Détecter les intentions d'achat dans les forums de discussion du domaine automobile : une approche robuste à l'épreuve des expressions linguistiques peu répandues

Marguerite Leenhardt[1], Gaël Patin[2]

[1] XiKO / SYLED-Paris 3 – marguerite.leenhardt@xiko.fr
[2] XiKO / ERTIM-INaLCO – gael.patin@xiko.fr

Abstract

The current economic context has raised new expectations among organizations, seeking for new types of applications regarding the online chatter happening every day in online forums. This paper presents a corpus-based approach for discovering and acquiring qualitative knowledge on an uncommon type of expression nevertheless valued by firms, which is purchase intent. It aims at considering the process of knowledge discovery and acquisition for a text classification task at an industrial scale, based on over 30 000 messages gathered on various French online forums dedicated to the automobile domain. The displayed process avoids the costly steps of elaborating traditional linguistic resources such as knowledge cartridges, ensuring a maximum contextual accuracy while providing a robust and reliable method for purchase intent detection. Encouraging results after evaluating this method shows it achieves a satisfying performance, given the uncommon type of expression the system has to deal with and the short time span that an industrial context demands.

Résumé

La conjoncture actuelle mène les entreprises à former de nouvelles attentes quant aux systèmes dédiés à l'analyse des données conversationnelles issues du web, notamment des forums de discussion. Cet article présente une approche adossée sur le traitement d'un grand corpus de données, dont l'objectif est centré sur la découverte et l'acquisition de connaissances qualitatives à partir d'un type peu répandu d'expressions linguistiques disponibles : les intentions d'achat. L'exposé s'appuie sur l'analyse d'un corpus de plus de 30 000 messages récoltés sur différents forums français liés au domaine de l'automobile et traite des préalables pour la mise en place d'un système de catégorisation de textes à l'échelle industrielle. La démarche proposée évite le coûteux travail induit par l'élaboration des ressources linguistiques traditionnelles, telles que les cartouches de connaissances, tout en garantissant une pertinence contextuelle maximale et en proposant une méthode robuste et fiable pour la détection des intentions d'achat. Les résultats sont encourageants : la méthode proposée obtient des performances satisfaisantes, étant donnés le type peu répandu d'expressions que le système doit prendre en charge et les délais de production courts imposés par la mise en œuvre d'applications en contexte industriel.

Mots-clés : analyse des données textuelles, découverte de connaissances, conversations web, forums

1. Introduction

Dans la phase actuelle, différents secteurs économiques hexagonaux sont confrontés à une conjoncture peu favorable. C'est notamment le cas du commerce de détail de la vente automobile[1]. L'acquisition de nouveaux clients constitue, à cet égard, un levier important pour améliorer le climat des affaires. Les entreprises s'intéressent aux prises de parole spontanées de leurs clients, notamment sur le Web (Dutrey et al., 2012), afin d'améliorer la qualité de leur offre de services. Or, de nombreux usagers du Web s'expriment

[1] Source : enquête de conjoncture INSEE du mois d'octobre 2013 dans le commerce de détail et l'automobile, accessible au lien suivant : http://www.insee.fr/fr/themes/info-rapide.asp?id=86 (consulté le 21.11.2013).

quotidiennement sur les biens de consommation qu'ils cherchent à se procurer. Il apparaît nécessaire de fournir des solutions d'analyse de données textuelles adaptées, qui permettent non seulement de traiter les grands volumes de données disponibles dans les conversations spontanées sur le Web, mais aussi d'y découvrir les intentions d'achat qui y sont exprimées chaque jour. Le fonctionnement de telles solutions, implantées en contexte industriel, implique de mettre en œuvre des procédés spécifiques ; ces systèmes informatisés doivent être capables :

(i) de récolter des données dans des flux d'information structurés de façon hétérogène ;

(ii) de prendre en charge des productions qui divergent de la norme linguistique documentée dans les règles de grammaire ou d'orthographe d'une langue ;

(iii) d'effectuer des analyses précises, c'est-à-dire identifier les intentions d'achat portant sur un type de bien de consommation particulier en produisant le moins d'erreurs possible ;

(iv) d'opérer ces analyses de façon robuste, c'est-à-dire sans diminution de performance lorsque le système est confronté à des données bruitées.

Nous présentons dans cet article la façon dont la plateforme de la société XiKO répond à ces différentes problématiques de récolte, de gestion des données linguistiques tout-venant, de précision et de robustesse des analyses. L'objectif de notre démarche est de parvenir à élaborer, dans les délais courts qui sont ceux de l'industrie, des ressources pertinentes pour la mise en œuvre d'un système de catégorisation suffisamment performant ; la tâche que nous abordons est complexifiée car elle porte sur une forme de production rare, que sont les expressions d'intentions d'achat liées aux véhicules automobiles dans les forums de discussion. Après avoir abordé la question de la collecte de données hétérogènes issues des conversations de forums Web en vue d'une application industrielle spécifique (section 2), nous évoquerons plus particulièrement l'analyse des productions linguistiques spontanées axée sur l'identification des intentions d'achat (section 3), ainsi que son incidence sur la méthode mise en œuvre pour l'élaboration de ressources linguistiques incrémentales (section 4), établies à partir des connaissances découvertes dans les conversations analysées.

2. Récolter des données à partir de forums de discussion

Tout système de récolte automatique de données textuelles sur le Web est confronté à un problème d'envergure, en ce qu'il doit faire face à de grands ensembles de données hétérogènes, faiblement structurées, bruitées et en constante évolution. La collecte de données issues de conversations de forums Web pose quant à elle le problème spécifique de l'identification et de la réplication de la structure conversationnelle telle qu'elle est construite par les internautes qui y participent. Avant d'aborder plus avant ces aspects et leur incidence sur l'implémentation d'un système automatisé adéquat, nous nous proposons de brosser un rapide portrait de ces supports Internet que sont les forums Web, en tant qu'espaces communautaires jouant un rôle prépondérant – par leur pérennité comme par leur densité – dans le partage et la diffusion des connaissances collectivement construites et partagées par les utilisateurs, toujours plus nombreux, du réseau Internet.

2.1. Quelques généralités sur les forums : la question de la structure « froide »

Il est couramment admis que le premier forum est WIT – WWW Interactive Talk, lancé en 1994 par le W3C pour faciliter les discussions sur des problèmes techniques précis et les

archiver de façon structurée. Ce projet est lié à la naissance du principe du WWW (World Wide Web), introduit à l'aube des années 1990 au Centre Européen de Recherche Nucléaire (CERN). Il faudra attendre 2000 pour qu'un dispositif analogue, mais gratuit et dont le code est librement accessible (*open source*), soit disponible : il s'agit de phpBB, un moteur de forums développé en PHP, toujours très utilisé aujourd'hui sur la toile[2]. Sans en dresser ici une chronologie détaillée, il faut néanmoins souligner la culture fortement communautaire des forums, qui trouve très probablement son origine dans la méthodologie des RFC (*Request for Comments*) héritée d'ARPANET[3]. Deux traits spécifiques la caractérisent :

- rechercher et mettre en commun des informations ;
- faciliter les discussions ouvertes sur un réseau de télécommunications.

D'un point de vue technique, un forum est une arborescence de pages web dont le dispositif technique permet une gestion aisée des utilisateurs, de l'organisation des discussions et de la publication des messages qu'ils produisent. Lorsque nous parlons de structure, c'est d'abord au sens de la structure arborescente dans laquelle sont organisés les échanges des internautes[4] : nous en présentons un exemple factice ci-après.

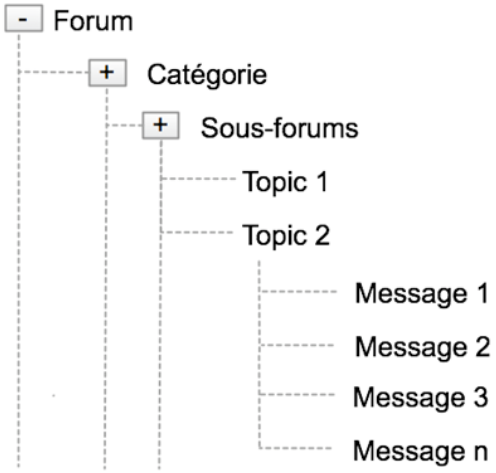

Figure 1. Exemple de structure arborescente d'un forum

En pratique, la structure formelle de la page web sur laquelle les utilisateurs publient leurs contributions fournit l'information « froide », c'est-à-dire la structure de l'observable lorsqu'on l'appréhende tel quel. Il s'agit de savoir comment distinguer entre la « structure froide » des données et la vision analytique de la structuration des échanges, projetée éventuellement sur les données lorsque l'on met en place un système de récolte de données

[2] Les dispositifs techniques permettant la mise en place de forums se sont depuis multipliés. Une infographie sur l'évolution des logiciels de forums de 1994 à nos jours a été établie récemment par le spécialiste Forum Software Reviews. Pour plus de détails sur les dispositifs techniques les plus répandus, voir http://www.scriptol.fr/cms/classement.php (liens consultés le 23 octobre 2013).

[3] En 1962, l'ARPA soutient des initiatives de projets de recherche en informatique et contribue à développer un réseau de collaboration actif entre la recherche et les industries. ARPANET est lancé en 1969 ; avec lui apparaît la méthodologie des *Requests For Comments* (RFC) dont le but est d'organiser les échanges entre les différents contributeurs du projet. Le principe des RFC consiste à faire part de ses commentaires et à donner un avis argumenté pour le documenter.

[4] Et non pas au sens de l'*infra*-structure logicielle (format de fichiers, protocoles d'échange de données).

textuelles[5]. Ce niveau d'information doit être codé dans le corpus, au même titre que les autres informations renvoyant à la situation d'interaction, telles que la date et l'heure de publication d'un message, par exemple. Cela a une incidence sur le type d'outillage adéquat à mobiliser :

- il doit permettre de distinguer entre « structure froide » et « structure analytique » ;
- dans le cadre des échanges médiatisés par ordinateur, qui ne sont pas toujours en adéquation avec le « standard de la langue »[6], le système devrait être indépendant de ressources linguistiques préexistantes[7], définies dans la grande majorité à partir de productions linguistiques standard.

2.1. L'extraction à l'épreuve de l'hétérogénéité des forums : identifier et conserver la structure des conversations extraites

L'extraction de données textuelles en contexte hétérogène implique la mise au point de stratégies adaptées, c'est-à-dire permettant d'obtenir des données non bruitées. Par exemple, lorsque l'on s'intéresse au contenu conversationnel d'un forum Web, les encarts publicitaires ou certains artefacts – tel que le texte présent dans un bouton « Répondre » – présents sur une page, sont considérés comme facteur de bruit. Un certain nombre d'informations complémentaires doivent être récupérées, afin de conserver le maximum de contexte sur la situation dans laquelle apparaît une production linguistique. A cet égard, la date de publication d'un message, le pseudonyme de son auteur, constituent, entre autres, des informations contextuelles pertinentes. Enfin, des éléments contextuels plus complexes semblent indispensables à prendre en compte : on pense notamment à la disposition particulière du signifié dans l'ordre de succession des messages – par exemple, l'indentation d'un message répondant à un message précédent.

L'outil informatique a ses limites et peut difficilement distinguer seul entre les informations pertinentes et les éléments de bruit. Il pourrait donc sembler adéquat de recourir au travail humain pour le recueil de ces informations. Cependant, la collecte de corpus manuelle, intégralement réalisée par l'humain, est inenvisageable en contexte industriel, notamment parce qu'elle induit un coût important en temps et en main d'œuvre. Pour contribuer à pallier cette problématique, en particulier dans l'objectif de réduire les coûts et le temps de production des corpus, la plateforme XiKO réduit les tâches manuelles à la seule sélection des structures des pages web dont les contenus sont pertinents à extraire. La stratégie adoptée est fondée sur le recours à XPath[8]. Ce langage a l'avantage de permettre la localisation précise de portions de pages web, en permettant une recherche séquentielle dans la structure de ces pages. Cette approche s'avère économe en temps de mise en place et bien adaptée à l'extraction de données à partir de forums ne disposant pas de flux RSS.

[5] Cette question est d'ailleurs abordée par (Marcoccia, 2004), du point de vue de l'enjeu méthodologique qu'elle implique pour l'analyse linguistique menée sur corpus.

[6] On pense en particulier à l'utilisation des smileys, aux abréviations de type « SMS », à des usages alternatifs de la ponctuation (ponctuation expressive par exemple), hormis les problématiques liées à la performance linguistique des internautes, du point de vue du standard orthographique et syntaxique notamment.

[7] Tout du moins, le système devrait donner le choix d'y recourir ou non.

[8] XPath est un langage d'indication de chemin dans du contenu plus ou moins structuré, de type XML, à l'aide duquel on exprime une requête afin d'extraire un élément quelconque pointé par le chemin indiqué.

3. Analyser les productions linguistiques spontanées

Notre objectif est de parvenir à traiter le matériau linguistique tel qu'il est donné à voir dans la situation d'expression dans laquelle il existe, c'est-à-dire en adoptant une posture non normative sur les productions linguistiques que le système d'analyse doit prendre en charge. De ce point de vue, la statistique textuelle offre des méthodes robustes pour l'exploration des données linguistiques issues des conversations spontanées du Web. En effet, à la différence de certaines technologies de Traitement Automatique des Langues Naturelles, elle s'affranchit de l'usage de ressources linguistiques externes.

La catégorisation automatique de textes (« classification » en anglais) correspond au fait d'associer une catégorie à un texte, en fonction des informations linguistiques ou para-linguistiques du texte en question. Plus précisément, la tâche de catégorisation de textes a pour objectif de chercher à identifier une relation entre un ensemble de textes et un ensemble de catégories disponibles, en s'appuyant sur un modèle chargé de prédire l'appartenance d'un texte à une catégorie, dont la mise en œuvre est généralement associée à un algorithme d'apprentissage automatique. Cela implique donc de disposer d'un ensemble d'apprentissage, c'est-à-dire d'un groupe de textes dont la catégorie est préalablement connue – en ce sens, la tâche de catégorisation est supervisée – lequel ensemble permet d'estimer la performance des paramètres du modèle de catégorisation, c'est-à-dire le modèle produisant le moins d'erreurs possible. Les principales difficultés de la catégorisation automatique appliquée à des textes produits en langage naturel, relèvent de l'équivocité et de la variabilité du matériau linguistique considéré. Le modèle de prédiction est en outre élaboré à partir d'une vision subjective : il s'agit d'une décision motivée par l'interprétation sémantique d'un contenu textuel par un expert ; sur cet aspect particulier, la littérature indique que la consistance de la décision peut d'ailleurs varier d'un expert à l'autre. L'étape de l'élaboration du modèle de prédiction revêt une importance particulière : elle consiste à représenter les paramètres à partir desquels l'algorithme d'apprentissage automatique va fonctionner. Différentes stratégies peuvent être appliquées, notamment :

(i) la représentation du texte dite en « sac de mots », où un texte est représenté par le vecteur des mots qui le constituent ; cette approche est directement confrontée au problème de pouvoir définir ce qu'est un « mot ». Loin d'être trivial à résoudre, il s'accompagne généralement de stratégies de pré-traitements des textes, pour par exemple identifier les mots composés.

(ii) la représentation du texte par un vecteur de phrases ou de syntagmes, dont les performances sur le plan de la significativité statistique sont faibles, comme l'illustrent par exemple les expérimentations menées par (Lewis, 1992) ; les travaux de (Caropreso et al. 2001) y proposent une alternative, qui consiste en l'exploitation des « phrases statistiques », ensemble de mots cooccurrents dans un texte, identifiés à l'issue de pré-traitements tels que la suppression des mots grammaticaux et le recours à la racinisation.

(iii) la représentation du texte par des racines lexicales et des lemmes, qui sous-tendent une réduction de la complexité induite par les différentes flexions et dérivations de mots, impliquant donc nécessairement un recours à des étapes de pré-traitement des textes fondés sur des caractéristiques morphologiques ou morphosyntaxiques.

(iv) la représentation du texte en séquences de n caractères, ou « n-grammes », où le texte est modélisé par le profil des n-grammes les plus fréquents qu'il contient ; cette approche s'affranchit, par définition, de la prise en compte des caractéristiques linguistiques des textes, ce qui la rend opératoire indépendamment de la langue ou de la performance linguistique thématisée dans le texte.

La catégorisation automatique de textes peut par exemple être exploitée pour la recherche documentaire, de veille d'information et plus généralement dans les systèmes de Recherche d'Information. Ces dernières années, les applications d'analyse des opinions et des sentiments ont connu un important développement (Pang et Lee, 2008), portées par la demande croissante du marché des entreprises et des organisations, préoccupées par les avis exprimés à leur égard par le grand public sur le Web (Dutrey et al., 2012). Dans ce sillon, les entreprises, qui cherchent à améliorer leurs performances commerciales, nourrissent des attentes quant à un nouveau type d'application de catégorisation de textes, lié à la détection des intentions d'achat. C'est précisément à ce cas de figure spécifique que nous nous intéressons ici. Nous présentons la stratégie retenue dans le cadre de l'implémentation de la plateforme XiKO pour l'identification de ces phénomènes particuliers au sein des conversations de forums Web (section 4), ainsi que l'évaluation des résultats obtenus (section 5).

4. Identifier les caractéristiques de l'intention d'achat

4.1. La méthode des spécificités pour la représentation des connaissances

La stratégie de représentation des connaissances retenue rappelle celle des « phrases statistiques » proposée par (Caropreso et al. 2001). Cependant, il faut souligner avec insistance le fait qu'aucun pré-traitement du matériau linguistique n'est opéré dans notre cas. Le parti que nous retenons est de proposer une méthode de représentation des connaissances capable de conserver toute la complexité, par là même, la richesse linguistique des conversations spontanées sur le Web, sans chercher à les normaliser ou à projeter sur elles une vision normative de la langue. En effet, la stratégie que nous proposons s'appuie sur la méthode des *spécificités*, bien connue dans le domaine de la statistique textuelle (Lebart et Salem, 1994). Elle consiste à définir des volets d'un corpus de travail afin d'appliquer à chacun d'entre eux des calculs probabilistes dans le but d'y mettre en évidence les éléments caractéristiques. Les différents volets du corpus procèdent généralement du choix d'un analyste, lequel choix est motivé par le contexte et l'objectif de l'analyse. Concrètement, le calcul des spécificités de l'un des volets du corpus fait émerger des occurrences de segments textuels particulièrement distinctifs du volet donné, par rapport au reste du corpus. Autrement dit, le caractère sur ou sous-spécifique de ces segments n'y est très probablement pas le fruit du hasard et peut trouver différentes explications, d'un point de vue stylistique, thématique, sémantique ou encore discursif. L'analyste s'appuie sur les résultats de ce calcul, pour identifier les segments textuels qui semblent le mieux caractériser les connaissances contenues dans le corpus. Il convient de souligner l'importance des choix présidant à la segmentation de volets dans le corpus de travail, étant donné leur forte incidence sur la richesse et la pertinence des segments textuels identifiés, donc des connaissances découvertes en corpus. Cette stratégie de découverte pour la représentation des connaissances présente l'avantage de pouvoir être appliquée indépendamment du degré de standardisation des données linguistiques du corpus, sans impliquer de pré-traitement visant à en réduire la complexité. Au contraire, la qualité du modèle de connaissances découle de la prise en compte de cette complexité intrinsèque au matériau linguistique. C'est donc cette stratégie de

représentation des connaissances vers laquelle nous nous sommes tournés pour identifier les caractéristiques de l'expression de l'intention d'achat dans les conversations spontanées des internautes.

4.2. L'identification de caractéristiques de l'expression de l'intention d'achat pour la catégorisation automatique

Dans le cadre dans lequel nous nous situons, l'objectif est de mettre au point un système informatisé, capable de déterminer automatiquement dans quelle catégorie classer les messages qu'il reçoit en entrée. Nous abordons dans cette section, premièrement le mode de représentation des messages et des règles, ensuite la question de la sélection des attributs utilisés pour concevoir les règles, et enfin la création en elle-même des règles de catégorisation. Les messages sont représentés sous la forme d'un vecteur de booléens, chaque booléen du vecteur indiquant la présence ou l'absence d'une caractéristique dans un message donné. Le système de catégorisation est constitué d'un ensemble de règles fondées sur des fonctions logiques booléennes, les variables des fonctions étant les valeurs booléennes du vecteur.

Il faut ensuite aborder l'étape de la sélection des attributs utilisés par le système. Pour identifier les messages manifestant l'expression d'une intention d'achat, deux types de caractéristiques sont utilisés dans la procédure que nous proposons : d'une part, les paradigmes de segments textuels ; d'autre part, leurs éventuelles cooccurrences au sein d'un même message. Nous appliquons la démarche évoquée plus haut, fondée sur un découpage du corpus en deux volets, le premier volet contenant les messages porteurs d'une intention d'achat, le second, tous les autres messages disponibles. Les segments qui émergent suite au calcul des spécificités sont, soit spécifiques, soit sous-spécifiques des messages contenant une intention d'achat. Les premiers sont alors considérés comme des indices pertinents pour l'identification des intentions d'achat dans des messages du corpus, par contraste avec tous les autres messages du corpus. Nous présentons ci-dessous, un exemple d'une caractéristique spécifique de l'expression de l'intention d'achat, matérialisée sous la forme d'une cooccurrence :

je suis … à la recherche … d'une voiture

Il s'agit donc, dans ce cas, de trouver les cooccurrences des composantes « *je suis* », « *à la recherche* » et « *d'une voiture* » dans la fenêtre d'un message du corpus. Voici maintenant un exemple de caractéristique sous-spécifique, matérialisée sous la forme d'un paradigme de segments textuels, permettant d'identifier les messages qui ne sont pas porteurs de l'expression d'une l'intention d'achat :

réparation / expert / sous garantie / symptôme / rouille / corrosion / diagnostic

Le calcul des spécificités permet donc de faire émerger, à partir d'ensembles de données importants, des caractéristiques ayant un fort potentiel discriminant pour mener à bien la tâche de catégorisation automatique à laquelle nous nous intéressons ; ce, de façon rapide et adossée sur un corpus de productions linguistiques réelles.

L'ultime étape consiste à concevoir les règles du système sous le contrôle d'un analyste, qui garantit la pertinence de l'application des règles sur un corpus d'entraînement dont la composition est détaillée dans la section suivante. Dans une perspective d'industrialisation, l'étape de conception des règles et des caractéristiques à partir desquelles elles sont élaborées,

doit satisfaire à la double contrainte du temps et de la maintenabilité. En effet, le nombre de règles doit être restreint pour limiter la complexité du système, favoriser un temps de traitement rapide et participer d'une maintenance la moins coûteuse possible.

5. Evaluation de la méthode proposée

5.1. Corpus et procédure d'évaluation

Le système de catégorisation présenté a pour objectif l'identification des messages exprimant des intentions d'achat de véhicules automobiles, produits par des internautes dans des forums spécialisés. Etant donné le contexte industriel de l'application présentée – dans lequel il s'agit de pouvoir mettre en œuvre une procédure opératoire sur des conversations spontanées du Web telles qu'elles se manifestent, tout en prenant en compte des contraintes de coût et de temps de traitement – le choix d'une évaluation sur des données réelles, sans pré-traitement linguistique, s'est imposé. Pour ce faire, la procédure d'évaluation a été appliquée à un corpus constitué dans le cadre d'une application client réelle. Les données ainsi exploitées proviennent de dix[9] forums spécialisés du Web français, dont ont été extraites, sur une période d'un mois, l'ensemble des conversations associées à des marques et modèles d'automobiles. Le corpus est constitué de 30 341 messages, comprend 50 841 formes, 1 138 084 occurrences et 26 423 hapax. Les messages récoltés, répartis en 4214 conversations, ont, dans un premier temps, fait l'objet d'analyses préliminaires au moyen d'outils d'exploration de corpus, afin d'en identifier les principales caractéristiques. Ce préalable a permis de gagner un temps précieux pour la réalisation d'une tâche d'annotation manuelle, effectuée au niveau global du message, qui a consisté à repérer ceux exprimant une intention ferme d'achat d'une automobile, à court ou moyen terme. Sur l'ensemble du corpus, 193 messages ont été identifiés comme contenant une intention d'achat ; en voici quelques exemples :

« Suis a la recherche d'un monospace. Ma voiture est tombée en carafe. Avec un déménagement et l'arrivée de 2 bb, le budget est tout petit. J'ai vu une annonce pour un espace phase 2, 2.1 TD de 1994 avec 230000 km pour 700€. Qu'en pensez-vous ? Quels sont les pièges à éviter ? »

Forum 321 Auto, nat39800, septembre 2013.

« Bonjour,
Je suis a la recherche d'une voiture, j'ai testé une Audi A3 1.9 tdi 110 et elle m'a beaucoup plus cependant il y a des réparations esthétique à faire (aile arrière) elle a 180 m. Km
Qu'est ce que vous en pensez et quel prix vous y mettriez ? Je suis tombé aussi sur une annonce d'une A3 a 180m. Km (Moteur) et a 325m. (Caisse)
Pour le même prix mais sans réparation esthétique pour 2500€ Est ce que ça vaut le coup ?
Merci. »

Forum France, Jooker243, septembre2013

« Bonjour, je créais ce topic car ma sœur désire changer de voiture.
Son choix c'est porté sur le grand scénic 3 neuf en diesel. Et elle me demandai des conseilles sur la motorisation. Alors elle hésite entre le dci 130 ou le dci 150 en BVA ?
Pour moi je pense que le 130 doit être un très bon moteur et suffisant pour le scénic. Donc je demande quelques conseilles juste pour aider un peux
Merci »

lacoste71 FORUM AUTO

[9] Les dix forums pris en compte sont les suivants : 321 Auto, Auto Evasion, Auto Passion, AutoPlus, Autosportive, Autotitre, Forum Auto, Forum France, Lotus ZE, Motor Legend.

Premier constat à l'issue de cette étape de travail : le nombre de messages contenant une intention d'achat manifeste, est très faible en regard du nombre total de messages. En effet, le ratio d'intention d'achat sur le nombre total de messages est de 0,63% : cela contribue à indiquer que l'on se trouve en présence d'un signal quantitativement faible, donc difficile à capter dans le flux des conversations des internautes. Concernant l'évaluation de la méthode proposée, nous nous sommes conformés aux pratiques en vigueur pour l'évaluation des systèmes de catégorisation. A cette fin, le corpus de travail a été divisé en deux ensembles, le premier dédié à d'entrainement et le second à l'évaluation proprement dite. L'ensemble d'entrainement, dont la vocation est de servir à la conception du système de catégorisation, est constitué de 23 871 messages soit approximativement 80 % du corpus ; il comporte 130 intentions d'achat. L'ensemble d'évaluation, qui doit servir à évaluer la qualité de notre système, est constitué des 6 470 messages restants, dont 63 intentions d'achat. L'évaluation de notre système a été effectuée en appliquant la procédure d'identification des caractéristiques d'intentions d'achat (décrite dans la section 4) sur l'ensemble d'entraînement. Le système de catégorisation, qui comprend 9 règles simples, est constitué par un analyste à partir des caractéristiques d'intention d'achat jugée comme pertinentes. Ce système catégorise ensuite automatiquement les messages de l'ensemble d'évaluation. Les résultats, présentés ci-après pour évaluer la performance du système, sont indiqués en termes de rappel et de précision. Dans le cas d'une catégorisation mono-catégorie, le rappel est le ratio du nombre de messages correctement catégorisés comme « intention d'achat » sur le nombre total de messages annotés « intention d'achat » ; la précision est le ratio du nombre de messages correctement catégorisés « intention d'achat » sur le nombre total de messages catégorisé « intention d'achat ».

5.2 Résultats de l'expérimentation

Nous rappelons ici que le système de catégorisation présenté appartient au paradigme des *systèmes experts*. L'un des principaux problèmes de ce genre de dispositifs informatisés est celui du coût de conception des ressources linguistiques qu'ils exploitent, que sont, notamment, les règles de catégorisation, les lexiques d'expert, les ontologies ou encore les cartouches de connaissances. En effet, l'élaboration de ce type de ressources, largement utilisées à ce jour en contexte industriel, peuvent prendre plusieurs mois à mettre en place et s'avèrent difficiles à maintenir dans le temps (Ezzat, à paraître). Pour la constitution des ressources exploitées par notre système, nous nous sommes restreints à l'intervention d'un analyste pendant trois jours, afin de mettre en avant les performances de notre méthode pour la conception de connaissances pertinentes et légères à mettre en place, en vue d'une exploitation optimisée en contexte industriel. Le système obtient les résultats suivants : un rappel de 74,6 %, 59,3 % en précision et 66,1 en F-mesure. Ces résultats sont encourageants en regard de la tâche et de la difficulté de traiter un sociolecte fluctuant et peu normalisé du point de vue du standard de la langue. Enfin, il faut également rappeler la difficulté de maintenir un niveau de F-mesure acceptable, c'est-à-dire supérieur à 50, pour une catégorie de message très faible représentée dans le corpus, en l'occurrence 0,63 % du nombre total de messages. (Laza et al., 2011) ont constaté une forte dégradation des performances de leur système, qui implémente un algorithme de réseau bayésien adossé à des ressources linguistiques de type ontologique. En effet, selon leurs évaluations, le système mis en place n'a pu obtenir de performance supérieure à une F-mesure de 40.

6. Discussion et perspectives

Les données textuelles générées par les internautes, notamment dans les forums de discussion, sont devenues l'une des mines d'information privilégiées par les entreprises pour se mettre à l'écoute de leurs clients. Si les particularités linguistiques des conversations issues du Web ont déjà fait l'objet de nombreux travaux, les solutions existantes sont confrontées à la gestion de l'écart entre les ressources linguistiques implémentées dans les systèmes d'analyse et la matérialité des productions spontanées des internautes. De surcroît, alors que les recherches liées à l'analyse des opinions ont connu un important essor ces dernières années, les applications liées à des objectifs métiers plus spécifiques telles que l'identification de prospects pour l'acquisition de nouveaux clients, restent peu investiguées à ce jour.

Dans la phase actuelle, les attentes des entreprises évoluent quant aux applications d'analyse des conversations Web, dans un mouvement qui va vers la détection des intentions d'achat dans les conversations spontanées des internautes s'exprimant dans les forums de discussion. Détecter ce type d'expressions spontanées représente un potentiel économique non négligeable, qui intéresse directement les processus d'acquisition de clients et d'avant-vente. En nous intéressant à ce phénomène dans les forums de discussion liés au domaine de l'automobile, nous avons été confrontés à la rareté des données, qui ont généralement un impact très négatif sur la performance des systèmes. Cependant, la méthode proposée, qui consiste à exploiter les statistiques textuelles, en particulier le calcul des spécificités pour la découverte et l'acquisition de connaissances pertinentes exploitables par un système de catégorisation automatique industriel, obtient des résultats encourageants, tout en satisfaisant à l'impératif de délais courts pour le déploiement d'un système de catégorisation dédié à une demande spécifique. Elle présente en outre l'avantage d'être facilement maintenable et peu coûteuse à élaborer, sans recourir à des étapes de pré-traitement linguistiques, donc en conservant intacte la complexité du matériau linguistique disponible dans les conversations spontanées des internautes.

Références

Barthel T., Beaudouin V., Collin O., Fleury S. et Vié C. (2000). Les forums publics sur Intranoo (en 1999), RP/FTR&D/6861, CNET, Paris

Biyani P., Bhatia S., Caragea C. et Mitra P. (2012). Thread specific features are helpful for identifying subjectivity orientation of online forum threads. In *Proceedings of COLING 2012 : Technicak Papers,* pp. 295-310

Boullier D. et Lohard A. (2012). Opinion mining et Sentiment analysis. Méthodes et outils. *OpenEdition Press*, 236 p.

Caropreso M-F., Matwin S. et Sebastiani F. (2001). A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In Amita G. Chin (Ed.), *Text Databases and Document Management : Theory and Practice*, Idea Group Publishing, pp. 78-102

Dutrey C., Peradotto A. et Clavel, C. (2012). Analyse de forums de discussion pour la relation client : du Text Mining au Web Content Mining. In *Actes des JADT 2012*, pp. 445-457

Ezzat M. Acquisition de relations entre entités nommées à partir de corpus (Thèse de Doctorat menée à l'INaLCO sous la direction de T. Poibeau, à paraître)

Habert B. (2005). *Instruments et ressources électroniques pour le français.* Ophrys, Paris, 176p.

Hampton K. et Wellman B. (2002). Neighboring in Netville : How the Internet Supports Community and Social Capital in a Wired Suburb, *City and Community* 2, 4 : 277-311

Jalam R. (2003). *Apprentissage automatique et catégorisation de textes multilingues*, Thèse de Doctorat en Informatique, Université Lumière Lyon 2

Lafon P. (1980). Sur la variabilité de la fréquence des forms dans un corpus. *Mots*, 1 : 127-165

Laza R., Pavon R., Reboiro-Jato M. et Fdez-Riverola F. (2011). Evaluating the effect of unbalanced data in biomedical document classification. Journal of Integrative Bioinformatics, 8(3):177, 2011

Lebart L. et Salem A. (1994). *Analyse statistique des données textuelles*. Dunod, Paris, 344 p.

Lewis D. (1992). An evaluation of phrasal and clustered representations on a texte categorization task. In Croft et al. (Eds), *Proceedings of SIGIR-1995, 15th ACM International Conference on Research and Development in Information Retrieval*, pp. 37-50

Marcoccia M. (2004). L'analyse conversationnelle des forums de discussion : questionnements méthodologiques, Les carnets du CEDISCOR (en ligne), n°8, pp.22-37, accessible en ligne : http://cediscor.revues.org/220 (consulté le 5 février 2011)

Mourlhon-Dallies F. (2007). Communication électronique et genres du discours. Regards sur l'internet, dans ses dimensions langagières. Penser les continuités et discontinuités. GLOTTOPOL, Revue de sociolinguistique en ligne, 10 : 11–23.

Pang B. et Lee L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, vol.2 : 1-135.