Classification d'opinions par méthodes symbolique, statistique et hybride

Sigrid Maurel, Paolo Curtoni et Luca Dini

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

Résumé : La classification automatique de textes d'opinion est le DÉfi Fouille de Texte 2007. CELI-France présente trois méthodes différentes pour effectuer cette classification de textes de quatre corpus différents du DEFT'07. La première méthode est symbolique, la seconde statistique et la dernière, hybride, est une combinaison des deux premières. La combinaison permet de tirer parti des avantages des deux méthodes, à savoir la robustesse de l'apprentissage automatique statistique et la configuration manuelle symbolique orientée par rapport à l'utilisation d'applications réelles. L'*opinion mining* se fait au niveau de phrase, puis le texte entier est classé selon sa polarité en positif, moyen ou négatif.

Mots-clés : méthodes symbolique, statistique et hybride, classification d'opinions, analyse au niveau de phrase.

1 Introduction

DEFT'07 est la troisième édition de la conférence d'évaluation du DÉfi Fouille de Texte. Le thème de cette année est la classification de textes d'opinion, présents dans différents domaines de textes. Le corpus comprend des critiques de films, de livres et de disques, et de jeux vidéos. Dans un tout autre genre il y a des textes de débats politiques, et finalement des relectures d'articles scientifiques.

Un texte d'une critique de film, d'un article, etc. contient un jugement argumenté de l'auteur du texte, positif, moyen ou négatif, sur un sujet donné. Mais il contient aussi des parties sans sentiments, par exemple dans le résumé du livre sur lequel porte la critique. Le défi est de trouver avec précision les parties pertinentes pour la classification automatique du texte entier.

CELI-France a mis au point trois méthodes pour classer les textes des différents corpus mis à disposition par le comité d'organisation du DEFT'07. Ces trois méthodes ont servi pour traiter les corpus *aVoiraLire* (critiques de films, livres, disques, etc.), *jeuxvidéo* (critiques de jeux vidéo) et *relectures* (relectures d'articles scientifiques) ; seule la méthode statistique a été utilisée pour le corpus des *débats politiques* (notes de débats parlementaires).

La première est une méthode symbolique qui inclut un système d'extraction d'information (adapté aux corpus). Elle est basée sur les règles d'un analyseur syntaxico-sémantique. La deuxième est une méthode statistique basée sur des techniques d'apprentissage automatique. Enfin, une dernière méthode combine les techniques des deux précédentes.

Après une brève présentation des corpus, nous décrivons les trois méthodes développées.

2 Les corpus

Les corpus mis à disposition par DEFT'07 sont assez différents les uns des autres, que ce soit par la taille des corpus que par la taille de chaque texte. Certains sont structurés, d'autres non, et le nombre de fautes d'orthographe présent dans les textes varie aussi beaucoup (pour plus de détails c.f. section 3.2.1).

Les textes sur les jeux vidéo sont en général beaucoup plus longs que les textes de critiques de films et livres. Ils sont aussi plus structurés avec des parties différentes concernant le graphisme, le scénario, la jouabilité, etc. La longueur des critiques de films et livres varie beaucoup d'un texte à l'autre, on ne peut pas en dégager de structure interne.

Environ un tiers des relectures d'articles scientifiques est structuré en différentes parties : la rédaction, l'originalité, l'importance, etc., mais les deux-tiers restants ne le sont pas, et en général ces textes sont beaucoup plus denses que ceux du premier tiers.

La longueur des textes rapportant des débats politiques varie énormément. Ce corpus est le plus important en taille, il contient environ dix fois plus de textes que celui des jeux vidéo et des critiques de films et livres, et presque vingt fois plus que celui des relectures d'articles.

Aux trois premiers corpus sont associées trois classes de jugement (*positif*, *moyen* et *négatif*), alors que deux classes de jugement seulement (*pour* (c'est-à-dire *positif*) et *contre* (*négatif*)) sont associées au dernier corpus.

3 Méthode symbolique

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 (Aït-Mokhtar *et al.*, 2001; Basili *et al.*, 1999; Dini, 2002; Dini & Mazzini, 2002)). Cet analyseur traite un texte donné en entrée phrase par phrase et en extrait, pour chaque phrase, les relations syntaxiques présentes. Il s'agit de relations syntaxiques de base telles que le modifieur d'un nom, d'un verbe, le sujet et l'objet de la phrase, etc., et de relations plus complexes telles que la coréférence entre deux syntagmes de la phrase.

La possibilité est donnée à l'utilisateur d'élaborer une grammaire à sa guise et d'ajouter de nouvelles règles pour 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 de nouvelles relations), augmenter/diminuer les traits sur les mots dans le lexique qui agissent sur les règles, enlever certaines parties du traitement, etc.

À la fin de l'analyse un indice de confiance est calculé par la méthode symbolique. Il servira à la méthode hybride (c.f. section 5) pour déterminer le résultat final.

La polarité POS/NEG/MOY attribuée au texte entier dépend du rapport entre la quantité de relations d'opinions positives et négatives. Celle-ci est évaluée en relation avec la quantité de relations d'opinions moyennes. Une majorité de relations d'opinions positives détermine une polarité positive du texte, tandis qu'une majorité de relations d'opinions négatives provoque une polarité négative du texte entier. L'équilibre entre relations d'opinions positives et négatives entraine une classification moyenne du texte. La quantité de relations d'opinion moyenne influence la quantité de relations d'opinions positives et négatives nécessaire afin que soit attribué au texte une polarité positive ou négative.

L'algorithme pour la classification utilise les formules suivantes :

```
POS, NEG, MOY = nombre d'opinions positives, négatives et moyennes

BalanceInterval=((0.133)*(MOY/(POS+NEG+MOY)))+(0.066)

PosNegRatio=(POS/(POS+NEG))+((0.5-(POS/(POS+NEG)))*(MOY/(POS+NEG+MOY)))

IF((0.5-BalanceInterval) <= PosNegRatio <= (0.5+BalanceInterval))
     ASSIGNMENT=MOY
     SCORE=(BalanceInterval-ABS(PosNegRatio-0.5))/BalanceInterval
ELSE IF(POS>NEG)
     ASSIGNMENT=POS
     SCORE=POS/(POS+NEG+MOY)
ELSE
     ASSIGNMENT=NEG
     SCORE=NEG/(POS+NEG+MOY)
```

La méthode symbolique a été utilisée pour les corpus *aVoiraLire*, *jeuxvidéo* et *relectures d'articles*. Les textes du corpus des *débats politiques* contiennent des textes trop pauvres en sentiments exprimés directement, c'est-à-dire qu'ils contiennent trop peu, voire pas du tout, de mots ou de syntagmes qui expriment des sentiments et qui obéissent aux règles définies par la grammaire. La raison de ce comportement de l'analyseur est que le style des discours parlementaires est trop éloigné du style des textes du tourisme pour lequel la grammaire a initialement été développée (c.f. la section 3.1).

Les textes du corpus du tourisme utilisés pour configurer la grammaire de base proviennent d'un site de forums sur internet. Le style des forums et des newsgroups est assez particulier. Le style des discours

parlementaires est beaucoup plus formel, les phrases sont plus longues et le langage est plus soutenu. C'est pour cette raison que seule la méthode statistique a été utilisée pour ce dernier corpus.

3.1 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. Une relation de sentiment a en général deux arguments : le premier est l'expression linguistique qui véhicule le sentiment en question, le deuxième la cause 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.

L'analyse se base sur les mots du lexique qui ont reçu des traits spécifiques qui marquent le sentiment positif ou négatif. Les relations de sentiments moyens sont extraites sur la base de constructions de phrases. 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*). L'attribut de la relation (`positif` ou `négatif`[1]) d'un sentiment sera inversé en cas d'une négation dans la phrase. Si possible, les pronoms *qui* et *que* qui se rapportent à une entité présente ailleurs dans la même phrase, seront remplacés par cette entité (e.g. « Grenoble est une ville qui vaut vraiment le détour hiver comme été. »).

Cette 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[2] avec le plugin Knowtator[3]) de textes venant du domaine du tourisme. Ce corpus du tourisme contient une centaine de textes dont environ 75 ont été annotés. Chaque texte 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. Le corpus entier occupe un peu plus d'un mégaoctet sur le disque dur. L'annotation de ce corpus avec Protégé et Knowtator a été faite dans la lignée des travaux de (Riloff *et al.*, 2006; Wiebe & Mihalcea, 2006; Riloff *et al.*, 2005).

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. ».

L'annotation pour le tourisme ne contient pas seulement les valeurs *positif* et *négatif* pour classer les sentiments, mais est détaillée beaucoup plus finement (voir par exemple les travaux de (Mathieu, 2000, 2006)). Le schéma d'annotation choisi est plus fin que la simple opposition positif-négatif. En effet, nous avons repris la taxonomie d'(Ogorek, 2005) qui propose 33 sentiments (17 positifs et 16 négatifs) différents auxquels nous avons ajouté les pseudo-sentiments *bon-marché*, *conseil*, *cher* et *avertissement*. Ces sentiments sont classés en sous-groupes comme *amour*, *joie*, *tristesse*, *mépris*, etc.

3.2 Grammaires pour DEFT'07

Nous avons dû adapter la grammaire de base pour participer à DEFT'07. Elle 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 des sentiments et en particulier l'introduction de la notion de sentiment moyen. En effet pour le tourisme il n'y a que des sentiments positifs et négatifs. Notre approche pour le projet du tourisme 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 (c.f. ci-dessus) permet de nuancer suffisamment.

Nous avons donc construit deux nouvelles grammaires différentes de celle utilisée initialement. La première, commune aux corpus *aVoiraLire* et *jeuxvidéo*, la deuxième, pour le corpus des *relectures*.[4]

[1] L'attribut `moyen` évoqué dans l'introduction n'existe pas dans la grammaire de base, il a été ajouté seulement dans les grammaires DEFT'07 pour répondre aux besoins des corpus.

[2] `http://protege.stanford.edu/`

[3] `http://bionlp.sourceforge.net/Knowtator/index.shtml`

[4] Rappel : Il n'existe pas de grammaire spéciale pour le corpus des *débats politiques*, ce corpus n'a été traité uniquement que par la méthode statistique (c.f. section 4).

L'objectif de ces trois grammaires (celle de base (c.f. la section précédente 3.1) et celles pour DEFT'07) est d'extraire un maximum d'informations du texte, les sentiments positifs, moyens et négatifs. Au final, CELI-France souhaite utiliser les méthodes de l'*opinion mining* et pas seulement de la classification textuelle. L'analyse des textes au niveau des phrases convient bien à cette approche.

Les textes 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. À la fin de l'analyse du texte un sentiment global lui est attribué selon le nombre de relations positives, moyennes et négatives qu'il contient.

Les deux grammaires DEFT'07 se distinguent l'une de l'autre essentiellement par le lexique de mots qui reçoivent les traits `positif` et `négatif` correspondants aux valeurs des classes des textes (et par leur liste de termes, c.f. section 3.3). Par exemple, le lexique de la grammaire des *relectures* contient les mots *bon* et *bien* (« Cet article est *bien* écrit. », « Le papier est d'une *bonne* qualité. »). Ces mots ont été supprimés du lexique de la grammaire *aVoiraLire/jeuxvidéo* parce qu'ils produisent trop de relations éronnées (« Il a fait *bien* des choses avant de se mettre à faire des films. », « L'actrice est venue au *bon* moment. »). Souvent il ne s'agit pas de vrais sentiments.

Des règles spéciales pour créer des relations pour les textes avec un jugement de sentiment moyen ont été ajoutées. Par exemple quand une phrase contient un sentiment positif et un sentiment négatif coordonnés par *mais*, l'attribut des sentiments est remplacé par l'attribut `moyen` (« Ce jeu est *amusant* au début **mais** *ennuyant* la deuxième semaine. »). Ceci aide à classifier un texte qui contient des phrases avec des sentiments positifs et négatifs. Le texte entier sera alors classé comme moyen.

3.2.1 Les fautes d'orthographe dans les textes

Les fautes d'orthographe dans les textes des corpus posent parfois des problèmes d'analyse. Un mot où l'accent est mal mis peut changer complètement l'analyse. Prenons la phrase « Cet article me parait inacceptable car extrêmement confus et peu compréhensible, en tout cas par quelqu'un ne disposant que des informations contenues dans l'article. » Ici, il manque l'accent circonflexe sur le *i* de *parait*, le verbe. L'analyse retourne l'infinitif *parer* au lieu de *paraître*. La relation d'objet prédicatif entre *me* et *paraître* ne peut pas être extraite. C'est une condition indispensable pour l'extraction de la relation de sentiment entre *article* et *inacceptable*.

Heureusement, la grammaire permet d'intercepter certains des cas les plus fréquents, par exemple par l'ajout de mots mal orthographiés au lexique (*interressant*, sans accent et avec deux *r*, est apparu plusieurs fois dans les corpus.).

Les règles syntaxiques sont souvent assez souples pour faire l'accord entre un nom et un adjectif, ou un nom et un verbe même si le *e* ou le *s* manque. Mais malheureusement, il y a aussi des textes tellement mal écrits dans les corpus que l'analyse peut échouer. D'après ce que nous avons pu observer ces textes sont heureusement en minorité dans les corpus et ne modifient pas trop les résultats.

3.3 Listes de termes

Une liste de termes a été élaborée pour chaque corpus. Chaque liste contient les noms qui sont propres au domaine du corpus, voici quelques exemples. Pour le corpus *aVoiraLire* : *film*, *livre*, *album*, ... ; pour le corpus *jeuxvidéo* : *jeu*, *graphisme*, *soft*, ... ; et pour le corpus *relectures* : *article*, *papier*, *résultat*, ... Grâce à ces listes, des relations erronées, c'est-à-dire dont le deuxième argument n'est pas dans la liste parce qu'il n'appartient pas au domaine, peuvent être refusées.

Par exemple dans le corpus *aVoiraLire* cette mesure s'applique à la plupart des relations extraites de la partie résumé du film, du livre, etc. Prenons la phrase suivante : « Le héros a passé une *magnifique* journée. » Le mot *journée* n'étant pas dans la liste, la relation `SENTIMENT_POSITIF(magnifique,journée)` est ainsi refusée. Ce qui répond bien au défi car cette phrase ne contient pas un sentiment exprimé de l'auteur du texte (ce qui serait le cas par exemple la phrase « Ce livre est vraiment *magnifique*. »), mais fait partie du résumé de l'histoire.

Chaque liste ne contient que des noms communs, toutes les relations qui contiennent des noms propres sont gardées telles quelles (« Bref, inutile de dépenser le moindre euro pour ce *Yetisports* qui n'en vaut vraiment pas la chandelle. »). Les relations extraites à partir de mots qui ne sont pas des noms (« Je n'aime pas aller au cinéma. » ⇒ `SENTIMENT_NEGATIF(aimer,aller)`) sont gardées elles aussi.

Ces listes ont été élaborées automatiquement à partir des textes des corpus d'entrainement. Elles contiennent tous les noms en deuxième argument d'une relation, si cette relation a été extraite d'un texte de la même valeur de la classe et que l'indice de confiance calculé dépasse un seuil prédéfini. C'est-à-dire : un nom dans une relation de sentiment positif doit se trouver dans un texte avec la valeur *positif* de la classe dans le corpus d'entrainement.

À l'intérieur de chaque liste, les termes sont ordonnés par la fréquence avec laquelle ils ont satisfait les conditions d'extraction en relations. L'utilisation de ces listes permet d'augmenter le F-score des résultats d'environ 5-10%, selon le corpus.

4 Méthode statistique

La méthode statistique est une technique d'apprentissage automatique qui se base sur (Pang & Lee, 2004; Pang *et al.*, 2002; Pang & Lee, 2005)[5]. Nous l'avons adaptée aux corpus de langue française. Nous l'avons testé d'une part sur les textes du projet sur le tourisme, et d'autre part sur les textes des corpus du projet DEFT'07. (Pang & Lee, 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. La technique de base de la méthode de (Pang & Lee, 2004) ne considère donc pas les sentiments moyens. C'est la raison pour laquelle nous avons dû ajouter une méthode pour ceux-ci.

(Pang & Lee, 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'entrainement 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 (c.f. section 3) qui permet d'obtenir finalement des résultats plus nuancés. Les extraits peuvent être vus comme de bons résumés du texte.

La méthode statistique se base sur des n-gram. Pour les projets sur la langue française (DEFT'07 et le tourisme) 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'entrainement des textes pour ce projet, 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.

Nous avons travaillé dans un premier temps avec la partie qui distingue les phrases subjectives des phrases objectives. Malheureusement, les résultats ne sont pas très encourageants, ceci est dû au fait que dans les corpus d'entrainement, les parties subjectives ne sont pas clairement distinctes des parties objectives. Souvent, les deux parties sont mélées l'une à l'autre.

Des expérimentations ont été faites avec le corpus *aVoiraLire*, en prenant seulement la/les première(s) et/ou la/les dernière(s) phrase(s) du texte. 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 au début ou à la fin du texte, 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 texte en entier, mais cette technique n'a finalement pas été retenue, car elle ne sera pas facilement reproductible sur des textes provenants d'autres domaines que la critique de film et de livre, où il n'y a pas forcément un résumé au milieu du texte.

Pour le projet DEFT'07 l'entrainement du module statistique a donc été réalisé uniquement sur les phrases de chaque texte qui ont été sélectionnées par la méthode symbolique, qui contiennent donc des sentiments, et selon les valeurs de leur classe (positif, moyen ou négatif) attribuées à chaque corpus par le comité d'organisation. Les résultats soumis correspondent à cet entrainement. Ils sont ensuite confrontés aux résultats de la méthode symbolique pour donner un résultat final pour chaque texte.

5 Méthode hybride

La méthode hybride est une combinaison des deux méthodes précédentes (c.f. sections 3 et 4). 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, moyen ou négatif.

[5] `http://www.cs.cornell.edu/home/llee/papers.html`

Cette méthode a été utilisée pour trois des quatre corpus donnés (*aVoiraLire*, *jeuxvidéo* et *relectures*). Elle a donné les meilleurs résultats pour les corpus *jeuxvidéo* avec un F-score 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).[6] Pour le corpus *débats politiques* seule la méthode statistique a été utilisée.

La classification définitive est calculée avec les deux méthodes symbolique et statistique. Les résultats respectifs sont confrontés pour obtenir une classification finale. La pondération exacte est encore sujet d'expérimentations et en cours de travail.

Grâce au refus de certaines relations de sentiment erronées par la méthode symbolique, l'entrainement de la méthode statistique se fait sur un corpus plus homogène, et le résultat sera plus précis.

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'entrainement 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 et de listes de termes différents selon le domaine d'application).

6 Conclusion

La combinaison des méthodes symbolique et statistique a donné des résultats plus précis que chacune des méthodes employée séparément.

La figure 1 résume les résultats obtenus. Pour les *débats politiques* il n'y a qu'une valeur du F-score puisque ce corpus n'a été traité que par la méthode statistique.

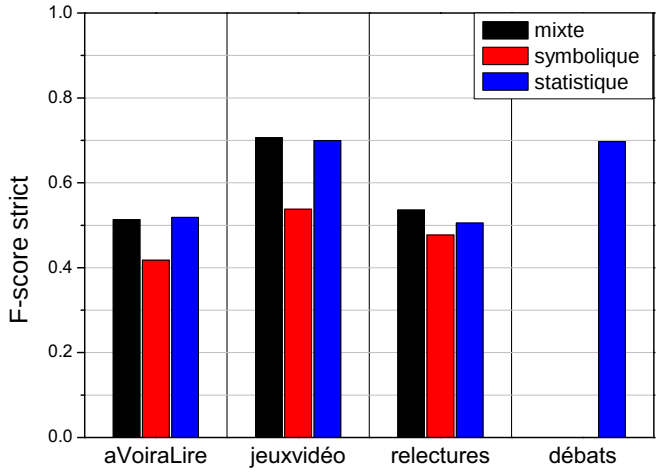

FIG. 1 – Les résultats, le F-score strict, de chaque corpus selon la méthode de classification utilisée.

L'intérêt de la méthode hybride repose dans la considération des contextes d'application de ses produits. 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 la création de lexiques spécifiques, de taxonomies etc. On voit donc que la méthode hybride obtient de meilleurs résultats, cela permet de profiter des avantages d'une approche symbolique sans pour autant en avoir tous les inconvénients.

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

[6] 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 faire une liste de termes plus performante.

par le client lui-même. Les algorithmes d'apprentissage automatique sont alors en mesure de donner des premiers jugements au niveau du texte entier.

Or, le jugement donné par rapport à un texte est souvent d'une utilité limitée. Par exemple savoir que, dans un forum sur la téléphonie, il y a 30.000 messages à polarité positive et 15.000 à polarité négative, n'est pas le type d'information qui peut être activement utilisée par une compagnie de téléphonie. Déjà en intégrant la polarité avec un système d'extraction d'entités nommées (marque, modèle, etc.), on peut avoir des résultats plus spécifiques et donc plus informatifs.

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 soucis de securité exprimé par les internautes sur un certain modèle de voiture).

Les exemples pourraient être multipliés. Ce qui est important est que, pour des raisons commerciales, la démarche hybride que nous avons tenue en DEFT'07 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 acceptation par le marché.

Références

AÏT-MOKHTAR S., CHANOD J.-P. & ROUX C. (2001). A multi-input dependency parser. In *Actes d' IWPT*.

BASILI R., PAZIENZA M. T. & ZANZOTTO F. M. (1999). Lexicalizing a shallow parser. In *Actes de TALN'99*.

DINI L. (2002). 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.

DINI L. & MAZZINI G. (2002). Opinion classification through information extraction. In ZANASI, BREBBIA, EBECKEN & MELLI, Eds., *Data Mining III*, p. 299–310. WIT Press.

MATHIEU Y. Y. (2000). *Les verbes de sentiment. De l'analyse linguistique au traitement automatique*. CNRS Editions.

MATHIEU Y. Y. (2006). 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*, p. 109–124. Springer.

OGOREK J. R. (2005). Normative picture categorization: Defining affective space in response to pictorial stimuli. In *Actes de REU'05*.

PANG B. & LEE L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Actes d' ACL'04*, p. 271–278.

PANG B. & LEE L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Actes d' ACL'05*, p. 115–124.

PANG B., LEE L. & VAITHYANATHAN S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Actes d' EMNLP'02*, p. 79–86.

RILOFF E., PATWARDHAN S. & WIEBE J. (2006). Feature subsumption for opinion analysis. In *Actes d' EMNLP'06*, p. 440–448.

RILOFF E., WIEBE J. & PHILLIPS W. (2005). Exploiting subjectivity classification to improve information extraction. In *Actes d' AAAI'05*.

WIEBE J. & MIHALCEA R. (2006). Word sense and subjectivity. In *Actes d' ACL'06*, p. 1065–1072.