANG, K., Yu, N. & ZHANG, H., « WIDIT in TREC 2007 Blog Track: Combining Lexicon-Based Methods to Detect Opinionated Blogs », In *Proceedings of the 16th Text REtrieval Conference (TREC 2007)*, 2007.

ZHOU, G., Joshi, H. & BAYRAK, C., « Topic categorization for relevancy and opinion detection », In *Proceedings of the 16th Text REtrieval Conference (TREC 2007)*, 2007.

Claude MARTINEAU
Stavroula VOYATZI
Lidia VARGA
Université Paris-Est
77454 Marne-la-Vallée Cedex 2

Stéphanie BRIZARD
Aurélie MIGEOTTE
Arisem – 1-5 rue Carnot
91883 Massy cedex