Représentation vectorielle de textes courts d'opinions
Analyse de traitements sémantiques pour la fouille d'opinions par clustering

Benoît Trouvilliez[1, 2]
(1) Centre de Recherche en Informatique de Lens, Université d'Artois, Rue Jean Souvraz, 62300 Lens, France
(2) Onyme SARL, 165 Avenue de Bretagne, 59000 Lille, France
btrouvilliez@onyme.com, benoit.trouvilliez@gmail.com

Résumé. Avec le développement d'internet et des sites d'échanges (forums, blogs, sondages en ligne, ...), l'exploitation de nouvelles sources d'informations dans le but d'en extraire des opinions sur des sujets précis (film, commerce,...) devient possible. Dans ce papier, nous présentons une approche de fouille d'opinions à partir de textes courts. Nous expliquons notamment en quoi notre choix d'utilisation de regroupements autour des idées exprimées nous a conduit à opter pour une représentation implicite telle que la représentation vectorielle. Nous voyons également les différents traitements sémantiques intégrés à notre chaîne de traitement (traitement de la négation, lemmatisation, stemmatisation, synonymie ou même polysémie des mots) et discutons leur impact sur la qualité des regroupements obtenus.

Abstract. With the internet and sharing web sites developement (forums, blogs, online surveys, ...), new data source exploitation in order to extract opinions about various subjects (film, business, ...) becomes possible. In this paper, we show an opinion mining approach from short texts. We explain how our choice of using opinions clustering have conducted us to use an implicit representation like vectorial representation. We present different semantic process that we have incorporated into our process chain (negation process, lemmatisation, stemmatisation, synonymy or polysemy) and we discut their impact on the cluster quality.

Mots-clés : représentation des textes, représentation vectorielle, traitement de textes courts, regroupements d'opinions.

Keywords: text representation, vectorial representation, short text processing, opinion clustering.

Introduction

Avec le développement d'internet, on observe la croissance de nouvelles sources d'informations reflétant des opinions sur des sujets variés : un film, une actualité, les valeurs d'une entreprise, les prestations d'un commerçant, un site internet, ... Ces sources prennent des formes diverses telles que des sites communautaires (forums, blogs) ou des sites de sondages en ligne. Les textes sont assez courts et dépassent rarement 4 phrases. Dans ce contexte, plusieurs stratégies d'extractions d'opinions («fouilles d'opinions» ou «*Opinion Mining*») ont vu le jour. Alors que certains travaux visent à une analyse de sentiments afin de déterminer si les auteurs sont plutôt favorables ou défavorables au sujet en question (Poirier *et al.*, 2008), nous nous intéressons à la problématique du regroupement afin d'extraire les principales idées développées. Nous faisons appel à deux techniques différentes : des méthodes d'analyse sémantique qui extraient l'information des textes dans un modèle («représentation de textes») et des méthodes de regroupements non supervisés («*clusterings*») qui transforment cette information en groupes d'opinions (Fig. 1).

1 Représentation et traitements sémantiques

1.1 Généralités et constats

Les textes que nous avons à traiter ont une taille moyenne de deux phrases excèdant rarement 5 mots chacune. Ils reflètent des opinions d'auteurs différents, n'utilisant pas le même niveau de registre de langue et n'écrivant pas non plus dans un langage fortement rigoureux. Deux constats peuvent être faits. D'une part, la taille des représentations est petite. Chaque texte n'exprime que peu d'idées / opinions, rarement répétées dans un même texte. D'autre part, il n'est pas rare de trouver des fautes d'orthographe ou des abréviations. Si celles-ci portent sur l'opinion du message, il est peu probable que l'on puisse récupérer l'information complète a posteriori d'après le premier constat.

1.2 Les représentations

Deux grands types de représentations de la sémantique d'un texte existent. Les «représentations explicites», dont les graphes sémantiques sont un exemple (Rastier, 1989) (Sowa, 1984), prennent en compte les liens sémantiques existants entre les mots du texte(«sémantique des mots»). Elles permettent d'obtenir une précision importante mais ne sont pas aisées à construire. Les «représentations implicites» représentent la sémantique en utilisant un ensemble de variables booléennes. Elles ne permettent pas de représenter la sémantique des mots mais offrent une métrique simple de comparaison entre les représentations («distance sémantique») par comparaisons booléennes. Cette distance peut ensuite être utilisée par les algorithmes de regroupements pour rapprocher les données. La représentation vectorielle standard est un exemple de représentation implicite : elle utilise comme variables booléennes, la présence ou l'absence des mots des textes («vecteurs de mots» (Salton *et al.*, 1975)). Cette représentation a fait l'objet de nombreuses études et est souvent utilisée en corrélation avec des méthodes statistiques tels que *Hyperspace Analogue to Language* (HAL) (Lund *et al.*, 1995) ou *Latent Semantic Analysis* (LSA) (Deerwester *et al.*, 1990) pour permettre la représentation de la sémantique des mots. Il est également courant d'utiliser des méthodes de pondération afin de mesurer la quantité d'information apportée au texte par chaque mot («pertinence des mots»). Cette information va nous servir à identifier les mots les plus propices aux rapprochements.

L'utilisation de tous les mots des textes engendre une taille de représentation considérable, difficile à traiter. Il est courant d'effectuer des traitements sémantiques visant à la simplifier en regroupant ou supprimant certaines composantes des vecteurs. Cette simplification des descripteurs a un impact sur la qualité finale du traitement : si les textes ne sont pas correctement représentés, il est peu vraisemblable que le résultat obtenu soit probant car la représentation ne reflétera pas le sens du texte. Certains chercheurs préconisent de ne pas systématiquement employer ces méthodes de simplification mais au contraire de réfléchir à leur pertinence et au fait que des mots supprimés traditionnellement des représentations peuvent être primordiaux pour des traitements proches de la sémantique (Riloff, 1995).

1.3 Pondération et distance sémantique des mots

La pondération vise à accorder plus d'importance à certaines variables booléennes, à en pénaliser d'autres en allant jusqu'à les retirer (équivalent à une pondération nulle) et s'oppose à l'équi-importance qui considère que tous les mots apportent la même quantité d'information. Utiliser une pondération est particulièrement intéressant en *clustering* pour distinguer les mots véhiculant beaucoup d'informations, de ceux n'en apportant que peu. Deux types d'approches existent. Les premières, méthodes statistiques, se basent sur les occurrences des mots dans le corpus pour déterminer l'importance des mots en contexte. La formule "*term frequency, inverse document frequency*" (TF.IDF) (Sparck Jones, 1972) en est un exemple. Dans le cadre de regroupements, un mot qui apparaît dans la majorité des textes n'est pas assez discriminant et conduit à rapprocher trop de textes. Un mot apparaissant très rarement conduit à en rapprocher trop peu sur des idées peu représentatives. Les deuxièmes, méthodes linguistiques, consistent à établir une liste des mots à éliminer, considérée comme peu évolutive et de taille raisonnable (Salton *et al.*, 1975) (Salem, 1987). Le contenu et la longueur de cette liste («*black list*», «*stop list*» ou «liste noire») sont différents en fonction de la nature du traitement. Quelle que soit l'approche retenue, la suppression à ce stade d'un descripteur important pour le corpus est définitive et provoque une dégradation sensible des résultats.

Dans notre contexte, les opinions étant rarement répétées, il est important de faire attention à ne pas supprimer de mots utiles pour éviter les pertes d'informations. Utiliser une liste noire pour supprimer les mots inutiles permet le contrôle des mots retirés sous réserve de ne pas y inclure de mots qui pourraient être informatifs. Considérer les autres mots comme équi-importants ne semble pas être une solution efficace car ceux qui véhiculent le plus par leur présence le sens global des textes sont plus importants que les autres. Utiliser une technique statistique de pondération semble donc judicieux. La formule du TF.IDF n'est pas appropriée car sa pertinence repose sur l'hypothèse de répétition des sens importants dans un texte, peu vérifiée si les textes sont très courts. De même, l'IDF seul privilégie les mots très rares. Utile dans un cadre de recherche de termes discriminants, cela a pour effet dans notre cas de provoquer des rapprochements sur des mots issus du niveau de langage du répondant plus que sur les idées exprimées. On peut par contre émettre l'hypothèse que l'importance des mots pour les textes est équivalente à l'importance des mots pour l'ensemble des textes. Nous proposons d'utiliser l'entropie de Shannon appliquée sur l'ensemble des

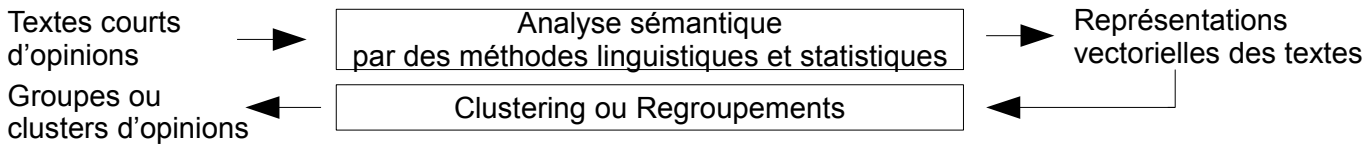


FIG. 1 – Notre processus de traitement

documents (formule 1) pour que les mots ayant une probabilité d'apparition moyenne dans l'ensemble du corpus soient privilégiés.

$$I_{mot} = -(P_{mot} * log(P_{mot})) \text{ , avec } P_{mot} = \frac{Occ_{mot}}{Nb_{mots}} \quad (1)$$

Dans (1), I_{mot}, P_{mot} et Occ_{mot} représentent l'importance, la probabilité d'apparition et l'occurrence du mot dans le corpus, et Nb_{mots}, le nombre de mots au total dans le corpus. Deux mots sémantiquement proches devraient obtenir des notes qui reflètent leurs liens. Pour cela, il est courant d'associer à la représentation vectorielle des méthodes permettant d'identifier les liens sémantiques entre les mots du corpus. Une fois encore, les méthodes sont soit statistiques (HAL, LSA) soit linguistiques (utilisation de ressources tels qu'un thésaurus ou une ontologie). Nous proposons d'utiliser une méthode linguistique basée sur une ontologie. La plus connue est le Wordnet de Princeton (Miller, 1995), créé pour la langue anglaise et fondateur des «Wordnets» : ressources aux caractéristiques similaires, développées pour plus de 70 langues et listées par la «Global Wordnet Association»[1]. Chaque unité de sens de la langue est représentée par un groupe de mots porteur de ce sens appelé synset. Des liens sémantiques tels que l'antonymie, l'hypéronymie ou l'hyponymie sont définis entre les synsets. Pour le français, on trouve le Wordnet Libre du Français (WOLF) (Sagot & Fišer, 2008), utilisé pour nos tests, et le projet francophone EuroWordnet (Vossen, 1998). Le problème majeur est alors la polysémie des mots car si l'ontologie fournit l'ensemble des sens du mot et les liens existants pour chacun d'eux, il est difficile de déterminer lequel est employé. Dans le cadre de textes longs, la répétition du sens employé dans le texte peut aider à l'identifier, fait non vérifié dans des textes courts. Nous avons donc opté pour une méthode qui calcule la probabilité d'apparition d'un mot t comme la somme des probabilités d'apparition des mots i partageant un de leurs sens en commun avec le mot t (formule 2).

$$I_{mot/sens} = -(P_{mot/sens} * log(P_{mot/sens})) \text{ , avec } P_{mot/sens} = \sum_{i \in Syn_{mot}} P_i \text{ , et } Syn_{mot} = \bigcup_{mot \in s} s \quad (2)$$

Dans (2), $I_{mot/sens}$ et $P_{mot/sens}$ représentent l'importance et la probabilité d'apparition d'un mot selon la représentativité de ses sens dans le corpus, i, un mot de l'ontologie, s, un synset de l'ontologie et P_i, la probabilité d'apparition du mot i calculée selon la sous-formule P_{mot} de la formule (1). La formule 2 a pour propriétés essentielles de provoquer la maximisation des probabilités d'apparition des mots en considérant chaque mot potentiellement porteur du sens comme porteur et de ne pas requérir de désambiguisation au préalable des sens des mots. Cette stratégie, peu justifiée dans un contexte pluri thématique, l'est dans notre cas car la probabilité d'emploi polysémique d'un terme au sein des textes du corpus se trouve considérablement réduite en l'absence de thèmes multiples.

Le tableau 1 montre les mots identifiés comme les plus importants dans deux corpus de textes différents selon cette méthode, mots propices aux regroupements d'opinions (ouverture de magasins, progression du chiffre d'affaire, formation à l'anglais, favoriser la mobilité, développer les connaissances culturelles, ...). Les mots «améliorer» et «promouvoir» ont reçu une note identique grâce à la détection d'une proximité sémantique dans leurs sens. Nous expliciterons ensuite les effets sur le clustering de ces pondérations.

L'ontologie permet également de tenir compte de la distance sémantique entre les mots lors des regroupements. En l'absence de thèmes multiples, nous considérons que deux mots sont sémantiquement liés si l'ontologie connaît un sens à ces mots partageant un lien. Dans le cas, où plusieurs liens pourraient être trouvés, le lien le plus fort est retenu. A savoir, par ordre décroissant d'importance, la synonymie forte,

[1] référence : http ://www.globalwordnet.org

la synonymie faible, l'hypéronymie et la fratrie par hypéronymie commune (formule 3). Ce raisonnement possède les mêmes propriétés que celles évoquées pour la pondération : maximisation de la probabilité d'identification d'un lien sémantique et non désambiguisation du contexte.

$$D(i,j) = \min_{i \in s, j \in s'} (d(s,s')) \tag{3}$$

Dans (3), i et j représentent deux mots du corpus, $D(i,j)$, la distance sémantique entre i et j, s et s', deux synsets de l'ontologie distincts ou non et $d(s,s')$, la distance entre s et s' selon l'ontologie.

1.4 Représentation de la négation

La représentation de la négation est un problème à cause notamment des formes multiples qu'elle peut prendre, simples (ne...pas, ne...plus, ...) ou complexes (double négation souvent par antonymie), et de sa «portée» dans le texte (sur une ou plusieurs phrases ou parties de phrases). Cette liste n'est pas exhaustive mais illustre bien les problèmes rencontrés. Alors que les chercheurs travaillant sur l'identification de la thématique des textes n'effectuent aucun traitement particulier et se contentent de traiter les mots de la négation comme des mots ordinaires voire de les inclure à la liste noire, d'autres, travaillant sur la sémantique préconisent une différenciation entre les phrases selon que l'idée exprimée est niée ou non (Poirier *et al.*, 2008). Traiter la négation en particulier, nécessite de se confronter au problème de sa portée. Comme les textes sont très courts, nous pouvons poser l'hypothèse que la négation, si elle existe, peut être appliquée sur l'ensemble du texte sans en dégrader le sens (Poirier *et al.*, 2008). Cela n'est bien sûr pas vérifié dans des textes complexes (phrases avec conjonctions, ...).

Afin de valider ces considérations, des tests ont été réalisés sur un jeu d'essai comportant deux sortes de messages : des messages affirmant aimer les chats et des messages niant aimer les chats. Ils ont consisté à effectuer des regroupements en considérant successivement les marques de la négation comme appartenant à la liste noire, comme des mots ordinaires pour lesquels aucun traitement n'est à faire et enfin comme des mots discriminants permettant de marquer les messages avec une valeur affirmée ou niée. Dans ce

Résultat sur le jeu «réussite de l'enseigne TrucMuche»							
mot	I_{sens}	mot	I_{sens}	mot	I_{sens}	mot	I_{sens}
magasin/s	0.1535	international/e	0.1357	chiffre d affaire	0.1315	expansion	0.1248
ouverture/s	0.1519	préférée	0.1357	progresse	0.1315	com	0.1248
france/çais	0.1415	TrucMuche	0.1357	formation/s	0.1315	vendus/ues	0.1224
client/s/e/es	0.1396	augmentation/s	0.1336	développement/s	0.1294		
enseigne	0.1377	nombre/s	0.1315	volume/s	0.1271		
Résultat sur le jeu des «choix stratégiques»							
mot	I_{sens}	mot	I_{sens}	mot	I_{sens}	mot	I_{sens}
formation/s	0.1668	international/ale/ales/aux	0.1489	communication/s	0.1296	inter	0.1186
anglais/se	0.1591	échange/s	0.1441	connaissance/s	0.1254		
mobilité/s	0.1591	langue/s	0.1372	différents/tes/ces	0.1254		
favoriser	0.1563	sites	0.1372	personnel/s	0.1232		
développer/ement	0.1519	étranger/ers/ères	0.1354	améliorer	0.1186		
		culture/es/el/elle/els/elles	0.1316	promouvoir	0.1186		

TAB. 1 – Extraits de l'importance des mots avec l'entropie de Shannon appliquée sur les sens

dernier cas, les marques de la négation et antonymes sont retirés. Puis, les messages sont marqués comme affirmés, s'ils contenaient un nombre pair d'expressions de négation et antonymes, et comme niés sinon. Dans le cas où la négation est dans la liste noire, aucune distinction n'est faite entre les messages affirmés et niés. Dans le cas du non traitement de la négation, la distinction est faite uniquement sur la présence ou l'absence des mots négatifs. Cela engendre le regroupement à tort des messages contenant une double négation avec ceux n'en contenant qu'une simple : «Rien ne me fera aimer les chats» et «Rien ne me fera détester les chats». Dans le cas du marquage, la distinction se fait sur la polarité induite aux messages par la négation. Cette dernière distinction est la meilleure dans notre cas puiqu'elle permet d'obtenir deux clusters d'opinions opposées : «ceux qui aiment les chats» et «ceux qui n'aiment pas les chats».

1.5 Lemmatisation / stemmatisation

Ces méthodes identifient les mots différents à cause des règles syntaxiques et grammaticales mais ayant une sémantique identique. La lemmatisation identifie la fonction grammaticale et le lemme du mot, soit à l'aide du contexte au moyen d'une désambiguïsation, solution étudiée dans cet article, soit en établissant une liste des différents lemmes possibles du mot («lemmatisation en ou hors contexte»). La stemmatisation ne résout jamais le contexte mais identifie selon la forme et la langue utilisée, la fonction grammaticale du mot et en déduit son radical. La lemmatisation simplifie les représentations de mots dont le radical varie mais pas celles de ceux occupant une fonction grammaticale différente. La stemmatisation, au contraire, est capable d'effectuer des rapprochements entre les mots occupant une fonction grammaticale différente mais pas entre ceux dont le radical varie et provoque de plus un rapprochement non désiré entre les mots sémantiquement différents mais ayant une racine commune.

Nos choix se sont portés sur TreeTagger (Schmid, 1994), étiqueteur multilingue réalisé par le laboratoire de l'Université de Stuttgart, pour la lemmatisation et sur Snowball, sous projet de Apache Lucène[2], pour la stemmatisation. Ils ont été testés sur différents jeux de tests. Ils ne commettent pratiquement aucune erreur en contexte orthographique correct. TreeTagger semble même parvenir à désambiguïser correctement dans des phrases comportant des homonymes comme le verbe «porter» et le nom «porte». Sur un corpus mal orthographié, les résultats de lemmatisation s'effondrent. Aucun lemme correct n'a été identifié sur les termes inconnus et peu de fonctions grammaticales l'ont été. Du côté de la stemmatisation, les résultats sur un corpus mal orthographié sont mitigés. Les mots sur lesquels la faute porte sur la partie considérée comme le radical n'ont pas été corrigés. De même, certaines fautes portant sur le suffixe ont modifié le radical retourné par Snowball (exemple : «jouet» donne le radical «jouet» alors que «jouer» donne le radical «jou»).

Ces tests sur la lemmatisation et la stemmatisation montrent que la lemmatisation est peu efficace dans un contexte orthographique difficile. Cela est surtout un handicap si le terme affecté est un terme décisif pour les regroupements. La stemmatisation se révèle être une solution si la faute ne modifie pas le radical du mot. Ces constats nous ont conduits à supposer que la combinaison des deux méthodes permettrait de tirer parti des avantages de chacune des deux stratégies : la lemmatisation permet de regrouper dans un premier temps les mots dont le radical évolue et la stemmatisation, de regrouper ensuite les différentes formes grammaticales d'un même concept tout en corrigeant éventuellement quelques fautes d'orthographe. Les résultats intermédiaires obtenus après la lemmatisation sont conservés et utilisés lors des rapprochements sémantiques à l'aide de l'ontologie. Il serait également possible d'utiliser les statistiques ainsi obtenues

[2]http ://lucene.apache.org/

Traitements sémantiques				
Négation	**Lemmatisation**	**Stemmatisation**	**Pondération**	**Distance sémantique des mots**
Ne rien faire	Ne rien faire	Ne rien faire	Equi-pondération simple	Similitude de formes
Black List	TreeTagger	SnowBall	*Black List* + Equi-pondération	Similitude de sens
Tags des textes			*Black List* + TF.IDF	
			Black List + Shannon avec sens	
			Black List + Shannon	

TAB. 2 – Les différentes stratégies de traitements sémantiques

pour déterminer si deux mots proches après stemmatisation le sont à cause de leur sémantique.

Dans le tableau 1, «formation»/«formations», «culturel»/«culturelle» ont la même importance grâce à la lemmatisation, et «culture»/«culturelles», «expatriation»/«expatriés» grâce à la stemmatisation.

2 Résultats expérimentaux sur les regroupements

Le tableau 2 reprend les 120 combinaisons de traitements sémantiques évoquées. Parmi elles, celle soulignée dans le tableau s'est démarquée dans nos réflexions et tests par son adéquation avec nos attentes. Une analyse des regroupements obtenus au moyen d'une méthode hiérarchique dans cette configuration est présentée dans le tableau 3. Les tests, conduits sur méthode par partitions (KMeans), ont montré des problèmes sémantiques similaires à ceux évoqués ci après pour la méthode hiérarchique.

Dans le premier extrait, les idées de «réunion régulière», «proximité, convivialité», «fêter les succès» ont été détectées. L'expression «mise en place» a été considérée comme une opinion au lieu des compléments de cette expression. Dans le deuxième extrait, les idées «d'écouter et de voir» aussi bien les clients que d'autres métiers et de «voyage» ont été trouvées. Cet exemple montre l'intérêt d'employer des techniques tels que la lemmatisation ou la stemmatisation : le rapprochement sémantique a ainsi pu être fait entre le nom «voyage» et le verbe «voyager». Dans le troisième extrait, «projets internationaux» et «projets mondiaux» ont été rapprochés par la détection de liens sémantiques entre les termes «internationaux» et «mondiaux». Par ailleurs, cet extrait porte sur l'un des deux jeux utilisés pour tester la pondération. On retrouve bien les idées faisant l'objet de la plus forte pondération : formation à l'anglais, favoriser la mobilité et développement des connaissances culturelles. Dans le dernier extrait, notre traitement de la négation en particulier a permis de rapprocher le message «je n'ai pas eu de problèmes avec les produits» du groupe «satisfaction produit». En revanche, sur des textes plus complexes, notre technique engendre comme prévu, des erreurs de compréhension : le message «je n ai rien a dire je suis tres heureuse de ma commande» a été rapproché du groupe «insatisfait commande» à cause de la négation qu'il comprend («n»). «mercie a vous touts» a été bien classé, malgré l'orthographe, en combinant la lemmatisation et la stemmatisation, mais «je crois savoir ma la respnsable du pont relais...» ne l'a pas été.

Une analyse statistique a été menée sur le clustering produit par cette méthode hiérarchique sur un jeu de test créé à partir de plusieurs corpus. Cette qualité a été évaluée, sur l'ecart entre le nombre de groupes attendu et produit, et sur l'homogénéité de ces derniers. Chaque message est codifié manuellement selon l'idée principale qu'il exprime. L'homogénéité d'un cluster correspond au pourcentage de messages codifiés avec l'idée majoritaire du cluster. Si aucune idée n'est majoritaire, le cluster est déclaré comme totalement hétérogène. L'homogénéité globale est ensuite calculée comme la moyenne des homogénéités

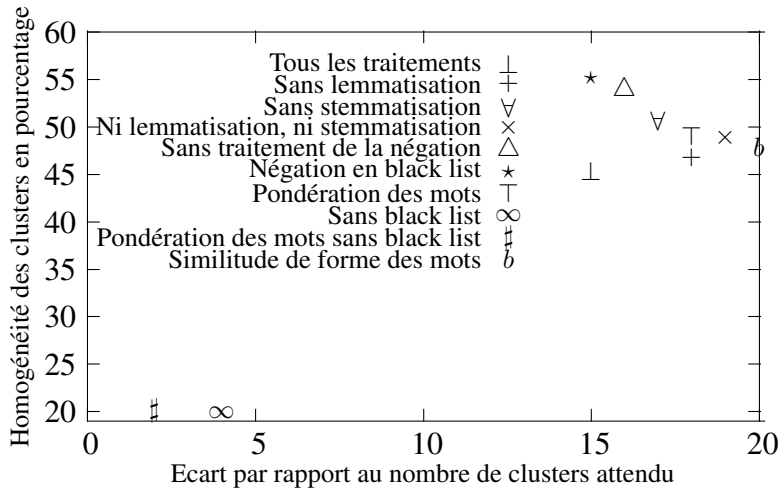
Premier extrait : jeu «entreprise»	
Titre du cluster	**Messages**
Proximité, convivialité	Développer (la proximité/la convivialité) (dans l'entreprise/avec et entre les collaborateurs)
Conventions	Convention semestrielle ou annuelle / Convention : réunion de tous les collaborateurs
Groupes de projets	(Travailler en/Créer des groupes de) projets transversaux / Favoriser les groupes de projets
Mettre en place	Mettre en place (des groupes de travail transversaux/des objectifs) / Solliciter des volontaires pour mettre en place des groupes de travail transversaux suivis et reconnus par le manager / Mise en place (d'outils de mesure précise de réalisation des objectifs/de réunions régulières d'information)
Fêter les succès	Savoir fêter les succès / Créer des événements réguliers et varies fêter les succès
Deuxième extrait : jeu «relationnel»	
Titre du cluster	**Messages**
Ecouter, observer	Je regarde et j'écoute mes clients / (Ecouter/Etre à l'écoute/Faire parler/Rêver pour) les clients, les fournisseurs / Rester connecté avec le client et la réalité du terrain / Ecouter son instinct / Je vais voir autre chose dans d'autres métiers
Moins d'opérationnel	(Pouvoir se dégager/Je me dégage) de l'opérationnel / Susciter un contre pouvoir non opérationnel
Veille	Se tenir informé du marché (visite clients, veille) / Je visite mes clients / S'informer, reste en veille
Voyager	Je voyage / Voyager, s'ouvrir au monde / Voyages en interne, échanges avec ses équipes
Troisième extrait : jeu «choix stratégiques»	
Titre du cluster	**Messages**
Projets internationaux	Faire des projets internationaux / Développer une culture mondiale à l'occasion de projet
Différences culturelles	(Sensibilisation/Sensibiliser le personnel) aux différences culturelles
Connaissance culturelle	Développer la connaissance des (cultures de nos clients/différentes "cultures business"/cultures locales/des cultures des autres pays/différents sites) / Se former à la culture
Ouverture international	(Développer les/Promouvoir l'exercice des) langues étrangères / Développer la maîtrise de la langue anglaise / Apprendre l'anglais / Favoriser l'internationalisation des recrutements
Mobilité	Développer (la mobilité/la mobilité internationale (dans les 2 sens))
Quatrième extrait : jeu «commande internet»	
Titre du cluster	**Messages**
Satisfait produits	comme d'habitude je n'ai pas eu de problème avec les produits / J'ai été ravie par ma commande et surtout la rapidité et le sérieux de l'envoi. / je suis contente de vos colis / Rien à dire tout est parfait ! ! / Frais de livraison un peu élevés je trouve. Les produits sont conformes et de bonne qualité pour le prix. / Je suis très contente / mercie a vous touts
Insatisfait commande	commande soit disant livrée en 48 heures mais livraison seulement après relance le 24/12 ça fait un peu long heureusement qu'il n'y avair pas d'échange à faire / J'ai passé commande le 12/12 et je n'ai pu récupérer ma commande au relais que le 21/12 au lieu de 48h après car il était livré mais pas enregistré. / Annonce par mail commande livrée au point relai et quand je suis allée la chercher elle n'était pas là. / à ce jour je n'ai toujours pas ma commande. / aucun suivi de cette commande. / je n ai rien a dire je suis tres heureuse de ma commande
Non classé	je crois savoir ma la respnsable du pont relais queles livraisons ne sont effectuées que 2 fois par semaine ce qui repousse le délai de 48h prévu. Que popuvez-vous faire pour remédier à ce soucis ? ?

TAB. 3 – Extraits de regroupements sur quatre jeux de tests

de chaque cluster pondérée selon le nombre de messages qu'il contient. Les deux critères étant complémentaires, on considérera un résultat meilleur qu'un autre s'il l'est sur l'un des critères et est au moins aussi bon sur l'autre. On admettra par ailleurs de manière non formelle qu'un résultat n'est plus vraiment acceptable en dessous d'une homogénéité moyenne de 30% et au delà d'un écart total de 20 rangs ainsi que le fait qu'un gain de 3 rangs sur le critère d'écart permet de compenser une perte moyenne de 10% sur

le critère d'homogénéité. Nous avons pris comme référence la combinaison mise en évidence précédemment («solution avec tous les traitements»), puis nous l'avons comparée avec quelques autres proches en terme de traitements (Fig. 2).

La solution de référence obtient un bon résultat mais les autres traitements de la négation (en liste noire et sans traitement) semblent meilleurs, le traitement en liste noire l'étant davantage que la solution sans traitement. Parmi les messages comportant une négation dans ce jeu, seul 44% répondent aux critères de complexité évoqués. Cela illustre la faible pertinence de cette solution en présence de messages complexes. Il est également intéressant de souligner que toutes les solutions ont produit un nombre de clusters plus élevé que celui attendu. Nos solutions n'arriveraient donc pas à regrouper les idées autant qu'un expert. Ce test met également en évidence la très forte pertinence de la liste noire et de la distance sémantique des mots calculée selon les sens. Tous les tests réalisés contradictoirement obtiennent des résultats nettement inférieurs. La pondération des sens semble meilleure que la pondération des mots car elle permet de produire moins de groupes même s'ils semblent légèrement moins homogènes. L'emploi de la lemmatisation ou de la stemmatisation semble indispensable sous peine de produire un nombre de clusters trop élevé. La lemmatisation se révèle toutefois plus efficace que la stemmatisation et de qualité presque égale voire légérement supérieure à la combinaison des deux méthodes. Ce constat tend à prouver que la stemmatisation peut dégrader les résultats par rapprochements de mots ayant seulement une racine commune.



Symbole	Ecart	Homogénéité
♯	2	20,12
∞	4	19,94
⊥	15	45,62
⋆	15	55,15
△	16	53,94
∀	17	50,91
+	18	46,83
⊤	18	49,1
×	19	48,92
b	20	47,83

FIG. 2 – Résultats statistiques de différentes méthodes

3 Conclusion

Nous avons étudié le regroupement de textes courts d'opinions. Dans ce cadre, La représentation vectorielle se révéle pertinente car elle fournit une distance sémantique simple à calculer et la possibilité d'y adjoindre un système de pondération favorisant les dimensions les plus intéressantes pour les rapprochements. L'entropie de Shannon donne de bons résultats pour le calcul de ces poids surtout après la prise en compte de la sémantique des mots à l'aide d'une ontologie et sans avoir recours à une désambiguisation du corpus. La recherche de marques de négation en vue d'un marquage des messages dans leur intégralité comme niées ou affirmées nous permet d'obtenir généralement de meilleurs résultats sur les regroupements que les autres solutions envisagées dans le cadre de textes très courts. A l'inverse, cette méthode

se révèle être médiocre dans un cadre contraire. Enfin, la lemmatisation se révéle très peu efficace dans un contexte orthographique difficile. Ce constat n'est pénalisant dans notre application que si les fautes portent sur l'opinion évoquée. La stemmatisation des formes lemmatisées permet de les corriger dans le cas où le radical n'est pas affecté. Les axes de travaux futurs concernent la mise en place d'une correction orthographique automatique du corpus avant le traitement et une réflexion plus approfondie sur la gestion des messages complexes comportant plusieurs idées.

Remerciements

Je remercie messieurs Pierre Marquis et Vincent Dubois, directeur et co-encadrant de thèse, et messieurs Antoine Serniclay et Thibaud Vibes, responsables d'Onyme, pour leurs conseils dans mes recherches.

Références

DEERWESTER S., DUMAIS S., FURNAS G., LANDAUER T. & HARSHMAN R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, **41**(6), 391–407.

LUND K., BURGESS C. & ATCHLEY R. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th annual conference of the Cognitive Science Society*, volume 17, p. 660–665.

MILLER G. (1995). Wordnet : a lexical database for english. *Communications of the ACM*, **38**(11), 41.

POIRIER D., BOTHOREL C., GUIMIER DE NEEF E. & BOULLÉ M. (2008). Automating opinion analysis in film reviews : the case of statistic versus linguistic approach. In *Proceedings of the LREC 2008 Workshop on Sentiment Analysis : Emotion, Metaphor, Ontology and Terminology*, p. 94–101.

RASTIER F. (1989). Sens et textualité. *Paris, Hachette*.

RILOFF E. (1995). Little words can make a big difference for text classification. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*, p. 130–136 : ACM New York, NY, USA.

SAGOT B. & FIŠER D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. In *Proceedings of TALN 2008*.

SALEM A. (1987). *Pratique des segments répétés : essai de statistique textuelle*. Klincksieck.

SALTON G., WONG A. & YANG C. (1975). A vector space model for automatic indexing. *Communications of the ACM*, **18**(11), 620.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, volume 12 : Manchester, UK.

SOWA J. (1984). *Conceptual structures*. Addison-Wesley Reading, MA.

SPARCK JONES K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, **28**(1), 11–20.

VOSSEN P. (1998). Eurowordnet a multilingual database with lexical semantic networks. *Computational Linguistics*, **25**(4).