L'analyse des sentiments dans les forums

Sigrid Maurel, Paolo Curtoni et Luca Dini

CELI France, SAS
12-14, rue Claude Genin
38000 Grenoble
http://www.celi-france.com
{maurel, curtoni, dini}@celi-france.com

RÉSUMÉ. *Nous présentons ici trois méthodes différentes pour effectuer une classification automatique de textes d'opinion. La première méthode est symbolique, la seconde statistique et la dernière, hybride, est une combinaison des deux premières. Nous montrons comment la combinaison des méthodes symbolique et statistique permet de tirer parti des avantages des deux méthodes, à savoir la robustesse de l'apprentissage automatique statistique et la possibilité de configuration manuelle offerte par la méthode symbolique, permettant une utilisation dans des applications réelles. Les textes classés par ces méthodes viennent de sources informationnelles non structurées de type forum sur Internet.*

ABSTRACT. *We present three different methods to perform an automatic classification of texts which include opinions. The first method is symbolic, the second statistic and the last, hybrid, is a combination of the first two. We will show how the combination makes it possible to exploit the advantages of both methods, namely robustness of statistical machine learning and the possibility of a manual configuration given by the symbolic method allowing the use of real-life applications. The classified texts by these methods come from non structured information sources such as internet forums.*

MOTS-CLÉS : *méthodes symbolique, statistique et hybride, classification d'opinions de forums d'Internet, analyse au niveau de phrase.*

KEYWORDS: *symbolic, statistic and hybrid methods, classification of opinions of internet forums, analysis on sentence level.*

Atelier FODOP'08, pages 9 à 22

1. Introduction

Cet article[1] s'intéresse à la classification de textes d'opinion en langue française. Dans ce cas précis, la classification a pour objectif l'analyse de sentiments exprimés dans différents types de textes comme par exemple dans des forums de discussion sur Internet où les internautes échangent des avis et s'entraident. Les textes issus de forums sur Internet constituent des sources d'informations spontanées et récentes, incontournables pour acquérir, au jour le jour, des connaissances sur les consommateurs, pour anticiper leurs besoins et leurs attentes afin de tenter d'améliorer la relation client/fournisseur. En analysant ces textes d'opinion le fournisseur d'un produit ou d'un service peut mieux réagir aux désidérata de ses clients, le client peut de son côté s'inspirer des sentiments et opinions d'autres clients sur le produit auquel il s'intéresse et profiter ainsi d'une aide à la décision (acquérir ou pas le produit, choisir plutôt le produit A ou le produit B, etc.). Comme le montrent de nombreux travaux de socio- et psycho-linguistique (c.f. (Sproull *et al.,* 1991)), la communication médiée par ordinateur favorise l'expression des émotions, sentiments et opinions souvent contrôlés ou réprimés dans des cadres de communication plus traditionnels visant à étudier le point de vue des consommateurs (interviews face à face, enquêtes fermées, enquêtes ouvertes, etc.). De là, naît l'intérêt des analystes pour ces sources d'informations.

Une des difficultés de la classification en « positif et négatif » réside dans la nécessité d'une bonne analyse syntaxique du texte, analyse qui peut se révéler particulièrement difficile dans des cas de coordination entre plusieurs parties d'une phrase, d'anaphore ou de coréférence (la reprise d'un argument présent plus loin dans le document). Une autre difficulté du langage naturel pour l'analyse automatique de sentiments sont les contextes intentionnels, pour lesquels l'expression d'opinion n'est pas un vrai sentiment. C'est le cas dans une phrase comme :

« Je croyais que la France était un beau pays. »

(Dini *et al.,* 2002) ont montré le lien qui existe entre les structures syntaxiques et sémantiques d'une phrase et l'expression de l'opinion qu'elle véhicule. Ainsi l'analyse de la phrase par « paquets de mots » donne des résultats peu satisfaisants alors qu'une analyse syntaxique du texte peut aider à trouver les expressions qui contiennent des opinions. Les deux phrases suivantes contiennent les mêmes paquets de mots sans pour autant exprimer les mêmes sentiments. En effet, la première phrase contient un sentiment positif alors que la deuxième est négative :

« Je l'ai apprécié pas seulement à cause de ... »
« Je l'ai pas apprécié seulement à cause de ... »

Dans le cadre de sa participation à la campagne d'évaluation DEFT'07 (*cf.* section 5) CELI France a mis au point trois méthodes pour classer les textes des différents corpus. La première est une méthode symbolique qui inclut un système d'extraction d'information adapté aux corpus. Elle est basée sur des règles d'un analyseur

1. Le présent texte est une version résumée de l'article *Sybille, anatomie d'un système automatique d'extraction de sentiments* des mêmes auteurs, à paraître bientôt aux Éditions de l'Université Stendhal de Grenoble, c.f. (Maurel *et al.,* 2008, à paraître).

syntaxico-sémantique. Cet analyseur contient un lexique de mots qui véhiculent des sentiments sur lesquels réagissent les règles de la grammaire. La deuxième est une méthode statistique basée sur des techniques d'apprentissage automatique. Enfin, la dernière, SYBILLE, est une méthode hybride qui combine les techniques des deux précédentes pour aboutir à des résultats très précis. L'analyse des textes se fait au niveau de la phrase, les sentiments d'un document sont extraits phrase par phrase, et c'est seulement ensuite qu'une valeur globale est attribuée au message entier. Ceci permet d'extraire une information contextuelle qui est donc très précise.

Les sections suivantes présentent brièvement l'état de l'art et les corpus utilisés pour s'attarder ensuite sur les trois méthodes développées et en fournir une première évaluation.

1.1. *État de l'art*

Aujourd'hui l'analyse de sentiments se concentre sur l'attribution d'une polarité à des expressions subjectives (les mots et les phrases qui expriment des opinions, des émotions, des sentiments, etc.) afin de décider de l'orientation d'un document (c.f. (Turney, 2002), (Wilson *et al.,* 2004)) ou de la polarité positive/négative/neutre d'une opinion dans un document (c.f. (Hatzivassiloglou *et al.,* 1997), (Yu *et al.,* 2003), (Kim *et al.,* 2004)). Des travaux allant au-delà ont mis l'accent sur la force d'une opinion d'expression où chaque proposition dans une phrase peut avoir un fond neutre, faible, moyen ou élevé (c.f. (Wilson *et al.,* 2004)). Des catégories grammaticales ont été utilisées pour l'analyse de sentiments dans (Bethard *et al.,* 2004) où des syntagmes adjectivaux comme *trop riche* ont été utilisés afin d'extraire des opinions véhiculant des sentiments. (Bethard *et al.,* 2004) utilisent une évaluation basée sur la somme de scores des adjectifs et des adverbes classés manuellement, tandis que (Chklovski, 2006) utilise des méthodes fondées sur un modèle pour représenter des expressions adverbiales de degré telles que *parfois*, *beaucoup*, *assez* ou *très fort*.

L'approche que nous avons adoptée pour la classification de textes d'opinion est caractérisée par une utilisation mixte d'une technologie symbolique fondée sur des règles et d'une technologie statistique reposant sur l'apprentissage automatique, approche dans laquelle la méthode symbolique a un poids plus important (c.f. (Dini, 2002), (Dini *et al.,* 2002), (Maurel *et al.,* 2007)). La technologie symbolique fait d'abord une analyse du texte phrase par phrase et en extrait ensuite les relations qui véhiculent des sentiments, tandis que la technologie statistique traite les textes en une seule phase et attribue un sentiment global au texte entier à la fin du traitement. Il convient de remarquer que, contrairement à d'autres approches actuelles, la technologie de l'analyse de sentiments développée à CELI France (SYBILLE) ne se limite pas à une analyse lexicale (c'est-à-dire identification et pondération de mots positifs et négatifs), mais s'étend à une analyse syntaxique et sémantique. L'analyse syntaxique est effectuée par le biais d'une analyse robuste de surface telle que celles décrites par (Aït-Mokhtar *et al.,* 1997), (Basili *et al.,* 1999), (Aït-Mokhtar *et al.,* 2001), donnant ainsi un résultat très proche de celui produit par des grammaires de dépendance.

1.2. *Les corpus*

Les données de type forums de discussion sur Internet s'articulent comme un flux d'interactions, comme par exemple : demande-réponse, argument-contre argument, commentaire-désaccord, etc. Ce flux est distribué sur une dimension temporelle qui nécessite un traitement chronologique du fil de discussion. Contrairement aux corpus utilisés par (Wilson *et al.,* 2004) il n'est pas nécessaire ici d'identifier la personne à qui est associé un sentiment, car dans 95 % des cas les discours analysés sont des discours à la première personne.

Les corpus utilisés sont assez différents les uns des autres, que ce soit par la taille des corpus eux-mêmes que par la taille de chaque *thread* (fil de discussion). Nous avons utilisé les corpus de DEFT'07 auxquels nous avons ajoutés des corpus collectés sur Internet. Certains corpus sont structurés, d'autres contiennent beaucoup de messages en style *texto*, et le nombre de fautes d'orthographe présentes dans les messages varie aussi beaucoup.

Le comité d'organisation de DEFT'07 a pris le soin de nettoyer ses corpus (c.f. (Grouin *et al.,* 2007)). Ainsi, les fins de ligne ont été normalisées, les caractères encodés en ISO-Latin, et les textes ont été annotés manuellement.[2] Les corpus de DEFT'07 contiennent des critiques de films, de livres et de spectacles, des tests de jeux vidéo, des relectures d'articles scientifiques (de différentes conférences sur l'intelligence artificielle) et des notes de débats parlementaires (sur la loi de l'énergie). Les textes de nos corpus portent sur le tourisme (en France et ailleurs dans le monde), les jeux vidéo (critiques et problèmes) et les imprimantes (conseils d'achats). Ils comprennent d'un côté des aides à la solution de problèmes, mais aussi des avis sur des produits achetés et des lieux visités.

2. Méthode symbolique

Comme nous l'avons dit plus haut, la méthode symbolique se base sur une analyse syntaxique du texte faite par un analyseur fonctionnel et relationnel (c.f. les travaux sur l'analyse syntaxique et sémantique de (Basili *et al.,* 1999), (Aït-Mokhtar *et al.,* 2001), (Dini, 2002), (Dini *et al.,* 2002), (Dini *et al.,* 2007)). Cet analyseur traite, phrase par phrase, un texte donné en entrée et en extrait, pour chaque phrase, les relations syntaxiques présentes. Il s'agit de relations syntaxiques fonctionnelles de base, telles que le modifieur d'un nom, d'un verbe, sujet et objet d'une phrase, ainsi que de relations plus complexes telles que la coréférence entre deux syntagmes au sein d'une même phrase.

L'utilisateur a la possibilité d'élaborer une grammaire à sa guise et d'ajouter de nouvelles règles afin d'extraire les relations auxquelles il s'intéresse. Pour ce faire il peut modifier les règles d'extraction de relations (par exemple ajouter des règles pour

2. Pour ce qui concerne nos propres corpus ils sont encodés en UTF-8 et nous n'avons effectué aucun nettoyage. Tous les corpus sont disponibles au format XML.

de nouvelles relations), augmenter/diminuer les traits sur les mots dans le lexique qui agissent sur les règles, enlever certaines parties du traitement, etc. Un algorithme permet de calculer un indice de confiance qui servira à la méthode hybride (*cf.* section 4) pour déterminer le résultat final. Cet algorithme prend en compte la valeur (`NMOD`, `OBJ`, `COORD`, etc.) du trait `confianceRegle` qui est attribuée à la relation par la règle du traitement qui s'applique à la phrase. La polarité positive ou négative attribuée au message entier[3] dépend du rapport entre la quantité de relations d'opinions positives et négatives.

2.1. *Grammaire*

La grammaire utilisée a été initialement développée afin d'extraire les relations de sentiments exprimés dans une phrase dans le cadre d'un projet sur le tourisme en France. Elle a été ensuite modifiée et améliorée en vue de la participation à DEFT'07 ((*cf.* section 5), (Maurel *et al.,* 2007)). Dans un deuxième temps, la grammaire a été divisée en deux parties : une première partie de base (la grammaire « générique ») s'appliquant à tous les textes qui contiennent des sentiments, et une deuxième partie pour chaque domaine différent, selon le sujet du corpus : tourisme, jeux vidéo, imprimantes, etc. Les différences se situent essentiellement dans les lexiques appliqués, chaque domaine ayant ses propres mots et expressions. Ainsi les mots se rattachant à la vitesse (*lent*, *rapide*, etc.) ont des polarités différentes selon qu'ils qualifient une imprimante ou un voyage. De même, comme le montrent les phrases ci-dessous, l'adjectif *effrayant* est plutôt perçu comme positif dans une description romanesque alors qu'il est perçu comme négatif dans le domaine des assurances ou du tourisme :

« Dans *Ghost*, les habitants du village sont vraiment effrayants ! »
« C'est effrayant de voir comment la côte est de plus en plus bétonnée. »

En général, une relation de sentiment a deux arguments : le premier est l'expression linguistique qui véhicule le sentiment en question, le deuxième est la cause ou l'objet du sentiment (si la cause est exprimée dans la phrase). Ceci donne pour la phrase

« J'aime beaucoup Grenoble. »

la relation `SENTIMENT_POSITIF(aimer,Grenoble)`. L'attribut `POSITIF` de la relation, c'est-à-dire la valeur de sa classe, indique qu'il s'agit d'un sentiment positif dont la cause est *Grenoble*. Dans le cas d'une phrase comme

« Je déteste ! ! ! ! ! »

la relation n'aura qu'un seul argument : `SENTIMENT_NEGATIF(détester)`, dans la mesure où l'objet du sentiment n'est pas exprimé dans la phrase. L'objectif de la grammaire est d'extraire un maximum d'informations du *thread*, en particulier les

3. L'attribution d'un sentiment global au message entier est utilisée dans des contextes spécifiques, comme par exemple pour l'évaluation DEFT'07 (*cf.* section 5). Sinon nous n'attribuons pas de sentiment global mais gardons les sentiments attribuées à chaque phrase.

sentiments positifs et négatifs, les lieux et produits. Pour ceci les *threads* sont analysés phrase par phrase. Chaque phrase peut contenir zéro, une ou plusieurs relations de sentiment. Il est tout à fait possible d'avoir des relations de sentiments positifs et négatifs dans une même phrase :

> « En qualité d'impression, la Epson est meilleure, en texte comme en photo, malheureusement c'est aussi la plus chère. »

Les parties de la grammaire qui varient selon le corpus se distinguent essentiellement par le lexique de mots qui reçoivent les traits `positif` et `négatif` correspondant aux valeurs des classes des textes. Par exemple, le lexique de la grammaire du *tourisme* contient les mots *joli* et *beau* :

> « Ce monument est vraiment *beau*. »

Pourtant, dans un corpus qui porte sur le cinéma, les livres ou les jeux vidéo, ces mêmes mots n'expriment pas des sentiments. Ils ont donc été supprimés du lexique de la grammaire des *jeuxvidéo* parce qu'ils produisent trop de relations éronnées :

> « Cela dépendra moins de vous que de l'imbécillité contagieuse des ennemis qui attendent sagement derrière un petit muret, leur *beau* visage buriné dépassant allègrement. »

Comme on le voit dans la phrase précédente, dans ce contexte les mots de type *joli* ou *beau* sont utilisés pour décrire une action ou un personnage, mais pas un sentiment. La difficulté réside dans le fait de pouvoir distinguer les parties subjectives des parties objectives d'un texte. La description d'une action peut contenir des phrases avec des sentiments, donc subjectives, qui se réfèrent au déroulement de l'histoire. Cependant ces phrases devront être considérées comme étant objectives pour l'évaluation.

2.2. *Lexique de sentiments*

L'analyse du texte se base sur les mots du lexique qui ont reçu des traits spécifiques marquant le sentiment positif ou négatif. Il s'agit pour la plupart de verbes (*aimer*, *apprécier*, *détester*, ...) et d'adjectifs (*magnifique*, *superbe*, *insupportable*, ...), mais aussi de quelques noms communs (*plaisir*) et d'adverbes (*malheureusement*). Par exemple, quand une relation de modifieur du nom est extraite (*paysage magnifique*) et que le modifieur (*magnifique*) porte le trait `sents`, la relation de sentiment ($\Rightarrow$ `SENTIMENT_POSITIF(magnifique,paysage)`) est extraite ensuite entre le nom et son modifieur. Après il y a évidemment des règles plus complexes pour extraire les relations des phrases plus compliquées.

L'attribut de la relation (`positif` ou `négatif`) d'un sentiment sera inversé quand une négation est présente dans la phrase, comme par exemple :

> « J'aime pas du tout les randonnées en montagne ! »
> « Ce n'est pas un mauvais restaurant. »

Quand cela est possible, les pronoms *qui* et *que* se rapportant à une entité présente ailleurs dans la même phrase, seront remplacés par cette même entité :

« Grenoble est une ville qui vaut vraiment le détour hiver comme été. »

Certains noms communs ainsi que des verbes de type interrogatif ont reçu un trait (`no-sents`) pour empêcher l'extraction de relations. Dans *Je cherche un bon hôtel.*, *Bon voyage !* ou *Bonne journée !* il ne s'agit pas de sentiments proprement dit exprimés par l'auteur du texte, mais plutôt de souhaits comme on peut les trouver surtout au début ou à la fin de messages. C'est pour cette raison que nous essayons d'éviter d'extraire ces relations.

2.3. *Annotation manuelle de textes*

La configuration de la grammaire générique a été faite sur la base d'un travail d'annotation manuelle (à l'aide du logiciel Protégé 3.2[4] avec le plugin Knowtator[5]) de *threads* venant du domaine du tourisme. Ce corpus du *tourisme* contient une centaine de *threads* annotés (avec comme sujet différentes régions et destinations en France). Chaque *thread* est composé de messages des utilisateurs du forum ; la longueur varie entre dix et 55 messages par document. Un message peut ne contenir qu'une phrase ou plusieurs paragraphes. L'annotation de ce corpus avec Protégé et Knowtator a été faite dans la lignée des travaux de (Riloff *et al.,* 2005), (Riloff *et al.,* 2006), (Wiebe *et al.,* 2006). L'annotation inclut les informations de cause, d'intensité et de l'émetteur du sentiment. Dans

« J'aime énormément Grenoble. »

aimer véhicule le sentiment, *Grenoble* est la cause du sentiment et *je* est l'émetteur du sentiment. L'adverbe *énormément* exprime l'intensité, le sentiment ici est plus intense que dans la phrase

« J'aime bien Grenoble. »

Cette phase d'annotation sert ensuite aussi à la méthode statistique pour l'élaboration d'un modèle à l'aide de l'entraînement des textes. Les phrases annotées positives seront séparées des négatives et un modèle statistique est créé ainsi. L'annotation correcte et précise des phrases est donc très importante pour les deux méthodes de traitement. L'annotation pour le *tourisme* ne contient pas seulement les deux valeurs *positif* et *négatif* pour classer les sentiments, mais est détaillée beaucoup plus finement (c.f. par exemple les travaux de (Mathieu, 2000), (Mathieu, 2006)). Le schéma d'annotation choisi est même plus fin et on voit donc que la classification des sentiments que l'on propose permet un grand nombre de modalités et va au delà de la simple opposition positif-négatif. En effet, nous avons repris la taxonomie d'(Ogorek, 2005) qui propose 33 sentiments différents (17 positifs et 16 négatifs) auxquels nous avons ajouté les pseudo-sentiments comme *bon-marché*, *conseil*, *cher* et *avertissement*, car dans le domaine du *tourisme* il y a beaucoup de messages concernant les prix des prestations dont les auteurs des messages sont contents (ou pas). Les sentiments de la taxonomie

4. `http://protege.stanford.edu/`
5. `http://bionlp.sourceforge.net/Knowtator/index.shtml`

d'Ogorek sont classés en groupes[6] comme AMOUR-DÉSIR (*amour, envie, tendresse, désir*), JOIE (*enchanté, excité, heureux, joyeux*), TRISTESSE-DÉTRESSE (*découragé, bouleversé, démoralisé, triste*), COLÈRE-DÉGOÛT-MÉPRIS (*colère, mépris, désapprobation*), etc.

3. Méthode statistique

Pour la méthode statistique, nous utilisons une technique d'apprentissage automatique qui se base sur les travaux de (Pang *et al.,* 2002), (Pang *et al.,* 2004), (Pang *et al.,* 2005). Nous l'avons adaptée aux corpus de langue française. Nous l'avons testée d'une part sur les *threads* du corpus sur le *tourisme*, et d'autre part sur les textes des corpus de DEFT'07. (Pang *et al.,* 2004) proposent deux axes de classification possibles, soit dans l'opposition subjectif-objectif, soit dans la distinction des opinions subjectives dans l'opposition positif-négatif. (Pang *et al.,* 2004) améliorent la classification de l'axe positif-négatif en supprimant d'abord du texte toutes les phrases objectives et en faisant la classification seulement sur la partie subjective. Cet *extract* correspond dans leurs expérimentations à 60 % du texte original. Nous n'avons pas retenu cette façon de faire à cause du manque d'un corpus d'entraînement ayant des parties subjectives et objectives bien distinctes. Nous avons choisi de faire la séparation de texte subjectif-objectif à l'aide de la méthode symbolique (*cf.* section 2) qui permet d'obtenir finalement des résultats plus nuancés. Les extraits peuvent être vus comme de bons résumés du texte au niveau des sentiments qu'ils expriment.

Des expérimentations ont été faites avec un des corpus de DEFT'07 qui contient des critiques de livres et films, en prenant seulement la/les première(s) et/ou la/les dernière(s) phrase(s) du message. Nous sommes partis de l'hypothèse que le jugement de l'auteur dans une critique de livre ou de film se trouve la plupart du temps en début ou en fin du message, la place du milieu étant vraisemblablement occupée par le résumé du livre ou du film. Les résultats de classification positif ou négatif avec cette technique sont meilleurs qu'en prenant le message en entier. Pourtant cette technique n'a finalement pas été retenue, car elle ne sera pas facilement reproductible sur des messages provenant d'autres domaines que la critique de film et de livre, où il n'y a pas forcément un résumé au milieu du message.

La méthode statistique se base sur des n-gram de caractères. Pour les projets sur la langue française (le tourisme, les jeux vidéo et DEFT'07) nous avons choisi $n = 12$. Comme pour la méthode symbolique, un indice de confiance est attribué aux textes. Il permet de comparer le résultat avec celui de la méthode symbolique pour en conclure le résultat final avec la méthode hybride. Pour l'entraînement des textes les techniques de *support vector machines* (SVM) et de *naive bayes* (NB) ont été utilisées. Les résultats sont légèrement meilleurs avec NB, mais ceci reste négligeable. L'entraînement du module statistique est donc réalisé uniquement sur les phrases de chaque *thread*

6. Sauf les pseudo-sentiments concernant les prix et conseils introduits par notre équipe comme *gratuit*, etc.

qui ont été sélectionnées par la méthode symbolique, qui contiennent donc des sentiments, et selon les valeurs de leur classe (positive ou négative) attribuées à chaque corpus par l'annotation manuelle. Les résultats sont ensuite confrontés aux résultats de la méthode symbolique pour donner un résultat final pour chaque message.

4. SYBILLE, la méthode hybride

La méthode hybride est une combinaison des deux méthodes précédentes. Elle prend en entrée les sorties des deux autres méthodes et calcule d'après les indices de confiance de chaque résultat, une moyenne qui sera traduite en positif ou négatif. La classification définitive est calculée avec les indices de confiance qui sont normalisés dans un premier temps pour donner une valeur entre 0 et 1. Ensuite les résultats respectifs sont confrontés pour obtenir une classification finale, le choix étant donné en priorité au résultat avec l'indice le plus élevé. La pondération exacte varie selon plusieurs facteurs, notamment la précision de l'annotation manuelle et la taille du corpus.

La méthode statistique permet de faire une première fouille dans les textes pour obtenir les messages positifs et négatifs. Ensuite l'utilisateur qui a configuré la grammaire peut modifier et améliorer celle-ci pour obtenir de meilleurs résultats. Le travail prend alors la forme d'un cycle où les résultats s'améliorent constamment. L'analyse du *thread* se fait au niveau des phrases et permet d'améliorer le résultat en ajoutant ou supprimant par exemple des mots au lexique. Ceci a l'avantage de montrer exactement quelles phrases du document expriment un sentiment.

C'est une approche qui permet de garder la robustesse de l'apprentissage automatique de la méthode statistique et d'orienter en même temps la base de l'entraînement sur une configuration manuelle de la méthode symbolique. Ceci permet de corriger de façon significative les erreurs de l'apprentissage automatique et d'intégrer les spécificités du cahier des charges, c'est-à-dire les particularités de chaque corpus (à l'aide de lexiques différents selon le domaine d'application).

Pour conclure, la figure 1 montre l'interface graphique du système SYBILLE, ici dans le domaine des imprimantes. En haut à droite il y a un champ dans lequel l'utilisateur peut faire une recherche **(1)** de messages qui contiennent un mot de son choix, sinon il peut choisir que les messages positifs ou négatifs **(2)**. Une autre façon de faire une recherche serait de se limiter aux messages qui n'évoquent qu'une marque précise **(3)**, ou en dessous un domaine d'application plus spécialisé **(4)**, ou encore un mot précis d'un domaine. On offre aussi l'option de sélectionner un forum donné parmi tous ceux qui ont été analysés. Les options de recherche peuvent être combinées à volonté pour limiter le nombre de réponses souhaitées. La relation de sentiment est indiquée avec ses arguments **(5)**, suivie de la phrase qui contient le sentiment et un lien (*external link* **(6)**) vers le *thread* entier qui permet de visualiser le contexte.

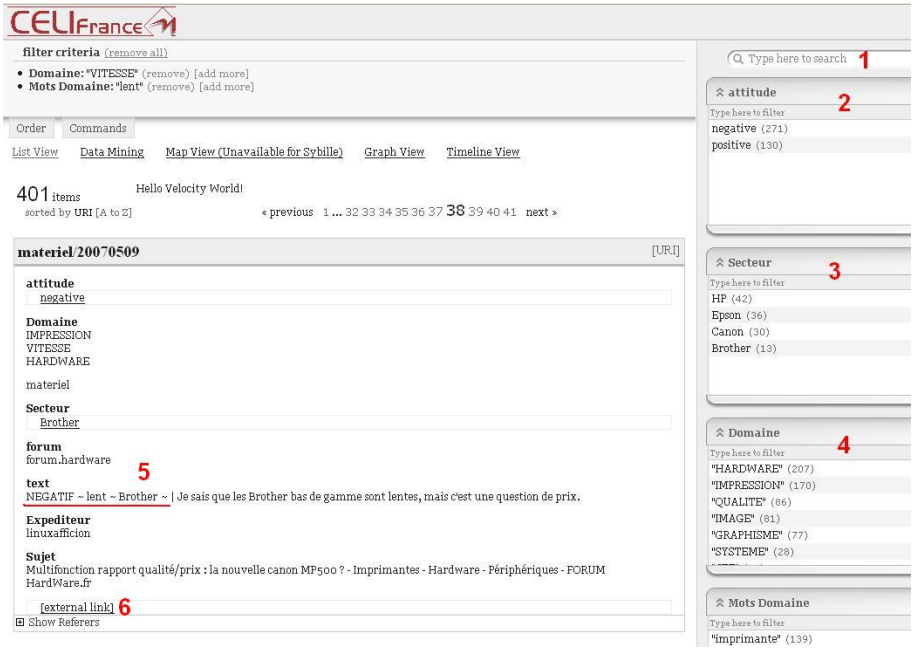

Figure 1. *L'interface graphique* SYBILLE*, ici pour le domaine des imprimantes. Différentes options de recherche dans la partie droite permettent d'accéder aux textes d'opinion.*

5. Évaluation

DEFT'07[7] (le DÉfi Fouille de Texte) est une campagne d'évaluation dont le thème était en 2007 la classification de textes d'opinion, présents dans différents types de textes. Plusieurs groupes de recherche (laboratoires universitaires et entreprises privées) ont pu tester leurs systèmes de classification sur les mêmes textes. Dans la phase initiale chaque groupe inscrit a reçu les deux tiers de chacun des quatre corpus différents qui avaient comme sujet des critiques de films et de livres, des tests de jeux vidéo, des relectures d'articles scientifiques et des notes de débats parlementaires. Pour les trois premiers corpus une note à trois valeurs (positif, moyen ou négatif) a été attribuée à chaque texte par le comité des organisateurs, une note à deux valeurs seulement (positif ou négatif) pour le dernier corpus. Après un certain temps pendant lequel chaque groupe a mis au point son ou ses systèmes de classification un troisième tiers de chaque corpus a été envoyé pour faire les tests dont les résultats ont dû être soumis quelques jours plus tard. Dix équipes ont participé à l'édition 2007, CELI France est

7. `http://deft07.limsi.fr/`

arrivée à la troisième place, les deux premières équipes sont issues de laboratoires de recherche universitaires.

La grammaire de l'analyseur a été paramétrée pour répondre aux besoins des différents corpus DEFT'07, du point de vue lexical mais aussi pour résister aux fautes d'orthographes répétitives. Le point le plus important à modifier a été la classification du message entier qui peut contenir plusieurs sentiments avec une seule valeur globale, et en particulier l'introduction de la notion de sentiment moyen. Notre approche standard est qu'au niveau des phrases les sentiments sont positifs ou négatifs. Il n'est pas nécessaire d'utiliser des sentiments moyens dans le domaine du tourisme, dans la mesure où la taxonomie utilisée (*cf.* section 2.3) permet de nuancer suffisamment. Les sentiments moyens pour DEFT'07 n'ont pas été extraits à l'aide de mots dans le lexique avec un trait moyen, mais d'après des structures de phrase. Par exemple à une phrase qui contient un sentiment positif et un sentiment négatif coordonnés par *mais* est attribué un sentiment moyen à la place :

« Ce jeu est *amusant* au début **mais** *ennuyant* la deuxième semaine. »

Quelques mots clés (surtout des adverbes comme *malgré, pourtant, ...*) sont utilisés pour aider à classifier un texte qui contient des phrases avec des sentiments positifs et négatifs (c.f. les travaux de (Sándor, 2005)). Le texte entier est alors classé comme moyen.

La méthode hybride a été utilisée pour les corpus de DEFT'07 *aVoiraLire*, *jeuxvidéo* et *relectures*. Elle a donné les meilleurs résultats pour les corpus *jeuxvidéo* avec un F-score[8] de 0,71, contre 0,54 (méthode symbolique) et 0,70 (méthode statistique) et *relectures* avec un F-score de 0,54, contre 0,48 (méthode symbolique) et 0,51 (méthode statistique).[9]

6. Conclusion

Dans cet article nous avons présenté comment le traitement automatique du langage naturel peut améliorer la qualité d'un système d'extraction de sentiments. Nous avons décrit une méthode symbolique avec une grammaire adaptée au domaine des textes, et une méthode d'apprentissage automatique. L'évaluation de notre système de classification SYBILLE a montré que la combinaison des méthodes symbolique et statistique donne des résultats plus précis que chacune des méthodes employée séparément.

---

8. Le F-score a été calculé de la manière suivante : $F_{score}(\beta) = \frac{(\beta^2+1)\times Précision\times Rappel}{\beta^2\times Précision+Rappel}$, avec $\beta = 1$; c.f. `http://deft07.limsi.fr/resultat.php#evaluation`.

9. Pour le corpus *aVoiraLire* le meilleur résultat a été obtenu par la méthode statistique avec un F-score de 0,52, contre 0,51 (méthode hybride) et 0,42 (méthode symbolique). Ce corpus n'est probablement pas assez uniforme (il parle de livres, films actuels au cinéma, disques, films plus anciens enregistrés, ...) pour pouvoir constituer un lexique plus performant.

L'intérêt de la méthode hybride repose sur la prise en compte des contextes d'application de ses résultats. Il est bien connu que la méthode purement symbolique a souvent pour le client un coût d'entrée plutôt élevé. Cette considération est liée au temps de configuration, de repérage ou de création de lexiques spécifiques, de taxonomies etc. L'utilisation d'une méthode hybride permet, au contraire, de minimiser les coûts de configuration, en réduisant une partie du travail à l'annotation de textes, une tâche qui dans la plupart des cas peut être réalisée par le client lui-même. Les algorithmes d'apprentissage automatique sont alors en mesure de donner des premiers jugements au niveau du texte entier.

Ce qui est le plus important, c'est qu'avec ce type de système statistique on peut ajouter, selon la méthode exposée dans cet article, une couche *symbolique* au fur et à mesure, de plus en plus importante dès que les exigences d'une application deviennent plus précises. On peut par exemple superposer une couche d'identification de jugement, qui permet d'avoir une visibilité sur les jugements sans devoir lire le texte dans son entier. On peut identifier certains patrons sémantiques qui sont d'importance capitale pour une application donnée et qui doivent avoir la priorité sur les résultats statistiques (par exemple le souci de securité exprimé par les internautes sur un certain modèle de voiture). Les exemples pourraient être multipliés. Ce qui est intéressant c'est que la démarche hybride est importante non seulement pour des raisons scientifiques de performance (le meilleur résultat entre les technologies que nous avons adoptées) mais, aussi et surtout pour des raisons de développement et d'acceptation par le marché.

7. Bibliographie

Aït-Mokhtar S., Chanod J.-P., « Subject and object dependency extraction using finite-state transducers », *in* , P. Vossen, , G. Adriaens, , N. Calzolari, , A. Sanfilippo, , Y. Wilks (eds), *Automatic information extraction and building of lexical semantic resources for NLP applications*, Association for Computational Linguistics, p. 71-77, 1997.

Aït-Mokhtar S., Chanod J.-P., Roux C., « A multi-input dependency parser », *Actes d' IWPT'01*, 2001.

Basili R., Pazienza M. T., Zanzotto F. M., « Lexicalizing a shallow parser », *Actes de TALN'99*, 1999.

Bethard S., Yu H., Thornton A., Hatzivassiloglou V., Jurafsky D., « Automatic extraction of opinion propositions and their holders », *Actes d' AAAI'04*, 2004.

Chklovski T., « Deriving quantitative overviews of free text assessments on the web », *Actes d' IUI'06*, p. 155-162, 2006.

Dini L., « Compréhension multilingue et extraction de l'information », *in* , F. Segond (ed.), *Multilinguisme et traitement de l'information (Traité des sciences et techniques de l'information)*, Editions Hermes Science, 2002.

Dini L., Mazzini G., « Opinion classification through information extraction », *in* , A. Zanasi, , C. A. Brebbia, , N. F. F. Ebecken, , P. Melli (eds), *Data Mining III*, WIT Press, p. 299-310, 2002.

Dini L., Segond F., « La linguistique informatique au service des sentiments », *Revue de l'électricité et de l'électronique*, Editions SEE, p. 66-77, 2007.

Grouin C., Berthelin J.-B., El Ayari S., Heitz T., Hurault-Plantet M., Jardino M., Khalis Z., Lastes M., « Présentation de DEFT'07 (DÉfi Fouille de Textes) », *Actes de DEFT'07*, p. 1-8, 2007.

Hatzivassiloglou V., McKeown K. R., « Predicting the semantic orientation of adjectives », *Actes d' ACL'97*, p. 174-181, 1997.

Kim S.-M., Hovy E., « Determining the sentiment of opinions », *Actes de COLING'04*, p. 1267-1373, 2004.

Mathieu Y. Y., *Les verbes de sentiment. De l'analyse linguistique au traitement automatique*, CNRS Editions, 2000.

Mathieu Y. Y., « A computational semantic lexicon of french verbs of emotion », *in* , J. G. Shanahan, , Y. Qu, , J. Wiebe (eds), *Computing attitude and affect in text*: *Theorie and applications*, Springer, p. 109-124, 2006.

Maurel S., Curtoni P., Dini L., « Classification d'opinions par méthodes symbolique, statistique et hybride », *Actes de DEFT'07*, p. 111-117, 2007.

Maurel S., Curtoni P., Dini L., « Sybille, anatomie d'un système automatique d'extraction de sentiments », *in* , A. Tutin, , I. Novakova (eds), *Le lexique des émotions et sa combinatoire syntaxique et lexicale*, Éditions de l'Université Stendhal, 2008, à paraître.

Ogorek J. R., « Normative picture categorization: Defining affective space in response to pictorial stimuli », *Actes de REU'05*, 2005.

Pang B., Lee L., « A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts », *Actes d' ACL'04*, p. 271-278, 2004.

Pang B., Lee L., « Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales », *Actes d' ACL'05*, p. 115-124, 2005.

Pang B., Lee L., Vaithyanathan S., « Thumbs up? Sentiment classification using machine learning techniques », *Actes d' EMNLP'02*, p. 79-86, 2002.

Riloff E., Patwardhan S., Wiebe J., « Feature subsumption for opinion analysis », *Actes d' EMNLP'06*, p. 440-448, 2006.

Riloff E., Wiebe J., Phillips W., « Exploiting subjectivity classification to improve information extraction », *Actes d' AAAI'05*, 2005.

Sándor A., « A framework for detecting contextual concepts in texts », *Actes du Electra Workshop*, 2005.

Sproull L., Kiesler S., *Connections*: *New ways of working in the networked organization*, Cambridge: MIT Press, 1991.

Turney P. D., « Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews », *Actes d' ACL'02*, 2002.

Wiebe J., Mihalcea R., « Word sense and subjectivity », *Actes d' ACL'06*, p. 1065-1072, 2006.

Wilson T., Wiebe J., Hwa R., « Just how mad are you? Finding strong and weak opinion clauses », *Actes d' AAAI'04*, 2004.

Yu H., Hatzivassiloglou V., « Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences », *Actes d' EMNLP'03*, p. 129-136, 2003.