Sémantique textuelle et TAL : un exemple d'application à l'analyse des sentiments

Egle Eensoo, Mathieu Valette

Sémantique textuelle et TAL : un exemple d'application à l'analyse des sentiments

Egle Eensoo Mathieu Valette

Équipe de Recherche Textes, Informatique, Multilinguisme (EA 2520 ERTIM)

INALCO, 2 rue de Lille, 75343 Paris Cedex 07

egle.eensoo@inalco.fr, mathieu.valette.inalco.fr

Résumé

Notre objectif est de faire rencontrer les pratiques courantes de Traitement Automatique des Langues (TAL) avec celles de la sémantique textuelle outillée par la textométrie. Nous dressons un panorama de ces deux pratiques qui établissent peu de liens alors qu'elles travaillent avec le même matériau textuel. Afin de provoquer une rencontre, nous nous appuyons sur l'exemple d'une application d'analyse des sentiments. Comme c'est le cas pour beaucoup d'autres applications du domaine, nous utilisons les algorithmes d'apprentissage automatique mais les descripteurs choisis sont issus d'une analyse textométrique du corpus et structurés suivant une grille de lecture inspirée de la sémantique textuelle. Ainsi, nous mettons en évidence que les critères axiologiques (thymiques) habituellement utilisés dans l'analyse de sentiments ne contribuent que marginalement dans la classification. D'autres critères (dialogiques et dialectiques), liés au déroulement temporo-aspectuel du récit, aux rapports entre les protagonistes, etc. apportent en revanche une amélioration significative par rapport à la ligne de comparaison.

1. Introduction

Le TAL a connu ces 20 dernières années l'essor des méthodes par apprentissage automatique et le déclin corrélatif des méthodes symboliques. On estime que la proportion d'articles de traitement automatique du langage (TAL) intégrant une section apprentissage a progressé de 30 à 90 % du début des années 90 à la fin des années 2000 (Church 2011, cité par Tanguy 2012). Ces méthodes au succès incontestable dans bon nombre d'applications telles que la traduction automatique, la reconnaissance vocale, la fouille de texte ou la recherche d'information, tendent à écarter progressivement la linguistique du champ disciplinaire : non seulement l'apprentissage automatique permet d'obtenir de meilleurs résultats que les méthodes symboliques, mais sa mise en œuvre est rendue aisée par des outils variés et accessibles (par exemple le logiciel WEKA[1]). Pour une même application – par exemple la traduction automatique – les méthodes symboliques, jadis, mobilisaient pendant plusieurs années une armada de linguistes pour l'écriture de règles ; aujourd'hui, pour peu que des corpus parallèles de taille suffisante soient disponibles, un système par apprentissage nécessitera très peu de ressources humaines et de temps.

Dans ce contexte, la linguistique intervient essentiellement sur deux types de tâches : (i) les linguistes effectuent en amont une lecture « experte » des textes afin d'annoter manuellement les données qui seront ensuite traitées automatiquement par des algorithmes d'apprentissage ; (ii) ils sont parfois sollicités en aval de l'expérimentation pour discuter les résultats. Cette seconde tâche demeure toutefois épilogique et facultative, dans la mesure où l'évaluation proprement dite repose sur des mesures de congruence entre annotation manuelle

[1] http://www.cs.waikato.ac.nz/ml/weka/

et résultat de l'apprentissage[2]. En bref, le savoir-faire linguistique n'est pas ignoré mais peut relever de la sous-traitance.

La théorie linguistique est marginalisée quand elle n'est pas suspecte. Une application phare, la fouille de textes, s'inspire rarement d'une théorie linguistique et encore moins d'une théorie de la sous-discipline que serait la linguistique des textes. Pourtant, leur objet d'étude est le même. En élaborant des méthodes d'extraction des connaissances et de classification de textes, c'est en effet le TAL qui, aujourd'hui, traite massivement des corpus et est confronté à la complexité des textes, à la textualité. En somme, la science des textes actuellement, c'est le TAL informatique.

Fort de ce constat quelque peu pessimiste, nous nous sommes fixé l'objectif d'évaluer en quoi une linguistique des textes – et en particulier la sémantique des textes – est à même de participer à ce déploiement d'applications. La linguistique en effet n'est pas complètement démunie : la linguistique de corpus peut en effet être perçue comme un pré-outillage pour la fouille de textes, notamment, sa sous-discipline que constitue la textométrie. Cette dernière présente en outre des affinités méthodologique et théorique avec la sémantique textuelle (Pincemin 2010).

Bien que cela n'ait pas toujours été le cas, il serait aujourd'hui abusif de considérer la linguistique de corpus et la textométrie comme relevant du TAL. En dépit de quelques traits communs (les corpus numériques, les algorithmes mathématiques informatisés), elles se distinguent sur les points suivants : le TAL, fondamentalement, vise l'automatisation des processus, l'élimination de la part de l'humain dans les traitements, tandis que la textométrie repose sur une itération entre l'analyse des sorties logicielles et la consultation des textes ou de fragments ; en cela, il s'agit davantage d'une linguistique assistée par ordinateur. Par ailleurs, un part significative du TAL est utilitariste et a pour finalité les applications informatiques, ce qui implique une recherche de performance et d'optimisation ; la textométrie a des objectifs épistémiques : accroître les connaissance et participer à l'interprétation d'un corpus. Enfin, à la différence du TAL où la mise en place d'un protocole d'évaluation est indispensable, l'évaluation et la reproductibilité ne sont pas un enjeu en textométrie. Les études textométriques sont validées par homologation, c'est-à-dire par l'assentiment d'une communauté qui, dans le meilleur des cas, est distante (par exemple, communauté de la critique littéraire pour l'analyse textométrique de textes littéraires), mais parfois n'est peut-être qu'un avatar récent du jugement d'acceptabilité pourtant honni de ladite communauté.

Prenant acte de ces différences de culture, nous faisons le projet de jeter un pont entre la sémantique textuelle et le TAL par le truchement de la textométrie, afin de mutualiser les avantages d'une association entre celles-ci et une méthode de classification automatique basée sur l'apprentissage supervisé. Il s'agira d'évaluer la pertinence d'une analyse linguistique au moyen d'un protocole TAL. Nous illustrerons notre propos à partir d'une application en analyse des sentiments.

2. Panorama des méthodologies utilisées en analyse de la

[2] Nous rejoignons ici l'analyse d'Yvon 2006 : « on déplorera […] que l'évaluation quantitative des modules de traitements, réalisée le plus souvent de manière indépendante des autres modules, soit devenu le « mètre étalon » unique des apprentis linguistiques. Les autres évaluations possibles des travaux liant apprentissage et traitement automatique des langues : leur adéquation avec une théorie linguistique ou avec des observations comportementales, leur apprenabilité, leur plausibilité cognitive, leur stabilité (qu'on pense à des tâches non supervisées), l'interprétabilité des modèles induits ou de leurs paramètres, sont ainsi progressivement passés au second plan des préoccupations (Yvon, 2006, cité par Tanguy 2012).

subjectivité

Les différentes approches de l'analyse de la subjectivité ont un facteur commun, l'intérêt pour les valeurs, qui les différencient des applications TAL historiques qui s'intéressent au contenu factuel. Cependant, elles sont très variées tant au niveau des objectifs que des méthodes.

Les objectifs sont souvent liés à une demande sociale et économique croissante de gestion (et de contrôle) des flux d'information sur le Web en plein expansion depuis la généralisation des espaces d'expression individuelle : blogs, forums, réseaux sociaux, etc. Beaucoup d'applications TAL se focalisent sur la détection des opinions des consommateurs concernant des produits de consommation : automobiles, cinéma, jeux vidéo, voyages, téléphones portables. D'autres traitent des questions politiques ou sociétales : quel est la cote de popularité de telle personnalité politique (Twittoscope) ? les électeurs sont-ils favorables à une ligne politique (Hopkins et King, 2007) ? quelles positions politiques émergent des textes politiques (Benoit *et al.*, 2009) ? Il existe aussi des applications *ad hoc* comme la détection des attitudes (ex : agressivité) pour la modération automatique des forums ou des newsgroups (Spertus, 1997) ou la détection des émotions pour les systèmes interaction homme-machine (Liscombe *et al.*, 2005).

Quant aux méthodes, on peut diviser ces applications en quatre grandes catégories du point de vue de l'implication de la linguistique :

2.1. Les approches cognitivistes font appel à des ressources lexicales qui reposent sur des modèles supposant l'existence de catégories cognitives préétablies et indépendantes des langues. Il s'agit notamment des dérivés du WordNet (Ghorbel et Jacot, 2011 ; Kim et *al.*, 2010), des ressources basées sur la théorie Appraisal (Whitelaw et al., 2005 ; Gardin, 2009) et sur la taxinomie d'Ogorek (Maurel et Dini, 2009). Ces modèles contiennent un jeu plus ou moins complexe de catégories formalisant des états privés (émotions, sentiments, jugements, etc.). Si SentiWordNet (Esuli et Sebastiani, 2006) se contentent de deux axes – objectif/subjectif et positif/négatif – pour caractériser les unités lexicales, WordNet-Affect (Strapparava et Valitutti, 2004) emploient 11 a-labels (ex : *emotion*, *mood*, *cognitive state*, *behaviour*, *attitude*, *sensation*, etc.) et la théorie Appraisal un système complexe de 4 axes (*attitude*, *graduation*, *orientation*, *polarity*) avec des subdivisions qui caractérisent chaque état. Cependant, malgré la généricité ou la précision descriptive recherchée, les unités lexicales ne sont que les réalisations linguistiques d'états privés universaux qui viennent combler les lacunes des modèles conceptuels. Malgré les problèmes évidents d'ambiguïté et d'inadaptablilité aux différents domaines et buts applicatifs, cette vision continue à séduire le traitement automatique parce qu'elle satisfait son exigence de formalisation.

2.2. Les approches linguistiques « théoriques » se caractérisent par la revendication d'un cadre linguistique théorique. Parmi celles-ci les plus productives sont dans la lignée de linguistique énonciative de Benveniste (Ducrot, Anscombre, Kerbrat-Orecchioni, Charaudeau). Ces théories en analysant la langue du point de vue des sujets et de leur positionnement dans le discours (marques d'énonciation, axiologie, argumentation) se prêtent à la modélisation de la subjectivité. Les méthodes TAL ne prétendent pas modéliser ces théories mais s'inspirent des notions théoriques proposées pour les adapter à un besoin applicatif donné. A la différence des cognitivistes dont l'approche est essentiellement du type *descendante*, nous rencontrons ici une combinaison des approches *descendantes* et *ascendantes* que l'on peut décrire comme un va-et-vient entre la réalité textuelle et les catégories théoriques. Par exemple (Vernier *et al.*, 2009a) utilisent un lexique extrait de façon semi-automatique à partir d'un corpus annoté manuellement avec des catégories évaluatives de Charaudeau pour le projeter sur un nouveau corpus. Dans le même esprit, (Vernier *et al.*,

2009b) recourent à l'usage des déictiques et d'un lexique axiologique en y ajoutant cependant des descripteurs « empiriques » qui rendent compte des différents genres présents dans le corpus (erratum, courrier, interview).

2.3. Les approches linguistiques « opportunistes » exploitent des phénomènes linguistiques de surface accessibles aux moyens d'outils TAL, et de diverses ressources lexicales. Ces outils, traitant un des paliers linguistiques (étiqueteurs morphosyntaxiques, analyseurs syntaxiques) permettent une exploitation des phénomènes linguistiques de surface comme l'extraction des patrons morphosyntaxiques (Turney, 2002 ; Yi *et al.*, 2003), la sélection de certaines parties du discours comme les verbes ou les adjectifs (Hatzivassiloglou et Wiebe, 2000) ou l'utilisation des mots-pôles issus d'un lexique préétabli (Yu et Hatzivassiloglou, 2003). L'attribution des valeurs se base sur la mutualisation de divers mesures statistiques, qui sont au cœur du processus, et des ressources lexicales (Yi *et al.*, 2003) ou textuelles (Turney, 2002 ; Yu et Hatzivassiloglou, 2003).

2.4 Les approches purement apprentistes concernent les travaux où la linguistique intervient le moins et le texte est vu comme une suite de chaînes de caractères, ou au mieux comme la suite de *tokens*[3] ou de mots. Le cœur du système repose sur les algorithmes d'apprentissage supposés retenir les unités les plus pertinentes pour caractériser un texte par rapport à sa catégorie. Cela permet aussi une accumulation massive des divers descripteurs et exempte d'un choix reposant sur la connaissance des textes. Par exemple, (Pang *et al.*, 2002) utilisent avec des algorithmes d'apprentissage supervisé divers descripteurs comme des *tokens* simples incluant la ponctuation (*unigrammes*), des suites de *tokens* (*bigrammes*) ou encore des catégories morphosyntaxiques. Ils procèdent à plusieurs expérimentations en les isolant ou les accumulant. On notera que les meilleurs résultats ont été obtenus avec des *tokens* simples, ce qui valide deux hypothèses : a) les mots et la ponctuation représentent plutôt fidèlement la complexité textuelle souvent appauvrie par les techniques TAL (lemmatisation, utilisation des ontologies, etc.), b) dans le cas du corpus homogène, les algorithmes d'apprentissage sont suffisamment performants pour sélectionner les éléments textuels pertinents.

3. Présentation du corpus

Une étude de cas en analyse des sentiments confrontera une vision théorique linguistique aux besoins applicatifs du TAL. Nous mettons l'accent sur la textualité, jusqu'ici peu prise en compte en TAL et nous nous interrogeons sur les points suivants : Comment prendre en compte la textualité pour une application TAL traitant la subjectivité ? Qu'est-ce qu'une analyse textométrique du corpus peut apporter aux algorithmes de classification basés sur l'apprentissage automatique ? Qu'apportent ces algorithmes à la compréhension de la textualité ? En nous intéressant à l'impact de l'application sur des propositions théoriques, nous nous distinguons des travaux du TAL majoritaires qui se focalisent davantage sur les résultats numériques et la pertinence des mesures mathématiques adoptées.

3.1. Contexte applicatif de l'étude — Le corpus est constitué de 300 textes courts réunis par SAMESTORY (http://www.same-story.com), un service d'agrégation d'ego-documents. Il s'agit, en l'occurrence, de témoignages et récits d'histoires vécues postés par les internautes sur différents forums de discussion (aufeminin.com, doctissimo.fr, etc.). Les catégorisations sont multicritères : thématiques, tonalité, conseil *vs* demande, sexe de l'émetteur, situation familiale, etc. Nous traitons, dans des textes portant sur la santé, deux tonalités naïves mais qui correspondent à une attente en termes de filtrage de la part de l'agrégateur : « joyeuse » et

[3] Le *token* est souvent caractérisé comme une suite de signes alphanumériques séparés par des espaces ou des signes de ponctuation. Souvent, il ne s'agit pas même d'un mot ou d'une unité lexicale ayant un signe et signifié mais d'un segment identifiable selon les seuls critères informatiques.

« triste ». De prime abord, elle s'apparente à une analyse thymique, au sens de (Charaudeau, 1992), mais notre hypothèse est qu'il s'agit de catégories complexes où les phénomènes textuels, tels que la structure du récit, sont davantage encore que l'expression linguistique des sentiments, les véritables critères de classification.

3.2. Annotation tonale du corpus — L'annotation tonale du corpus, c'est-à-dire la classification initiale des textes, a été effectuée par SAMESTORY. Nous en avons analysé un échantillon pour en déduire la stratégie d'annotation de façon à caractériser plus finement l'opposition tonale joyeux/triste : un témoignage « triste » est (i) une histoire qui finit mal, (ii) un témoignage exprimant des doutes, des interrogations, ou sollicitant de l'aide. Un témoignage « joyeux » est (i) une histoire triste qui finit bien, (ii) un témoignage modulant la gravité d'une situation en en soulignant les points positifs (iii) un conseil. Voici deux exemples de témoignage, le premier, « triste », le deuxième « joyeux » :

> Témoignage « triste » : « Mauvaise nouvelle pour le début d'année ! — Je viens à vous car je souhaiterais quelques répones à mes questions. Voilà pour commencer cela faisait 1 voire 2 mois que mon père se plaignait de la gorge et plus ça allait plus ça empirer au point qu'il ne pouvait plus parler. Au début on a cru que ça venait du rhume qu'il avait eu, après des dents ou des oreilles. Et c'est quand il est allé chez l'ORL qu'il a apprit qu'il avait un cancer de la gorge. Les médecins vont lui prélever un échantillon. Je m'inquiète. J'habite loin de mes parents et je ne peux pas être auprès d'eux pour les épauler. Je voudrais savoir si ce genre de cancer ce soigne bien si on peut en réchapper et quel est le traitement et si il est lourd. Merci de vos réponses. Je ne vois pas vers qui je peux me tourner car je ne veux pas en parler avec mes parents ».[4]
>
> Témoignage « joyeux » : « Parkinson juvénile : 8 ans sous agoniste dopaminergique — J'ai 40 ans et mon diagnostique date de 2002, donc ça fait huit ans que je suis sous agoniste dopaminergique, mes doses vu mon "jeune age" sont tres élevees c a d 10 mg de siffrol (plus du double du max) et 40mg de requip ceci pour eviter d'avoir recours a la levodopa trop tot. Je me demene aussi en ce moment contre des douleurs dans le bras principalement le bras droit, mon neuro m'a envoyer chez un rumathologue qui m'a prescrit un antidepresseur qui est censé detendre mes muscles. Ca marche un peu pour les muscles mais surtout je dors enfin normalement, par contre je n'aime pas l'effet de ce medicament et en plus il m'empeche d'ejaculer ... ceci dit moi aussi j'ai un tres bon neuro tres a l'ecoute je vis a geneve. Mais j'avoue que des fois y en a ras le bol, c une maladie viscieuse qui ne cesse de marteler ses coups bas mais je garde le moral ! Ah oui je souffre d'une forme rare parait il ... parkinson juvenile c a d hereditaire ».[5]

4. Présentation de l'expérience

4.1. Élaboration textométrique des critères de catégorisation — Nous tentons de mettre en évidence les phénomènes textuels qui différencient les témoignages de nos deux catégories. Nous avons une double ambition : trouver des critères de classification linguistiquement explicables et suffisamment robustes pour servir comme descripteurs aux méthodes d'apprentissage supervisé. Nous faisons l'hypothèse que les critères de classification *interprétables* sont plus performants que les descripteurs trouvés par des méthodes d'apprentissage, souvent non signifiants d'un point de vue textuel et incidents au corpus d'apprentissage (ex : présence de fautes d'orthographe non pertinentes par rapport aux catégories de classification). Ainsi, lors de l'étape de sélection de critères, l'analyste écarte les critères liés à l'échantillon du corpus et choisit les critères textuels cohérents avec les composantes sémantiques (thématique, dialogique, etc.) actualisées dans le corpus (Rastier

---

[4] http://www.same-story.com/sante-maladies/cancer/cancers-orl/mauvaise-nouvelle-pour-le-debut-d-annee-214393

[5] http://www.same-story.com/sante-maladies/maladies-neurologiques/parkinson/parkinson-juvenile-8-ans-sous-agoniste-dopaminergique-112381b

2001). Pour l'expérience, nous avons utilisé trois types de critères : (i) unités isolées : un choix de formes, lemmes ou catégories morphosyntaxiques ; (ii) collocations de taille variée (de 2 à 4 unités) ; (iii) cooccurrences phrastiques multiniveaux (combinant les éléments de différents niveaux de description linguistique : formes, lemmes ou catégories morphosyntaxiques). Tous les critères sont sélectionnés selon 4 principes : leur caractère spécifique à un sous-corpus, leur répartition uniforme dans le sous-corpus, leur fréquence et leur pertinence linguistique.

L'analyse du corpus et l'extraction des critères a été effectué avec deux logiciels textométriques – Lexico3 (Salem *et al.* 2003)[6] et TXM (Heiden *et al.* 2010)[7] – qui implémentent les algorithmes de spécificités (Lafon, 1980) et de cooccurrences (Lafon, 1981). (Eensoo et Valette 2012) détaillent la méthode utilisée pour l'analyse du corpus.

4.2. Distribution des critères de catégorisation en composantes sémantiques — Nous avons ainsi identifié et inventorié 70 critères sémantiques à partir de l'analyse textométrique puis nous les avons caractérisés en fonction des composantes sémantiques. Il en résulte la construction de deux acteurs types, deux *agonistes* que nous avons nommés l'*internaute joyeux* et l'*internaute triste*. Une structuration en composantes sémantiques pourrait paraître un peu forcée par endroit, mais elle doit nous permettre d'évaluer chaque catégorie de critères, autrement dit chaque composante sémantique, au moyen d'une méthode d'apprentissage automatique de manière à connaître les plus performante dans une tâche d'analyse des sentiments.

4.2.1. L'agoniste triste.

L'agoniste triste est construit sur la noyau sémique :

/inaccompli/ + /dysphorique/

S'y agrègent des sèmes relatifs aux différentes composantes sémantiques. D'un point de vue dialogique, l'acteur-énonciateur apparaît

- égocentré (*PPERSI*[8]) et enclos sur son univers intime (+/ego/) : « *j'*ai retenu *mes* larmes pendant 1secondes ... et après tout a explosé ... non *ma* vie est foutue PQ *MOI* ... *JE* SUIS TROP JEUNE... »)
- exprimant un univers impressif et non factuel (« *Je ne sais pas* comment cela va évoluer » ; « quand je suis debout *j'ai l'impression* de tanguer ».

Du point de vu dialectique (c'est à dire de la représentation du temps et du déroulement aspectuel, des rôles et des interactions entre acteurs), on observe une excentration de l'action (+/passivité/) : « *On me dit* que les causes de cette maladie ne sont pas encore précises » ; « Le médecin *me dit que* ça doit être le fibrome et préfère attendre l'écho » ; « J'ai besoin *d'en parler* car les personnes autour de moi ne comprennent pas ! » ; « ça m'affecte et je ne pourrais sans doute jamais leur *parler de* ça ».

Nous avons essayé de limiter au mieux le recours à des critères issus de la composante thématique pour privilégier les critères pérennes susceptibles d'être exploiter sur d'autres corpus. Nous avons toutefois étudiés ceux relevant du domaine choisi pour cette étude (//domaine médical//), ils se révèlent intéressants pour caractériser l'opposition des agonistes joyeux et triste. Les isotopies et les formes sémantiques relèvent donc des taxèmes de ce domaine :

---

[6] http://www.tal.univ-paris3.fr/lexico/

[7] http://textometrie.ens-lyon.fr

[8] Désormais, tous les éléments en italique sont des exemples de critères de catégorisation.

- //médecine// : ‘médecin’, rendez-vous’, ‘être atteint de’ (« je viens d’apprendre que je *suis atteinte d’*un cancer du sein ») ;
- //hôpital// : ‘hôpital’, ‘urgences’, ‘analyses’ ;
- //diagnostic// : ‘syndrome’, ‘kg’ (« je suis pratiquement impuissant devant ce *syndrome* » ; « elle a perdu plus de 40 *kg* en 6 mois ») :
- //prescription// : ‘mg’, ‘chimio’ (« depuis février il prend 12.5 *mg* de cortancyl »).

Des critères relevant de la classe thymique ont également été identifiés : ‘peur’, ‘souffrir’, ‘stress’ : « j’ai *peur* que la douleur me revienne », « je *souffre* de porter la maladie de mon fils depuis ma grossesse ».

4.2.2. L’agoniste joyeux.

L’agoniste joyeux est élaboré sur le noyau sémique inverse de l’agoniste triste :

/accompli/ + /euphorique/

À l’inverse du triste, le joyeux est du point de vue de la composante dialogique, un acteur-énonciateur altruiste qui s’adresse à un tiers. La deuxième personne du singulier est en effet très présente dans le corpus (*PPERS2*) : « Alors *tu* vois il faut avoir espoir ». Le joyeux construit des univers alternatifs (i) en faisant part de son expérience à des fins d’édification ‘mon expériences’, ‘pour ma part’ : « Je tenais à faire part de *mon expérience* », « je me sens très *concernée* », « *pour ma part*, tous c’est très bien déroulé » ; (ii) en intertextualisant son témoignage, observable par la présence de lien ‘http’ : « J’ai trouver une photo sur un site pour les curieux(ses) *http://www*.lokaterre.com/… », « Je te file une adresse : *http://www*.linternaute.com/sante… J'en ai encore plein d'autres si tu veux qui peuvent éclairer tes questions »[9].

Le caractère le plus remarquable des textes joyeux réside au niveau de la composante dialectique. A la différence de l’agoniste triste, l’agoniste joyeux élabore un texte séquencé, descriptif ou argumentatif : ‘par contre’, ‘après’, ‘puis’ : « *Par contre* j’étais soignée à l’homéopathie », « J’ai choisi la deuxième solution, *après en avoir* discuté avec mon ami », « J’avais déjà quelques éruptions qui ont débuté *après avoir* pris la décision de déménager », « Après tu t’installes *puis* elle va te préparer la grosse piqure mdr ».

Au niveau de la composante thématique le taxème //médical// est actualisé par des critères issus des médecines douces (« l’*homéopathie*, ça marchait apparemment bien ») et celui des //traitements// comprend des hyperonymes lâches tels que ‘produit’ ou surtout ‘truc’ : « il m'a dit oralement des noms chimiques, du genre "Une injection de trichlobidule de broto*truc* de cétamol" », « Le *truc* pas mal c’est de ne pas attendre que le gel ne sèche totalement », « elle me file un *truc* genre doliprane ».

4.2.3. Évaluation des critères

Au total, 70 critères ont été construits : 30 critères relevant de la composante dialectique (interaction entre les acteurs, rôles) ; 16 relevant de la composante dialogique ; 17 relevant de la composante thématique et 6 critères thymiques. Un échantillon des critères avec les exemples se trouve dans l'annexe.

L’évaluation de la capacité classificatrice des critères qualifiés dans le paragraphe précédent, a été réalisée au moyen d’une classification de textes effectuée en utilisant un algorithme d'apprentissage automatique de la famille des *Machines à vecteurs de support* –

[9] Notons que le poids des liens hypertextuels est accentué par la présence de message crypto-commerciaux fréquent dans les forums de discussion.

SMO (Platt, 1998) implémenté sur WEKA (cf. *supra*). La description de l'algorithme n'étant pas l'objet de cet article, nous ne présentons que succinctement le fonctionnement général des algorithmes d'apprentissage supervisé en classification des textes : ils prennent en entrée un ensemble de valeurs numériques qui caractérise différents aspects des textes, ici l'ensemble de nos critères-descripteurs. À partir d'un corpus d'apprentissage déjà classé, ils construisent un modèle, c'est-à-dire une configuration optimale des caractéristiques données en entrée pour respecter le classement initial. L'évaluation de la pertinence du modèle se fait par l'application de ce dernier au corpus de test (qui est également déjà classé mais pour lequel le classement n'est pas visible pour l'algorithme). Le résultat de la classification est comparé au classement initial et plusieurs mesures d'évaluation (le plus couramment la précision, le rappel et la f-mesure) permettent de déterminer l'adéquation du modèle aux nouveaux textes.

Le tableau 1 donne à voir les résultats de la classification. Notre ligne de comparaison (*baseline*) est la classification sur formes simples, qui permet d'obtenir un taux de bon classement de 68,10 %. Il apparaît que les seuls critères thymiques, pourtant privilégiés dans les applications d'analyse des sentiments, donnent des résultats à peine supérieurs au hasard[10] (56,60 %). Cependant on doit admettre qu'ils sont ici très peu nombreux. Les 17 critères thématiques ne donnent pas des résultats exceptionnels, mais ce serait plutôt une bonne nouvelle dans la mesure où nous cherchions précisément à nous en affranchir. C'est le cumul des 45 critères dialectiques et dialogiques qui nous permet de nous élever significativement au dessus de notre ligne de comparaison (77,07%). Ce résultat est particulièrement intéressant car ce sont ces composantes qui se démarquent le plus nettement des pratiques en fouille de textes qui, en général, privilégient des descripteurs thématiques ou thymiques. Enfin, la totalité de nos 70 critères issus d'une analyse textométrique permettent d'atteindre une classification réussie à hauteur de 84,05%, soit 16 points de plus que la ligne de comparaison, ce que l'on peut considérer comme un résultat encourageant.

Types	**% f-mesure**	**Nb de critères**
Mots simples	68,10	10 700
Thymique (/dysphorique/)	56,80	6
Thématiques (//médical//)	61,46	17
Dialogique	63,80	16
Dialectique	73,09	30
Dialectique + Dialogique	77,07	46
Tous les critères	84,05	70

***Tableau 1** : résultat de la classification*

Conclusion

Il est admis que les méthodes efficaces en classification thématique (par exemple, l'apprentissage supervisé sur mots simples) sont peu performantes pour les tâches d'analyse de la subjectivité. La difficulté réside dans le fait que la subjectivité ne relève pas seulement du lexique, mais d'autres niveaux de description : organisation temporelle du récit, structure argumentative, etc. Nous avons proposé ici quelques éléments d'analyse inspirés de la sémantique interprétative pour la prise en compte de ces niveaux de description et leur implémentation pour la classification. Notre caractérisation différentielle des « sentiments » sur les forums de discussion nous amener à stéréotyper deux agonistes : le triste, qui se

[10] Étant donné que l'on a deux classes de la taille comparable, le hasard est un taux de bon classement à 50%

caractérise par les traits /inaccompli/, /impuissance/ et la clôture des univers ; et le joyeux qui se caractérise par les traits /accompli/, /interaction/ et la construction d'univers multiples.

Le coût en temps de notre méthode d'élaboration de critères n'a pas été quantifié mais en première approximation, il apparaît comparable à d'autres méthodes semi-automatiques. Le domaine manquant de méthodes éprouvées, cette étude permet de mieux comprendre la tâche et sa complexité et d'esquisser une proposition méthodologique tenant compte d'une caractérisation textuelle de la subjectivité.

Remerciements

Nous avons plaisir à remercier Evelyne Bourion et Monique Slodzian pour leurs leur aide lors de la rédaction de cet article.

Références

BENOIT, K.; LAVER, M. & MIKHAYLOV, S**.** (2009). Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions , *American Journal of Political Science,* Blackwell Publishing Inc, *, 53*, 495-513

CHARAUDEAU P. (1992). *Grammaire du sens et de l'expression*. Hachette Education.

EENSOO, E., VALETTE, M. (2012) « Sur l'application de méthodes textométriques à la construction de critères de classification en analyse des sentiments », Georges Antoniadis, Hervé Blanchon, Gilles Sérasset, éd., Actes de la conférence conjointe JEP-TALN-RECITAL 2012, Volume 2: TALN, 4-8 juin 2012, Grenoble, pp. 367-374

ESULI, A. & SEBASTIANI, F. (2006). SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining, *Proceedings of the 5th conference on International Language Resources and Evaluation (LREC'06).*

GARDIN, P. (2009). Application de la théorie de l'Appraisal à l'analyse d'opinion , *MajecSTIC.*

GHORBEL, H. & JACOT, D. (2011). Mugellini, E.; Szczepaniak, P.; Pettenati, M. & Sokhn, M. (Eds.) Further Experiments in Sentiment Analysis of French Movie Reviews, Advances in Intelligent Web Mastering – 3, Springer Berlin/Heidelberg, , 86, pp.19-28

HATZIVASSILOGLOU, V. & WIEBE, J. (2000). Effects of adjective orientation and gradability on sentence subjectivity , *Proceedings of the International Conference on Computational Linguistics (COLING).*

HEIDEN, S., MAGUE, J-P. et PINCEMIN, B. (2010). TXM : Une plateforme logicielle open-source pour la textométrie – conception et développement. In I. C. Sergio Bolasco (Ed.),*JADT 2010*, Vol. 2, pages 1021-1032. [logiciel disponible sur http://textometrie.ens-lyon.fr/]

HOPKINS, D. ET KING, G. (2007). Extracting systematic social science meaning from text. Disponible sur l'internet : http://gking.harvard.edu/files/words.pdf

KIM, S. M.; VALITUTTI, A. & CALVO, R. A. (2010). Evaluation of unsupervised emotion models to textual affect recognition, *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text,* Association for Computational Linguistics, , pp.62-70

LAFON, P. (1980). Sur la variabilité de la fréquence des formes dans un corpus, *Mots,* 1, pages 127-165.

LAFON, P. (1981). Analyse lexicométrique et recherche des cooccurrences, *Mots, 3*, pages 95-148.

LISCOMBE, J.; RICCARDI, G. & HAKKANI-TÜR, D. (2005). Using context to improve emotion detection in spoken dialog systems , *In Proceeding of Interpeech*, pp.1845-1848

MAUREL, S. ET DINI, L. (2009). Exploration de corpus pour l'analyse des sentiments , *Actes de DEFT'09 « DÉfi Fouille de Textes », Atelier de clôture.*

PANG, B., LEE, L. et VAITHYANATHAN, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79-86.

PINCEMINE, B. (2010) - « Semántica interpretativa y textometría », in Duteil-Mougel Carine & Cárdenas Viviana (éds), Semántica e interpretación, Tópicos del Seminario, 23, Enero-junio 2010, pp. 15-55. (ISSN 1665-1200 ;

trad. Sebastián Giorgi).

PLATT, J. (1998). Machines using Sequential Minimal Optimization. B. Schoelkopf, C. Burges and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning.*

SALEM A., LAMALLE C., MARTINEZ W., FLEURY S., FRACCHIOLLA B., *et al.* (2003). Lexico3 – Outils de statistique textuelle. Manuel d'utilisation. http://www.tal.univ-paris3.fr/lexico/]

SPERTUS, E. (1997). Smokey: Automatic Recognition of Hostile Messages , In *Proc. IAAI*, pp. 1058-1065

TANGUY, L. (2012) *Complexification des données et des techniques en linguistique : contributions du TAL aux solutions et aux problèmes*, Mémoire d'habilitation à diriger des recherches, Université Toulouse-Le Mirail, Toulouse.

STRAPPARAVA, C. & VALITUTTI, A. (2004). WordNet-Affect: an Affective Extension of WordNet , *In Proceedings of the 4th International Conference on Language Resources and Evaluation,* , pp.1083-1086

TURNEY, P. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. *Proceedings of the Association for Computational Linguistics (ACL)*, pages 417-424.

VERNIER, M., MONCEAUX, l. et DAILLE, B. (2009a). DEFT'09 : détection de la subjectivité et catégorisation de textes subjectifs par une approche mixte symbolique et statistique *Actes de l'atelier de clôture de la 5ème édition du Défi Fouille de Textes.*

VERNIER, M.; MONCEAUX, L.; DAILLE, B. & DUBREIL, E. (2009b). Catégorisation des évaluations dans un corpus de blogs multi-domaine *, Revue des nouvelles technologies de l'information (RNTI), pp.45-70*

WHITELAW, C.; GARG, N. & ARGAMON, S. *(*2005). ACM (Ed.) Using appraisal groups for sentiment analysis , *Proceedings of the 14th ACM international conference on Information and knowledge management, pp.*625-631

YI, J.; NASUKAWA, T.; BUNESCU, R. & NIBLACK, W. (2003). Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques , *Proceedings of the Third IEEE International Conference on Data Mining,* IEEE Computer Society.

YU, H. & HATZIVASSILOGLOU, V. (2003). Towards answering opinion questions: separating facts from opinions and identifying the polarity of opinion sentences , *Proceedings of the 2003 conference on Empirical methods in natural language processing,* Association for Computational Linguistics, , 129-136

YVON, F. (2006). *Des apprentis pour le traitement automatique des langues*. Mémoire d'habilitation à diriger des recherches, Université Pierre et Marie Curie, Paris.

Annexe A

		Triste	Joyeux
Composantes textuelles	Dialectique Organisation linéaire et temporelle du récit) « Rend compte des acteurs et des interactions entre acteurs (= rôles) »	***ne sais*** ***sais pas*** ***en parler*** ***parler de*** ***en plus*** ***me dire que*** ***arriver pas à*** ***du mal à*** (« j'ai du mal à arrêter le stilnox, elle avait du mal à respirer ») ***je avoir prendre*** ***pour me*** ***voilà*** ***avoir l'impression***	***dès*** ***Par contre*** ***, car*** ***, après*** ***après avoir*** ***puis*** ***par exemple*** ***etc*** ***essayer*** ***au moins*** (« j'ai au moins mon amour qui est là pour moi ») **Modalités conclusives** ***Bon*** (« Bon t'inquiète pas ») ***ça marche*** **Modalité irénique** ***faut pas*** ***ne faut*** ***il s'agit de*** **n'ai plus** (« je n'ai plus mal à mon bras gauche, je n'ai plus le moindre problème de dents ») **si PPERS2S**
	Dialogique (positionnement des auteurs) Les univers dans lesquels les acteurs Possible	**Genre appel à l'aide** ***PPER1ST*** ***je*** ***?***	**Genre conseil** ***PPER2SG*** ***tu*** ***!***

	Contrefactuel (faux) Factuel (triste :maladie) (gai :guérie)	***mon mari*** ***pper3s pper1s*** ***remercier*** «	***merci*** * oublié :allerpper2sG **mon expérience** concerner (« je me sens très concernée parce que mon fils, sans être autiste, etc. ») (« en ce qui concerne les hommes ») ***[Pp]our ma part*** (« Pour ma part, j'ai eu un cancer de l'estomac, »)(« pour ma part, tous c'est très bien déroulé »)
	Thématique	**<u>Domaine médical</u>** ***urgences*** ***hôpital*** ***analyse***	**<u>Domaine médical</u>**
		médecin ***rendez-vous*** (s ?) [analyses a l'air plus polarisé)	
		syndrome(s) ***kg*** (« elle a perdu plus de 40kg en 6 mois ») ***mg***	***dentiste*** ***visage***
		<u>Taxème des traitements</u> ***chimio***	***rémission*** ***truc*** (« un truc genre doliprane », « un truc de grand-mère ») ***homéopathie*** **produit naturel** **naturel** (« les huiles essentielles naturelles ») (« les moyens de guérir les maux rhumatismaux d'une façon naturelle »)

			produit(s) (« ce produit lave vraiment bien la peau ») (« c'est un produit contre l'acné qui coûte environ 8 $ ») ***utiliser*** (« j'utilise toujours le savon Avène ») (« une nouvelle sorte d'antalgique est utilisée directement sur les nerfs »)