Identification de questions pour traiter les courriels par une méthode question-réponse

Luc Bélanger, Guy Lapalme

RALI – DIRO – Université de Montréal
C.P. 6128, succ. Centre-Ville
Montréal – Québec – Canada H3C 3J7
{belanglu, lapalme}@iro.umontreal.ca

Abstract

Automating the process of responding to e-mails is a way of improving customer relation management. We use techniques developed in the area of question-answering for the automatic response to e-mails. This article shows how to identify the questions contained in e-mails coming from the investors relation service of an enterprise. We use a rule-based approach which identifies 81% of the questions found in a corpus of e-mails.

Résumé

La réponse automatisée aux courriels est une solution pour améliorer les services de relations avec la clientèle. Pour y arriver, nous utilisons des techniques issues du domaine de la question-réponse. Nous présentons comment identifier les questions contenues dans un corpus de courriels provenant d'un service de relation avec les investisseurs d'une grande entreprise. Avec une approche à base de règles, nous avons identifié 81% des questions du corpus de courriels.

Keywords: automatic email response, question-answering, question extraction, customer relations management, digital reference.

Mots-clés : réponse automatisée au courriel, question-réponse, extraction de question, gestion des relations avec la clientèle, référence numérique.

1. Introduction

Ces travaux ont été réalisés dans le cadre du projet Mercure dont le but est d'élaborer de nouvelles méthodes de traitement des courriels pour la réalisation de la réponse automatisée aux courriels. Le domaine d'application du projet est le service de relation avec les investisseurs d'une compagnie publique. Les courriels envoyés aux services de relation avec les investisseurs ont la particularité de contenir une grande proportion de courriels qui sont des questions ou qui expriment un besoin d'information s'apparentant à une question. L'approche choisie pour traiter le problème est d'utiliser les techniques issues du problème de la question-réponse. L'identification des questions est une tâche préalable et essentielle pour traiter les courriels par les techniques de la question-réponse car elle permet d'isoler une partie importante du discours nécessaire pour retrouver l'information demandée.

Notre tâche principale a été la création d'une grammaire identifiant les patrons lexicaux les plus susceptibles d'être une question ou une requête. Les patrons ont été conçus manuellement pour obtenir un taux de rappel élevé, au détriment d'un taux de précision plus faible et d'une identification incomplète de l'information pour le traitement automatique. Nous avons pu ainsi repérer 81% des courriels contenant des questions.

2. Présentation du problème

La réponse automatisée aux courriels est nécessaire pour les départements de gestion des relations avec la clientèle des grandes entreprises. Les communications par courriels sont devenues si importantes et si abondantes qu'il est maintenant indispensable d'avoir des moyens pour les traiter automatiquement afin d'améliorer le service. Des études (Jupiter Communications, 2000 ; Banter, 2001) démontrent qu'il est essentiel pour les compagnies d'offrir un service de relation avec la clientèle qui traite les communications électroniques. Pour satisfaire les attentes des clients utilisant ce mode de communication, un courriel doit être traité à l'intérieur d'un délai d'au plus 6 heures, rendant l'automatisation du traitement nécessaire.

Différentes approches ont été proposées pour effectuer le traitement automatisé des courriels. Parmi les plus simples, il y a celles où l'expéditeur du courriel doit lui-même faire la classification de son courriel, pour l'envoyer au bon service, selon des adresses de courriels distinctes. Il y a également les systèmes d'auto-réponse, popularisés par les gestionnaires de listes de discussion (e.g. LISTSERV[1], Mailman[2], Majordomo[3]), qui fonctionnent par activation par des mots-clés. Ces approches sont rudimentaires et elles ne peuvent pas être utilisées en pratique pour traiter efficacement le flot de communications par courriel d'un service avec la clientèle.

Une approche populaire dans les entreprises est l'utilisation de systèmes de gestion des courriels. Ces systèmes demandent une réponse manuelle, contribuant à diminuer le temps de traitement des courriels en automatisant certaines fonctionnalités : la catégorisation des courriels, l'envoi d'accusé de réception, le routage des messages, la suggestion de réponses, l'intégration du système dans l'environnement de travail du préposé, l'archivage des courriels et la production de rapports statistique et historique. Le système Kana Response[4] est un exemple de système de gestion des courriels où des traitements sont effectués automatiquement.

Dans le cadre de notre projet, nous nous intéressons aux approches traitant les courriels de manière autonome. Les systèmes de ce type sont la plupart du temps développés spécifiquement pour un client. Puisque le domaine de ces systèmes est restreint, ceux-ci peuvent plus facilement utiliser des techniques sophistiquées. Les composantes les plus intéressantes utilisées par ce type de système sont : les ontologies du domaine, l'utilisation de patrons de textes, raisonnement à base de cas et méthodes de catégorisation.

Dans cet article, nous abordons le traitement des courriels en considérant qu'ils contiennent des questions ; nous nous inspirons des méthodes des systèmes de question-réponse. Les communications provenant des clients sont souvent une question ou une requête. Au même titre que la question-réponse, la réponse automatisée aux courriels est une tâche apparentée à la recherche d'information. Par contre la requête n'est plus seulement une suite de mots-clés ou une question, mais bien un courriel contenant une ou plusieurs questions.

Watanabe *et al.* (2003) ont tenté de solutionner le problème du traitement automatisé des courriels à l'aide d'un système de question-réponse. Ils ont généré des réponses aux courriels envoyés dans une liste de discussion. Ces courriels ressemblent à une conversation car il y a des courriels de questions et des courriels contenant les réponses. L'étape principale de leur système est l'extraction de phrases significatives dans les courriels à l'aide d'un pointage tenant compte de la présence de noms significatifs, de patrons d'expressions typiques d'une question et d'une pondération liée au nombre d'occurrences de la phrase dans les courriels réponse. Une fois

[1] http//www.lsoft.com/
[2] http//www.list.org/
[3] http//www.greatcircle.com/majordomo/
[4] http//www.kana.com/

les phrases significatives extraites, un calcul de similarité est effectué avec la base de courriels réponse pour déterminer quel courriel retourner comme réponse, similaire à un raisonnement à base de cas.

Les courriels sont des données textuelles avec des propriétés différentes de celles utilisées pas les systèmes de question-réponse élaborés pour les compétitions TREC du NIST (Voorhees, 2001). Les questions utilisées par les systèmes de question-réponse TREC sont extraites à partir des relevés de requêtes envoyées aux engins de recherche, et ensuite nettoyées pour les rendre plus facilement intelligibles pour les systèmes, elles sont généralement courtes et précises. L'information contextuelle contenue dans les courriels est une différence qui doit être exploitée pour le traitement des courriels. Les questions contenues dans les courriels sont énoncées différemment, les formules de politesse et les énoncés d'intention compliquent l'identification et la compréhension des questions.

Dans les systèmes de question-réponse, l'analyse et la classification des questions est un élément essentiel pour chercher efficacement une réponse à la question (Hermjakob, 2001). Les courriels étant des données brutes, nous devons les catégoriser selon l'objet du courriel et la présence de questions. Les données utilisées pour l'élaboration de notre module d'identification des questions sont composées de 210 courriels préalablement nettoyés, analysés et catégorisés manuellement. Ces courriels contiennent tous au moins une question et ils sont rédigés en anglais. La catégorisation des courriels a été réalisée selon le sujet principal de la requête pour déterminer s'il y avait un lien entre le sujet et la syntaxe de la question. Les catégories qui ont été choisies sont : *contact*, *date*, *divers*, *finance*, *invest* et *share*, elles ont été utilisées lors de l'évaluation pour analyser la performance de l'identification. Ces catégories peuvent aussi être utilisées pour identifier la méthodologie à considérer pour répondre aux questions et déterminer quelle information doit être extraite pour trouver une réponse.

3. Solution proposée

L'extraction des questions est réalisée dans un module de traitement de la langue, intégré dans le système d'ingénierie linguistique GATE de l'Université Sheffield (Cunningham *et al.*, 2002), utilisant à la fois de l'information syntaxique et lexicale. Le procédé de détection des questions est exécuté en une série d'étapes. Les premières étapes de traitement sont réalisées par une suite de modules provenant du système GATE, la détection des questions est réalisée par le module d'identification des questions, exécuté en deux étapes. La première étape est l'identification, par un *gazetteer*, des mots dans les courriels qui sont utilisés comme symboles terminaux dans la grammaire. La deuxième étape du traitement est l'identification de patrons de questions, encodée comme une cascade de transducteurs et exprimée dans le formalisme JAPE. Les deux composantes du module sont présentées sous la forme synthétisée d'une grammaire dans la figure 1. Chacune des règles d'identification est numérotée et écrite en *italique*, les symboles écrits en majuscules et en *italiques* sont des symboles non-terminaux et les symboles écrits en police `courrier` sont des terminaux.

Chaque règle a la tâche d'identifier un patron correspondant à une façon de poser une question dans un courriel. La question identifiée correspond à la région du courriel débutant par le début du patron et se terminant à un symbole de fin de phrase, déterminé par le module de segmentation de phrase. Nous décrivons maintenant les règles de la figure 1.

(1) Les patrons `I was wondering` ou `I wonder` de la règle *wonder* sont utilisés dans les courriels pour introduire poliment une question ou une requête formulée indirectement. L'information recherchée pour identifier précisément la requête ne suit pas immédiatement le patron recherché, le patron `if you` *modals action* se retrouve régulièrement entre le

(1) *wonder* → `I` (`wonder` | `was wondering`)
(2) *be_have_mod* → (*MODALS* | *TOBE* | *TOHAVE*)
(`it` | `there` | `these` | *EX*)[*DET*]
(3) *modalsBegin* → (*MODALS* | *TODO*)(*PRP* | *NNP*)
(4) *actionAsking* → *ACTION_ASKING* (`me` | `us`)
(5) *wh_be_have_do* → *WH_WORDS* (*TOBE* | *TOHAVE* | *TODO*)
(6) *Please* → `please` *ACTION_ASKING* (category=VB)
(7) *would_like_to* → [`I` | `we` | `he` | `she` | `and`]`would like to` (category=PRP)
(8) *wh* → *WH_WORDS* [category = PRP]

MODALS → `can` | `could` | `may` | `might` | `must` | `ought to` | `shall` | `should` | `will` | `would`
ACTION_ASKING → `advise` | `forward` | `provide` | `confirm` | `give` | `send` | `direct` | `let` | `tell`
WH_WORDS → `what` | `where` | `when` | `which` | `who` | `why` | `how`

Figure 1. Grammaire d'identification des patrons

patron recherché par la règle et l'information recherchée.

(2) La règle *be_have_mod* identifie des questions portant sur la confirmation d'une information incomplète ou sur la vérification de l'existence de quelque chose (compagnie, méthode de calcul, produit financier, ...) Ces questions sont compliquées à analyser car il faut déterminer l'information cherchée par l'auteur. La création de la réponse est encore plus complexe car elle dépend directement de la confirmation de la requête.

(3) Les questions identifiées par la règle *modalsBegin* sont difficilement catégorisables, le patron utilisé correspond à une syntaxe peu spécifique pour introduire une question. Cette règle pourrait d'une certaine façon être considérée comme un cas général de la règle *wonder*. Les réponses sont déterminées par l'action indiquée par le verbe suivant le patron.

(4) La règle *actionAsking* identifie les questions où le préposé doit exécuter une action. Cette règle sert aussi pour certaines formulations de requêtes similaires aux règles précédentes, mais où la partie d'introduction de la requête est absente ou pas assez fréquente pour être considérée. Les verbes d'actions utilisés pour identifier les requêtes sont : `advise`, `forward`, `provide`, `confirm`, `give`, `send`, `direct`, `let` et `tell`.

(5) La règle *wh_be_have_do* identifie les questions qui débutent par un *wh-words* suivi d'une forme des verbes *be*, *have* ou *do*. Ces questions sont parmi les plus communes, leur forme est généralement celle à laquelle on fait référence lorsque l'on considère le domaine des questions.

(6) La règle *Please* est très similaire à la règle *actionAsking*, la composante importante pour les

deux règles est le verbe énonçant l'action demandée. La différence avec la règle *actionAsking* est que le verbe d'action n'est pas nécessairement utilisé de façon transitive.

(7) Les requêtes identifiées par la règle *would_like_to* sont des questions où le but n'apparaît pas clairement. Le but de la question dépend du verbe suivant `to` dans le patron, celui-ci peut être *ask*, *confirm*, *inquire*, *know*, *attend*, ...

(8) La règle *wh* est réalisée pour récupérer l'ensemble des questions contenant un *WH_WORD* et qui n'ont pas été identifiées auparavant. Les questions identifiées n'ont pas de caractéristiques particulières, elles devront être traitées avec des typologies de questions similaires à ce qui se fait dans systèmes de question-réponse pour pouvoir être répondues.

Lors de l'identification des questions, les règles sont essayées à tour de rôle pour déterminer si le patron de la règle concorde avec le texte. Lorsqu'un patron de règle concorde, l'intervalle de texte débutant au premier mot du patron et se terminant à la fin de la phrase est identifié comme une question. Le mot suivant la fin de la question est ensuite utilisé pour débuter la recherche de la question suivante. L'utilisation de cette stratégie de recherche permet d'appliquer un ordre de priorité sur les règles, qui est l'ordre d'apparition des règles dans la grammaire.

4. Résultats

La méthode d'identification des questions réalise bien la tâche pour laquelle elle a été conçue. En tenant compte du contexte du courriel et du domaine traité, la gestion des relations avec les investisseurs, on constate que les résultats sont très encourageants pour la réalisation de la réponse automatisée aux courriels.

4.1. *Présentation des résultats*

Les premiers résultats sont ceux concernant l'identification des courriels contenant des questions ou des requêtes, ils sont présentés dans le tableau 1, lorsqu'un courriel est considéré comme étant bien identifié, ceci signifie qu'il contient au moins une question identifiée correctement. L'extracteur de questions réalise bien sa tâche, 81% des courriels contenant une ou des questions sont identifiés comme tel. L'identification est particulièrement efficace pour les catégories *date* et *divers*, où les courriels sont identifiés à 95% et 89% respectivement, sans avoir de difficultés avec une catégorie en particulier. L'explication de la bonne performance pour les catégories *date* et *divers* s'explique par le fait que dans la catégorie *date* les questions sont énoncées sous une forme conventionnelle ; dans la catégorie *divers* les questions sont plus faciles à identifier car les courriels sont assurément des questions, où le but est soit multiple, soit carrément hors-catégorie, mais où la syntaxe de la question se conforme aux attentes.

Catégorie	Nbr. courriels	Bien identifiés	Mal identifié
contact	15	11	4
date	24	23	1
divers	19	17	2
finance	61	50	11
invest	32	26	6
share	59	44	15
Total	210	171	39

Tableau 1. Évaluation de l'identification de la (des) question(s) pour chaque catégorie

Le tableau 2 présente la distribution de l'efficacité du module d'identification de questions. Par rapport aux résultats précédents, les données sont maintenant exprimées en fonction de la distribution de la bonne ou mauvaise identification des questions et non plus seulement par rapport à l'identification des courriels. Le taux de succès de l'identification est encore de 81%, même si cette fois on considère le nombre de questions bien identifiées par rapport au nombre total de questions dans le corpus. En examinant les résultats obtenus, on peut constater que les courriels concernant les questions financières de la compagnie ainsi que ceux ayant un lien avec le prix des actions semblent être plus difficiles à traiter. Ceci s'explique par la complexité des énoncés de questions et par l'utilisation du contexte du courriel pour déterminer les sens interrogatifs du message. La catégorie finance possède la plus grande densité de questions annotées par courriel, soit 112 questions réparties à l'intérieur de 61 courriels. C'est la densité du nombre de questions par courriel (≈ 2.35 questions/courriel) qui complique le traitement.

Catégorie	Nombre de questions	Bien identifiées	Mal identifiées	Identifiées à tort			Non identifiées
				normal	sig.	rép.	
contact	16	11	3	1	2	0	2
date	30	27	0	2	3	0	0
divers	49	30	0	2	0	4	12
finance	143	112	10	6	10	2	25
invest	40	41	5	4	1	0	4
share	74	65	4	3	5	0	12
Total	353	286	22	18	21	6	55

Tableau 2. Distribution des identifications par l'extracteur de questions pour les courriels contenant des questions

Les questions identifiées à tort se répartissent en trois catégories dans le traitement automatisé des courriels relativement à l'endroit où elles apparaissent dans le courriel.

1. La première catégorie de questions identifiées à tort est celle où le patron de question apparaît dans le corps principal du courriel, mais où ce patron n'est pas une question. La plupart du temps c'est un *WH_WORD* agissant comme un pronom relatif pour introduire une clause relative, qui n'est pas utilisé dans un contexte interrogatif. L'item (1.1) est un exemple où une question est mal identifiée, l'identification est réalisée par la règle *modalsBegin*.

   (1.1) Please do not hesitate to contact me **should you need clarification or have any inquiries.**

2. La deuxième catégorie de questions identifiées à tort est attribuable aux questions extraites dans la partie signature du courriel, cette catégorie est identifiée par sig. dans le tableau 3. Parfois la phrase extraite doit être considérée comme une question, mais la plupart du temps ce n'est pas le cas, l'identification du patron ne correspond pas à une question où le patron est une formule toute faite du type *Do you Yahoo ! ?*, *Where do you want to go today ?* ou tout autre slogan publicitaire énoncé comme une question.
3. La troisième catégorie de questions qui devrait être ignorée lors de l'identification est celle où la question apparait dans la partie retour (*reply*) d'un courriel (rép.). Ces questions doivent être traitées de façon très particulière, quelques fois elles proviendront d'un courriel redirigé qui doit être répondu, tandis qu'à d'autres occasions ce sera seulement une information qui a suivi le fil d'une *conversation* par courriel.

Le tableau 3 présente pour chacune des catégories de courriel le nombre de questions qui sont annotées par chacune des règles. La distribution des patrons de questions n'est pas uniforme, en fait 270 des 373 questions annotées du corpus le sont avec seulement deux règles (*modalsBegin* et *wh-words*). De plus les règles *wonder* et *actionAsking* ne sont utilisées que pour les catégories de courriels *finance* et *share*, les autres règles semblent être activées de façon uniforme relativement à la quantité de questions des catégories de courriels. Ces résultats mettent en évidence une différence importante entre le traitement automatisé des courriels et les systèmes de questions-réponses de type TREC, les questions ne sont pas toujours introduites par un *WH_WORD* et la syntaxe peut être très diversifiée.

Catégorie	*wh*	*modalsBegin*	*Please*	*actionAsking*	*would_like_to*	*be_have_mod*	*wonder*	*wh_be_have_do*	Total
contact	4	7	2	1 sig.	3	2	0	0	19
date	8	22	5	1 sig.	2	0	0	0	38
divers	9	15	3	0	0	7	0	0	34
finance	87	32	4	13	1	4	0	0	147
invest	15	22	6	0	5	3	6	0	51
share	18	31	15	9	7	2	2	0	84
Total	141	129	35	24	18	18	8	0	373

Tableau 3. Distribution des patrons de questions retrouvées dans chacune des catégories

4.2. *Questions non identifiées*

Selon les résultats présentés auparavant, la grammaire d'extraction réalise bien la tâche pour laquelle elle a été construite, par contre elle ne couvre pas tous les cas de questions qui se retrouvent dans le corpus. Les résultats du tableau 2 indiquent que la grammaire n'a pas été en mesure d'identifier 55 questions, ce qui correspond à 15% du nombre total de questions.

Le nombre de questions non identifiées ne doit pas être considéré comme une contre-performance de la méthode d'identification des questions car il est possible d'en expliquer la provenance. Un examen plus approfondi des données montre qu'il y a cinq courriels dans le corpus qui contiennent près de la moitié des questions non-identifiées (26 questions). Ces cinq courriels ne sont pas représentatifs des données du corpus, ils ont une densité de questions très élevée et ils utilisent le formattage du texte et non la syntaxe pour distinguer les questions. Les autres questions non-identifiées sont dues à deux facteurs, le premier est la faible densité des patrons utilisés pour les questions non identifiées ; le deuxième est la présence de structures grammaticales incorrectes, de coquille et de phrases difficiles à comprendre.

5. Conclusion

L'identification des questions dans les courriels est une étape cruciale pour aborder le problème du traitement automatisé des courriels d'un point de vue similaire à celui adopté pour le problème de la question-réponse. Le processus d'identification des courriels peut être considéré comme une étape de classification pour déterminer les courriels qui pourront être traités automatiquement et un préambule à une catégorisation des questions selon d'autres critères comme le sujet

de la question ou le type attendu de la réponse (prix, date, endroit).

L'approche par patron de surface est celle qui semble être la plus adaptée à notre problème. Il était impossible d'utiliser des méthodes d'apprentissage pour réaliser l'identification des questions car la quantité de données n'est pas assez grande et le corpus de courriel n'est pas assez diversifié pour trouver des facteurs discriminants assez forts. Les résultats obtenus par le procédé d'identification de questions sont très bons si on tient compte des facteurs d'erreurs mentionnés précédemment. Le taux de rappel des questions pourrait être augmenté par l'ajout de règles et la spécialisation des règles existantes, ceci aurait par contre comme effet de diminuer la précision de l'identification.

La réalisation de l'identification est une des étapes d'un procédé pour répondre automatiquement aux courriels envoyés au service de relations avec les investisseurs. Le but du projet étant de traiter les courriels de leur réception jusqu'à la réponse, les questions identifiées devront être analysées pour ensuite être traitées par un système de question-réponse. Le système de question-réponse devra tenir compte du contexte de la question et de l'information extraite lors de l'analyse du courriel, pour faire suivre la question à un module spécialisé qui aura la tâche de trouver l'information pertinente à la rédaction d'une réponse au courriel.

Références

Banter Inc. (2001). Natural Language Engines for Advanced Customer Interaction. [http ://www.realmarket.com/required/banter1.pdf].

Cunningham H., Maynard D., Bontcheva K. et Tablan V. (2002). GATE : A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Hermjakob U. (2001). Parsing and Question Classification for Question Answering. In *Proceedings of the Association for Computational Linguistics 2001 Workshop on Open-Domain Question Answering* : 17-22.

Jupiter Communications (2000). *E-mail Customer Service : Taking Control of Rising Customer Demand*.

Voorhees E.M. (2001). Overview of the TREC 2001 Question Answering Track. In *Proceedings of the Text REtrieval Conference*.

Watanabe Y., Yokomizo K. et Okada Y. (2003). A Question Answer System Using Mail Posted to a Mailing List. In *Proceedings of the PACLING'03* : 335-342.