Détecter le besoin d'information dans des requêtes d'usagers d'agents virtuels : sélection de données pertinentes

Octavia Efraim[1] Fabienne Moreau[1]
(1) LIDILE EA 3874, Université Rennes 2, Place du recteur Henri Le Moal, CS 24307, 35043 Rennes CEDEX
octavia-edie.efraim@univ-rennes2.fr, fabienne.moreau@univ-rennes2.fr

RÉSUMÉ

Pour orienter efficacement les messages reçus par différents canaux de communication, dont l'agent virtuel (AV), un système de gestion de la relation client doit prendre en compte le besoin d'information de l'usager. En vue d'une tâche de classification par type de besoin d'information, il est utile de pouvoir en amont sélectionner dans les messages des utilisateurs, souvent de mauvaise qualité, les unités textuelles qui seront pertinentes pour représenter ce besoin d'information. Après avoir décrit les spécificités d'un corpus de requêtes d'AV nous expérimentons deux méthodes de sélection de segments informatifs : par extraction et par filtrage. Les résultats sont encourageants, mais des améliorations et une évaluation extrinsèque restent à faire.

ABSTRACT

Selecting relevant data for information need detection in virtual agent user queries

Customer relationship platforms offer a variety of communication channels, which include virtual agents (VA). Efficient routing of messages must take into account the user's information need. For a task of classification by information need, it may be useful first to select from these often noisy messages those units of text which relevantly represent the information need. We describe a corpus of VA user queries, and experiment on it with two methods for selecting relevant segments: one based on term extraction, the other on filtering. The results are encouraging but there is room for improvement, and extrinsic evaluation remains to be done.

MOTS-CLÉS : agent virtuel, besoin d'information, sélection de termes, données bruitées

KEYWORDS: virtual agent, information need, term selection, noisy data

1 Introduction

Les plateformes de relation client multicanales proposent aux usagers des modalités variées d'interaction avec les entreprises, dont l'agent virtuel (AV) – un outil d'assistance automatisée aux internautes. Par souci d'optimisation des ressources humaines (conseillers clientèle), il peut être intéressant que la plateforme qui reçoit de manière centralisée les messages envoyés par des clients par tout canal (e-mail, tchat, AV, SMS, application mobile, formulaire de contact Web, etc.) dispose d'une fonctionnalité permettant de classer ces messages par ordre de priorité, pour les acheminer vers la ressource la plus économique qui soit en mesure d'y apporter une réponse satisfaisante : réponse par un conseiller humain – en direct ou différée – ou réponse automatique – en direct (AV) ou différée (e-mail ou autre). Les critères de priorité pourront varier selon les besoins, mais ils

devront intégrer en tout cas une notion de besoin d'information de l'usager. Concept incontournable en recherche d'information (RI), le **besoin d'information** peut s'entendre comme le sujet sur lequel l'usager souhaite obtenir des informations (Manning et al., 2008). Dans le domaine des systèmes de question-réponse cette notion est complétée par celles d'objet et de focus d'une question (Ibekwe-SanJuan, 2007 ; Grau, 2004). Si la définition de ces concepts n'est pas toujours la même pour tous les chercheurs, tous s'accordent généralement à dire que le besoin d'information reste implicite et ne peut être appréhendé que par son expression langagière. Saisir ce besoin pour l'intégrer à un système de routage des messages passe donc nécessairement par l'analyse du message.

Dans ce contexte, nous souhaitons explorer des méthodes de classification supervisée des messages par catégories de besoin d'information. Ces techniques devront être souples et portables, pour pouvoir traiter des messages multicanaux et multi-domaines, ayant donc des caractéristiques très variées. Dans une première étape, nous avons constitué un corpus de requêtes d'usagers d'AV, dont l'une des principales caractéristiques est liée à leur nature fortement dégradée (erreurs, langage abrégé, etc.). Avant de pouvoir catégoriser le besoin d'information (la définition des catégories de besoin et l'étiquetage du corpus font en effet l'objet d'un autre volet de nos travaux, non abordé ici), il est nécessaire de pouvoir l'identifier à partir de la requête de l'utilisateur. Cet article se veut donc être une étude préliminaire visant à proposer et évaluer des méthodes pour sélectionner dans les messages potentiellement bruités des utilisateurs les unités textuelles les plus pertinentes pour représenter le besoin d'information. Concrètement, nous nous plaçons dans une approche vectorielle de la représentation textuelle, où les problèmes liés aux dimensions de l'espace de représentation sont bien connus (Amini, Gaussier, 2013). En partant en effet du constat que des requêtes comme *bonjour j'aimerai savoir comment faire pour saoir l'etat de ma comande* contiennent beaucoup de bruit autour du noyau informatif (*etat de ma comande*), nous souhaitons parvenir à réduire les requêtes à ce noyau avant de les inclure dans un modèle de classification. L'intérêt de réduire le nombre de termes représentant un texte est généralement reconnu (Luhn, 1958 ; Aseervatham et al., 2011 ; Almeida et al., 2011). Cependant, certains travaux remettent en cause l'utilité de cette approche (Riloff, 1995), tandis que d'autres font état de résultats amenant à la nuancer selon la taille du jeu de données (Mejova, Srinivasan, 2011), la méthode de classification utilisée (Poirier et al., 2009) ou encore la méthode de sélection de variables employée (Dolamic, Savoy, 2010). Notre objectif est précisément de juger de l'utilité de différentes techniques de réduction du texte sur des données complexes, telles que celles issues des nouveaux canaux de communication, et pour la tâche que nous envisageons, dans l'hypothèse que les traitements prévus (y compris la correction) pourraient être plus efficaces s'ils ne sont appliqués qu'aux données porteuses de sens. Sur notre corpus de requêtes d'AV, nous avons expérimenté deux méthodes de sélection, basées respectivement sur l'extraction et sur le filtrage d'unités textuelles. Nous nous proposons, dans un premier temps, d'évaluer empiriquement la qualité des représentations réduites ainsi obtenues. Si elles sont satisfaisantes, il s'agira à une étape ultérieure (lorsque nous disposerons du corpus étiqueté) d'évaluer leur apport potentiel pour la classification. Pour cette première évaluation, nous définissons une notion opérationnelle de **pertinence** par rapport à l'expression d'un besoin d'information, en empruntant à la typologie des actes de conversation de Traum et Hinkelman (1992) : est pertinent dans notre contexte tout segment de la requête qui est porteur d'un « acte de langage fondamental » appartenant aux deux catégories « information » et « demande » ; les autres catégories d'actes fondamentaux, ainsi que les actes argumentatifs, de prise de parole et de synchronisation entre locuteurs (en l'occurrence, l'usager et l'AV), ne contribuent pas à l'expression du besoin d'information. Le but des méthodes proposées n'est donc pas de classer les requêtes selon une typologie des besoins d'information – cette tâche sera confiée au classifieur – mais de fournir à ce dernier des données plus ciblées, où les unités textuelles, même erronées, associées à l'expression du besoin d'information sont mieux cernées.

2 Caractéristiques des requêtes d'agents virtuels

Généralités sur le corpus. Les données qui font l'objet de notre étude sont issues de dialogues menés par des internautes avec 18 AV, collectés sur une période de deux ans et demi. Les domaines des AV sont divers (mutuelle, agence de voyages, mairie, e-commerce, etc.). La langue des conversations est le français, à l'exception de quelques requêtes envoyées par des internautes étrangers. Dans cette étude nous nous intéressons uniquement aux tours de parole des usagers et ignorons les interventions des AV, qui sont prédéfinies. Notre analyse se situe donc au niveau des messages pris en isolation. Il est intéressant de chercher à identifier le besoin d'information dans chaque requête et non pas uniquement au niveau global de la conversation, ce besoin se retrouvant très souvent exprimé à plusieurs reprises au cours du dialogue, inchangé ou ayant évolué. Il est courant d'appliquer aux données certains **prétraitements** élémentaires avant de les soumettre à des traitements plus avancés (Amini, Gaussier, 2013 ; Ibekwe-SanJuan, 2007). Ces opérations de normalisation réalisent une première sélection de variables : elles consolident des formes ou suppriment des variables considérées comme dépourvues d'intérêt pour une tâche donnée. Si l'objectif est toujours de réduire les dimensions de l'espace de représentation du texte, les décisions prises à cette étape sont déterminantes pour la suite des traitements (Grefenstette, Tapanainen, 1994), dont elles doivent prendre en compte les spécificités et les besoins. Les opérations de normalisation que nous avons jugées pertinentes dans notre contexte, choisies ou adaptées pour prendre en compte l'orthographe souvent incohérente ou incorrecte des internautes, sont les suivantes : mise en minuscules ; suppression de la ponctuation ; suppression des traits d'union et des apostrophes (sauf dans *aujourd'hui*) ; suppression des *t* épenthétiques (là où les tirets existent) ; reconstruction des formes élidées (notamment les pronoms) ; remplacement par une étiquette respectivement des caractères numériques, adresses Web et adresses mail ; suppression des émoticônes (intéressantes pour l'analyse des sentiments, mais ne contribuant pas à exprimer un besoin d'information). Une réduction de la variation morphologique flexionnelle ou dérivationnelle est souvent pratiquée aussi. Cependant, les outils de lemmatisation ou racinisation, traditionnellement conçus pour analyser du texte propre, ont des difficultés à traiter des données bruitées comme celles issues de la communication en ligne. Pour évaluer ces difficultés sur notre corpus, nous l'avons lemmatisé et étiqueté par catégorie grammaticale à l'aide du TreeTagger (Schmidt, 1994). L'orthographe fautive est elle-même une autre source de dispersion des variables. De nombreux travaux ont cherché à nettoyer les erreurs spécifiques aux nouveaux corpus : contenus en ligne (Dey, Haque, 2008), tweets (Han, Baldwin, 2011), SMS (Guimier De Neef et al., 2007 ; Tagg et al., 2014). À cette étape de nos travaux nous avons souhaité non pas corriger, mais comprendre de manière plus précise la nature du bruit présent dans notre corpus, pour envisager des possibilités de correction à une étape future.

Description détaillée du corpus. Le corpus contient 79.698 requêtes d'usagers. On y compte 645.637 mots et 25.500 formes uniques. Plus de la moitié (55,27%) des formes uniques ont une seule occurrence – proportion similaire à celle rapportée par Poirier et al. (2009) pour un corpus de commentaires en anglais portant sur des films. En éliminant les mots à fréquence 1 on obtiendrait 11.406 formes uniques. Les messages sont plutôt courts (75% ont moins de 10 mots) et le nombre de requêtes par AV varie considérablement entre des limites extrêmes (64 et 20.000). Le nombre de mots par AV (médiane de 9793 mots) augmente de façon quasi-linéaire avec le nombre de requêtes (de 219 à 151.100 mots).

Pour analyser certaines caractéristiques de notre corpus et pour évaluer l'efficacité des deux méthodes de sélection testées, nous avons constitué un **échantillon** extrait aléatoirement du corpus et contenant à l'origine 1000 requêtes, dont une, en anglais, a été éliminée. Pour cette analyse nous

ignorons les 185 mots de l'échantillon constitués exclusivement de caractères numériques, ce qui en réduit la taille à 8051 mots. Nous avons annoté sur l'échantillon la complexité de la requête (cf. ci-dessous) et la présence d'erreurs (type et correction proposée). De nombreuses typologies décrivent les erreurs qui caractérisent les nouvelles formes de communication écrite (Anis, 2002, entre autres). Pour faciliter l'annotation, nous avons opté pour une typologie plus sommaire, limitée au niveau lexical : non-mot (*mototisé*), mot réel (*coordonnes* pour *coordonnées*) ; faute de grammaire (souvent donnant un mot réel : *quel* pour *quelle*) ; mots fusionnés (*pouravoirun*) ; mot segmenté (*télép honique*) ; diacritique (*recu*). Nous indiquons aussi pour chaque mot s'il s'agit d'une abréviation ou d'une entité nommée. Outre ces annotations descriptives, nous avons effectué des annotations évaluatives : correction de l'étiquetage morphosyntaxique et de la lemmatisation ; pertinence (au sens défini dans l'introduction) des résultats obtenus par les deux stratégies de sélection de données expérimentées.

Caractéristiques linguistiques des requêtes. Les requêtes se présentent comme des (enchaînements de) mots-clés, des phrases simples, des phrases complexes ou encore des séquences de phrases. Leur niveau de complexité structurelle est généralement associé à une **complexité** liée au besoin d'information, que nous définissons une fois de plus en termes d'actes de conversation (Traum, Hinkelman, 1992) : est complexe une requête qui contient plus d'un acte de langage fondamental d'information ou de demande. Selon cette définition, 14% des requêtes de notre échantillon sont complexes. Par exemple, *bonjour, je voudrais savoir si une hospitalisation pour esthétique mais validée et prise en charge par la ss est-ce que si j'ai l'assurance hospitalisation, j'ai droit aux indeminités de votre part et y a-t-il une prise en charge des dépassements d'honoraires* est décomposable en un acte d'information et deux de demande. Au contraire, 12,7% de nos requêtes n'expriment pas un vrai besoin d'information. En effet, il n'est pas rare que les internautes abordent l'AV avec des intentions ludiques, comme dans *veut tu une pizza ?* ou *parlez-vous chinois?* Ces requêtes ne présentent pas d'intérêt particulier dans le cadre de nos travaux, elles pourront donc être directement traitées par l'AV pour ne pas encombrer inutilement le module de routage. Les messages diffèrent également en termes de registre de langue (du courant-soutenu au vulgaire) et sont souvent agrémentés de marques d'expressivité (majuscules, signes de ponctuation redoublés), pourtant supprimées car sans intérêt pour notre tâche. Les erreurs d'orthographe affectent 10,38% des mots de notre échantillon. 36% des fautes produisent un mot réel, 22,13% sont des fautes grammaticales et presque la moitié sont liées aux diacritiques. Si nous nous intéressons à la présence et aux types d'erreur dans notre corpus, c'est aussi pour apprécier la faisabilité de différents traitements TAL sur ces données. Le lemmatiseur utilisé sur notre corpus s'avère extrêmement sensible aux fautes d'orthographe, même mineures (accentuation) : seuls 17,58% des mots erronés ont été associés au lemme correct. La performance est meilleure au niveau de l'étiquetage morphosyntaxique : 46,17% des mots erronés ont la bonne étiquette, et si l'on accepte des étiquettes partiellement correctes (*e.g.* le cas des confusions entre formes verbales), le taux de succès augmente considérablement. Ceci laisse penser que des traitements basés sur des motifs (lexico-)syntaxiques plutôt que sur des lemmes pourraient être viables.

3 Sélection de données pour la détection du besoin de l'usager

Pour réduire les requêtes à ses éléments pertinents en termes de besoin d'information, nous proposons ici de tester deux méthodes dont l'objectif est soit de cibler les segments du message qui contiennent de l'information utile, soit au contraire de filtrer les éléments non pertinents. Ces méthodes sont simples mais présentent l'avantage d'être assez souples et donc assez adaptées au traitement de données erronées.

3.1 Méthode par extraction

Des approches consistant à extraire comme traits, sur les textes à analyser, des termes connus d'avance se retrouvent, par exemple, dans la classification par sentiment basée sur des lexiques (Liu, 2015). D'autres méthodes proposent de calculer les termes à retenir directement sur le document analysé au sein d'une collection, selon des mesures de pondération issues de la RI, dont les plus simples restent *df* (fréquence documentaire) et *tf-idf* (Ibekwe-SanJuan, 2007 ; Amini, Gaussier, 2013). Ces mesures, utilisées en RI pour le classement des documents d'une collection en fonction de leur similarité avec une requête, peuvent aussi servir comme critères de classement d'unités de différentes granularité dans une collection, par pouvoir discriminant, sans égard à une éventuelle requête. Ainsi, *df* permet de classer les éléments du vocabulaire d'une collection, tandis que *tf-idf* permet de classer les termes d'un document de la collection et a été exploité également pour la sélection de phrases pertinentes à l'intérieur d'un document (Zechner, 1996 ; Orăşan et al., 2003 ; Farzindar et al., 2005). Notre objectif est de cibler la partie pertinente de la requête, dont nous faisons l'hypothèse qu'elle est représentée par ses termes les plus discriminants. En considérant chaque requête comme un document (très court) au sein de la collection constituée par le corpus AV, nous avons testé la sélection de termes au niveau de la requête par tf-idf. La difficulté à appliquer cette méthode pour réduire les requêtes réside dans la définition d'un critère de seuillage satisfaisant. À une évaluation qualitative des résultats obtenus sur plusieurs AV, différents seuils fixes testés se sont avérés trop agressifs, réduisant complètement des requêtes pertinentes, tandis qu'un seuil dynamique (la médiane des valeurs tf-idf des mots de la requête) s'est montré peu précis et imprévisible, tantôt retenant des mots pertinents, tantôt les supprimant.

Face à ce problème, nous avons souhaité essayer une méthode plus adaptée à notre tâche que le *tf-idf*. Elle repose sur l'extraction dans les requêtes d'éléments appartenant à un lexique prédéfini construit automatiquement, dont nous supposons qu'il reflète pertinemment les besoins d'information possibles des usagers. Une fois le lexique constitué, chaque nouvelle requête sera représentée par (une fonction de) la présence de termes appartenant au lexique. Pour constituer ce lexique, nous avons choisi de le compiler sur des critères statistiques à partir de l'ensemble des requêtes (duquel nous avons enlevé l'échantillon destiné à l'évaluation) plutôt que sur l'ensemble des réponses et modèles de questions dont dispose chaque AV. Cette approche privilégie des formulations réellement employées par les usagers pour exprimer leur besoin d'information. Pour identifier les termes spécifiques à notre corpus et susceptibles donc d'être pertinents, un premier essai a consisté à comparer l'ensemble de notre corpus avec un corpus de langue générale de grande taille (frTenTen [Jakubíček et al., 2013], accessible depuis SketchEngine [Kilgariff et al., 2014]). Les termes retournés sont globalement pertinents, mais l'applicabilité de cette méthode est tributaire de l'accès à des ressources externes. Par ailleurs, si cette méthode met en avant les spécificités globales de notre corpus, nous jugeons intéressant de pouvoir faire ressortir des différences plus fines entre les AV. Nous proposons donc une méthode de compilation du lexique qui ne se rapporte plus à un corpus externe, mais exploite la composition même du corpus à l'étude. Ainsi, le fait de connaître l'appartenance des requêtes à un AV (assimilable grosso modo à un domaine) nous permet de faire ressortir des termes fortement associés spécifiquement à chaque AV, en classant les termes du vocabulaire de l'AV par ordre décroissant du ratio entre leur fréquence dans le sous-corpus AV respectif et leur fréquence dans le corpus englobant (d'où l'intérêt de mettre en commun les AV). Ce ratio, qui prend des valeurs de 0 à 1 et est interprétable en termes de probabilité conditionnelle, pourrait aussi être intéressant dans un contexte de modélisation thématique (*e.g.* pour détecter le domaine d'une requête). Deux paramètres (ratio minimal et fréquence totale minimale) permettent de contrôler : la spécificité du terme pour l'AV (elle augmente avec la valeur du ratio) ; la permissivité de l'extraction face à des fautes d'orthographe (la baisse du seuil de

fréquence totale minimale entraîne l'inclusion de mots plus rares, dont des graphies fautives). Sur ce dernier point, il convient de noter que cette méthode d'extraction n'a pas de notion de correction, ce qui lui permet de sélectionner des termes significatifs comportant des fautes, dès lors que ces formes erronées sont suffisamment fréquentes dans les requêtes ; en effet, il arrive souvent qu'une forme erronée (surtout lorsqu'il s'agit d'accents manquants) soit non seulement fréquente dans les requêtes, mais plus fréquente même que la forme correcte (constat réalisé aussi par Suignard et Kerroua (2013) sur un corpus de notes de conseillers clientèle). Il n'est donc pas nécessaire de modéliser explicitement les erreurs. Cette méthode nous a permis d'extraire des termes n-grammes (n allant de 1 à 4) très représentatifs de chaque sous-corpus. Nous avons varié légèrement le ratio minimal et la fréquence totale minimale selon la taille du n-gramme : 0,7 et 5 pour les unigrammes ; 0,8 et 4 pour les bigrammes ; 0,8 et 5 pour les trigrammes et les quadrigrammes. L'extraction de termes est peu sensible à la présence dans les corpus de mots à faible valeur informative (mots vides ou autres, dès lors qu'ils sont communs aux sous-corpus), au niveau des unigrammes. Pour les n-grammes d'ordre supérieur à 1, une étape de filtrage consistant à éliminer tous les termes commençant ou se terminant par un mot vide nous a permis de réduire considérablement le bruit.

Uniquement pour apprécier la couverture des termes retenus, nous avons opéré sur les requêtes de chaque AV une sélection des termes de son lexique. Cette évaluation quantitative a révélé que 55,67% des requêtes ont eu au moins un terme extrait. Selon une évaluation qualitative effectuée sur l'échantillon pour juger de la pertinence des représentations réduites obtenues, même si 58,2% des requêtes exprimant un besoin d'information ont été modifiées par le traitement, dans 32,2% cas les termes repérés dans la requête ne sont pas suffisants pour décrire complètement ce besoin. Si la performance de la méthode par extraction reste assez faible, son principal atout est la spécificité élevée des termes, qui limite, d'autre part, la capacité de cette méthode à récupérer, sans augmenter le bruit, des unités lexicales pertinentes réparties plus uniformément entre les sous-corpus (expressions d'un besoin d'information commun à plusieurs AV).

3.2 Méthode par filtrage

La sélection de termes par suppression d'unités à faible pouvoir discriminant est une pratique courante en TAL. Elle peut prendre la forme du filtrage par un antidictionnaire compilé (automatiquement ou manuellement) en amont et contenant des mots « vides » (*cf.* Poirier et al., 2009 ; Riloff, 1995 ; Dolamic, Savoy, 2010, pour des questionnements sur l'impact de leur suppression) ou d'autres unités risquant de bruiter les analyses requises par la tâche considérée (*e.g.* en analyse des sentiments [Liu, 2015]). D'autres fois, les unités à supprimer ne sont pas énumérées, mais définies par des critères liés aux exigences de la tâche. Ainsi, en résumé automatique on a exploré le filtrage selon des critères syntaxiques (épuration des phrases par retranchement des compléments circonstanciels [Prince, Yousfi-Monod, 2006]), rhétoriques (Marcu, 1998) ou d'organisation textuelle (segments situés entre parenthèses [Farzindar et al., 2005]) ; le modèle du canal bruité a également été appliqué, dans l'idée de récupérer le segment court d'origine auquel du texte non-essentiel aurait été rajouté (Knight, Marcu, 2002). Le filtrage que nous proposons vise à se rapprocher le plus possible du cœur informatif du message, par élagage itératif de structures récurrentes non-porteuses d'information utile en termes de besoin d'information. À partir d'une analyse manuelle des fréquences de n-grammes dans le corpus, nous avons défini 77 patrons lexicaux simples mais le plus génériques possibles, pour qu'ils soient généralisables aux différents domaines traités par les AV, et que nous avons groupés selon les classes sémantiques suivantes : salutations (*bonjour, ça va ?*), clôture (*cordialement* (+ nom)), excuse (*excusez-moi*), remerciement (*merci*), souhait (*je souhaiterais*), recherche (*je cherche*, *j'ai besoin de*), capacité intellectuelle (*savoir*, *je ne comprends pas*), capacité (*puis-je*, *si possible*), demande (*dites-moi*, *merci de (bien*

vouloir)), question (*est-ce que*, *comment (faut-il faire pour)*). À la différence de l'utilisation des patrons lexicaux en résumé automatique, nous les exploitons comme filtre négatif et non comme indicateurs thématiques permettant de sélectionner les segments importants (Paice, 1980 ; Farzindar et al., 2005 ; Orăşan et al., 2003). De plus, certains mots interrogatifs n'ont pas la même valeur dans nos patrons qu'en question-réponse, où ils servent au typage de la réponse attendue (Grau, 2004 ; Li, Roth, 2002) : dans notre corpus *comment* n'indique pas un besoin d'informations procédurales, mais fonctionne le plus souvent comme un mot-outil servant à introduire l'objet de la demande. Une fois la liste des patrons constituée à partir du corpus, elle servira, pour tout nouveau message, de filtre à appliquer de l'extérieur vers l'intérieur de la requête et de manière itérative. Cette méthode a l'avantage de préserver la cohérence du segment retourné.

Comme pour la première méthode, uniquement pour apprécier la couverture des patrons retenus nous avons filtré l'ensemble du corpus : 45,45% des requêtes ont été transformées. Ce pourcentage est inférieur à celui obtenu par la méthode par extraction (55,67%). Pourtant la qualité du résultat est supérieure : 69% des requêtes pertinentes ont été traitées correctement, ce qui était prévisible, puisque la stratégie du filtrage est de retourner « moins ou tout » (un fragment, sinon le tout), alors que la méthode par extraction retourne « moins ou rien » (le(s) terme(s) ou rien). Or parfois la requête est déjà minimale (mots-clés) et la retourner telle quelle est la bonne réponse ; par contre, si ces mots-clés ne se retrouvent pas sur la liste des termes ils ne seront pas extraits. À titre d'exemple, le résultat du filtrage sur quelques requêtes (prétraitées) est indiqué entre crochets : *bonjour je voulait savoir comment faire pour* [*voir le suivi de ma demande de pret*] ; *bonjour serais t il possible de savoir* [*ou en ne et ma commande*] ; *bonjour je aimerais savoir* [*où en ai ma commande NUM*] *merci* ; *bonjour je aimerai savoir comment faire pour* [*saoir le etat de ma comande*]. On constate que les patrons sont capables de gérer certaines erreurs (grammaire plus qu'orthographe) grâce à une définition souple. Là où le filtrage échoue notablement c'est sur les requêtes complexes, puisqu'il n'opère que sur les marges du message, or les motifs se retrouvent souvent à l'intérieur. Une solution serait d'étendre les règles pour permettre la segmentation de la requête aux points d'ancrage correspondant aux motifs, pour ensuite procéder au filtrage de manière récursive.

4 Conclusions et perspectives

La finalité de cette étude était de vérifier la faisabilité de réduire des requêtes d'AV à leurs éléments pertinents en termes de besoin d'information, en vue d'une tâche de catégorisation. Nous avons proposé deux méthodes qui constituent d'abord, à partir d'un corpus de requêtes passées, une liste ou une description des éléments à rechercher dans des requêtes nouvelles, soit pour les y retenir (ils constitueront alors la représentation de la requête), soit pour les en supprimer (la requête sera alors représentée par les éléments restants). Même si les premiers résultats sont encourageants, les méthodes proposées demeurent encore insuffisantes puisqu'elles ne permettent pas de cerner le besoin d'information dans au moins 30% des requêtes de notre échantillon. Ces premiers travaux offrent néanmoins des perspectives intéressantes. Nous avons déjà mentionné plusieurs pistes pour améliorer nos méthodes de sélection de données pertinentes (ajout de patrons lexicaux et traitement des requêtes complexes pour la méthode de filtrage, optimisation des paramètres pour la méthode d'extraction) et souhaitons également proposer une approche hybride, qui combinerait les deux méthodes selon des pondérations et des règles de priorité. L'objectif sera d'évaluer le gain de performance que cette sélection de variables peut rapporter dans le contexte d'une tâche et d'un modèle de classification précis, par rapport au modèle de base (entraîné sur des requêtes non traitées). Dans une étape ultérieure nous envisageons également d'élargir le champ de notre étude pour prendre en compte l'ensemble de la conversation et ainsi relier la notion de besoin d'information à l'historique de l'échange entre l'AV et l'usager.

Ces travaux bénéficient du soutien de la Région Bretagne à travers un financement ARED.

Références

ALMEIDA T., JURANDY ALMEIDA A., YAMAKAMI A. (2011). Spam filtering: how the dimensionality reduction affects the accuracy of naive Bayes classifiers. *JISA* 1 (3), 183-200.

AMINI M.-R., GAUSSIER E. (2013). *Recherche d'information*. Paris : Eyrolles.

ANIS J. (2002). Communication électronique scripturale et formes langagières : chats et SMS. Actes des *Journées* « *S'écrire avec les outils d'aujourd'hui* ».

ASEERVATHAM S., GAUSSIER E., ANTONIADIS A. et al. (2011). Régression logistique et catégorisation de textes. *Modèles statistiques pour l'accès à l'information textuelle*, 97-122.

DEY L., HAQUE S. (2008). Opinion mining from noisy text data. Actes de *Second workshop on analytics for noisy unstructured text data*, 83-90.

DOLAMIC L., SAVOY J. (2010). When stopword lists make the difference. *JASIST* 61 (1), 200-203.

FARZINDAR A., ROZON F., LAPALME G. (2005) CATS a topic-oriented multi-document summarization system at DUC 2005. Actes de *2005 Document understanding workshop*.

GRAU B. Les systèmes de question-réponse. (2004). Ihadjadene M. (éd.) : *Méthodes avancées pour les systèmes de recherche d'informations*. Paris : Lavoisier, 189-218.

GREFENSTETTE G., TAPANAINEN P. (1994). What is a word ? What is a sentence ? Problems of tokenization. Actes de *3rd conference on computational lexicography and text research*, 79-87.

GUIMIER DE NEEF E., DEBEURME A., PARK J. (2007). TiLT correcteur de SMS : évaluation et bilan qualitatif. Actes de *14e conférence TALN*, 123-132.

HAN B., BALDWIN T. (2011). Lexical normalisation of short text messages: makn sens a #twitter. Actes de *49th Annual Meeting of the Association for Computational Linguistics*, 368-378.

IBEKWE-SANJUAN F. (2007). *Fouille de textes*. Paris : Lavoisier.

JAKUBICEK M., KILGARRIFF A., KOVAR V. et al. (2013). The TenTen corpus family. *7th International corpus linguistics conference*.

KILGARRIFF A., BAISA V., BUSTA J. et al. (2014). The Sketch Engine: ten years on. *Lexicography* 1 (1), 7-36.

KNIGHT K., MARCU D. (2002). Summarization beyond sentence extraction: a probabilistic approach to sentence compression'. *Artificial Intelligence* 139 (1), 91-107.

LI X., ROTH D. (2002). Learning question classifiers. Actes de *19th International Conference on Computational Linguistics (COLING)*, 556-562.

LIU B. (2015). *Sentiment analysis*. Cambridge : Cambridge University Press.

LUHN H.P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development* 2 (2), 159-65.

MANNING C., RAGHAVAN P., SCHÜTZE H. (2008). *Introduction to information retrieval*. Cambridge : Cambridge University Press.

MARCU D. (1998). Improving summarization through rhetorical parsing tuning. Actes de *6th Workshop on Very Large Corpora*, 206-15.

MEJOVA Y. SRINIVASAN P. (2011). Exploring feature definition and selection for sentiment classifiers. *ICWSM*.

ORĂSAN C., MITKOV R., HASLER L. (2003). CAST: a computer-aided summarisation tool. Actes de *10th Conference on European Chapter of the Association for Computational Linguistics* 2, 135-38.

PAICE C.D. (1980). The automatic generation of literature abstracts: an approach based on the identification of self-indicating phrases. Actes de *3rd Annual ACM Conference on Research and Development in Information Retrieval*, 172-91.

POIRIER D., FESSANT F., BOTHOREL C. et al. (2009). Approches statistique et linguistique pour la classification de textes d'opinion portant sur les films. *Revue des nouvelles technologies de l'information*, 147-169.

PRINCE V., YOUSFI-MONOD M. (2006). Compression de phrases par élagage de leur arbre morpho-syntaxique. *Revue des sciences et technologies de l'information* 25 (4), 437-468.

RILOFF E. (1995). Little words can make a big difference for text classification. Actes de *18th Annual ACM Conference on Research and Development in Information Retrieval*, 130-36.

SCHMIDT H. (1994). Probabilistic part-of-speech tagging using decision trees. Actes de *International Conference on New Methods in Language Processing*.

SUIGNARD P., KERROUA S. (2013). Utilisation de contextes pour la correction automatique ou semi-automatique de réclamations clients. Actes de *20e conférence TALN*, 699-706.

TAGG C., BARON A., RAYSON P. (2014). "i didn't spel that wrong did i. Oops": analysis and normalisation of SMS spelling variation. Cougnon L.-A., Fairon C. (éd.) : *SMS communication*. Amsterdam : John Benjamins, 217-237.

TRAUM D.R., HINKELMAN E.A. (1992). Conversation acts in task-oriented spoken dialogue. *Computational Intelligence* 8 (3), 575-599.

ZECHNER K. (1996). Fast generation of abstracts from general domain text corpora by extracting relevant sentences. Actes de *16th Conference on Computational Linguistics* 2, 986-89.