Sponsorisée par

TALN 2001 Récital

Tome 1

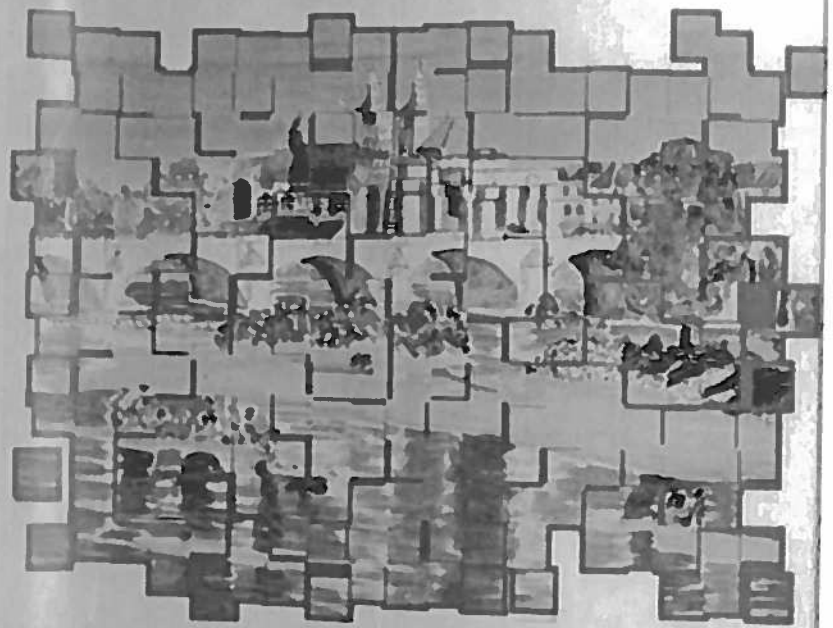
TALN

8e conférence sur le Traitement Automatique des Langues Naturelles

2001

8e rencontre des étudiants chercheurs en informatique pour le Traitement Automatique des la

Récital



Tours, 2-5 juillet 2001

Tome 1

sous l'égide de
l'association pour le Traitement Automatique des LAngues

organisée par le Laboratoire d'Informatique
et l'axe Langues et Technologies de l'Information
de l'Université de Tours

Critères de sélection d'une approche pour le suivi automatique du courriel

Leila Kosseim et Guy Lapalme
RALI, DIRO, Université de Montréal
{kosseim, lapalme}@iro.umontreal.ca

Résumé - Abstract

Cet article discute de différentes approches pour faire le suivi automatique du courrier-électronique. Nous présentons tout d'abord les méthodes de traitement automatique de la langue (TAL) les plus utilisées pour cette tâche, puis un ensemble de critères influençant le choix d'une approche. Ces critères ont été développés grâce à une étude de cas sur un corpus fourni par Bell Canada Entreprises. Avec notre corpus, il est apparu que si aucune méthode n'est complètement satisfaisante par elle-même, une approche combinée semble beaucoup plus prometteuse.

This paper discusses the strengths and weaknesses of different NLP techniques applied to the automatic follow-up of e-mail. First, we present the most widely-used methods, then we present a set of criteria for selecting the method to be used in particular circumstances. These criteria have been developed through a case study of a corpus provided by Bell Canada Enterprises. With our corpus, it appears that although no individual method is completely satisfactory, an approach combining several techniques is more appropriate.

Mots-Clefs: Réponse au courriel, Analyse de texte, Analyse de corpus.

1 Approches de TAL au suivi du courriel

Le courrier électronique est le mode de communication grandissant le plus rapidement aujourd'hui. Cependant, les entreprises ne sont généralement pas en mesure de faire face à un tel accroissement de demande de service à la clientèle par ce médium. Il devient donc intéressant d'explorer l'utilisation des outils de TAL pour faire le suivi automatique des courriels.

Les approches actuellement utilisées pour faire le suivi du courriel peuvent être classées en deux catégories: des approches *statiques* et des approches *dynamiques*. Les approches statiques utilisent des bases de *paires message-suivi*[1] préparées spécifiquement pour la tâche. Ces approches sont dites statiques, car elles doivent être réadaptées à chaque situation et les suivis restent fixes. De leur côté, les approches dynamiques utilisent des bases documentaires préexistantes telles que des listes de foires aux questions (FAQ) ou des sites Web. Ces systèmes ont l'avantage d'être adaptables à de nouvelles situations car la réponse est calculée dynamiquement.

[1]Nous utilisons le terme *message* plutôt que *question*, car un courriel peut transmettre différents buts communicatifs (ex. une question, une plainte, ...). Aussi, nous utilisons le terme *suivi* plutôt que *réponse*, car un courriel peut nécessiter différentes réactions de la part du récepteur (ex. une réponse, un classement ...).

Parmi les approches statiques, on retrouve la classification de texte, le raisonnement par cas et l'extraction d'information.

La classification de texte est appropriée lorsque les messages sont assez longs et stéréotypés. L'hypothèse de base de cette approche est que la majorité des messages sont appariés à un petit nombre de suivis. En catégorisant ces messages selon leur suivi, la majorité des messages peuvent être traités automatiquement soit en les transferrant au département approprié soit en leur associant une réponse pré-rédigée (fixe ou paramétrisée). De nombreux travaux ont porté sur la classification de texte, s'appuyant sur des techniques d'apprentissage automatique (Coh96; LF00; BSA00) ou sur des approches descriptives plus linguistiques (CLM+99).

L'extraction d'information (EI) consiste à identifier de l'information bien précise à partir d'un document d'un genre et d'un domaine bien spécifiques et à la représenter sous forme structurée (MUC98). Par exemple, à partir d'un rapport sur un accident automobile, un système d'extraction d'information sera capable d'identifier la date et le lieu de l'accident ainsi que le nom des victimes. Pour le traitement des courriels, lorsque le domaine de discours est restreint, l'extraction de scénarios peut être utilisé pour représenter le contenu du courriel (KBL00). Aussi, l'extraction d'entités nommées (noms de personnes, dates, ...) permet de paramétriser le contenu du message et ainsi personnaliser le suivi en fonction des données spécifiques retrouvées dans le message.

Le raisonnement par cas (RPC). À notre connaissance, aucun projet ne s'est penché sur le raisonnement par cas pour la réponse au courriel, pourtant cette approche semble tout à fait naturelle. En effet, en utilisant une base de cas formée de paires message-suivi déjà utilisées, on peut calculer la similarité entre un nouveau message et les messages déjà reçus et renvoyer la même réponse. Cette tâche est apparentée au travail de (CT98) sur la génération de texte et au travail de (BHK+97) sur la réponse à des questions fréquentes.

Parmi les approches dynamiques, on retrouve la recherche d'information et la question-réponse.

La recherche d'information (RI) considère le message comme une requête à faire correspondre à une base documentaire. Faire le suivi du courriel revient alors à identifier et renvoyer comme réponse le document ou le passage de la base documentaire qui ressemble le plus au message (ex. (Exp00)). Cette approche a l'avantage d'être indépendante du domaine de discours, mais n'offre pas une interaction naturelle ou personnelle. De plus, elle suppose que le message est assez court pour que le processus soit calculable efficacement.

La question-réponse (QR) est appropriée pour des questions courtes et explicites qui demandent une réponse factuelle qui se retrouve dans une base documentaire (ex. *Combien de succursales avez-vous en Amérique du Nord?*). Les systèmes de question-réponse effectuent généralement une recherche d'information à partir des termes de la question pour identifier les passages de la base documentaire les plus susceptibles de contenir la réponse. Ces passages sont ensuite analysés pour extraire des entités sémantiques particulières (noms de personnes, dates, ...) et, en fonction du type de réponse recherchée (*qui?*, *quand?*, ...), l'entité la plus probable est extraite (TRE99; TRE00).

2 Critères de sélection

Pour sélectionner l'approche la plus prometteuse dans un contexte particulier, nous avons identifié huit critères portant sur le message à traiter, le suivi et la paire message-suivi. L'influence de ces critères est illustré à figure 1.

Critère 1: La généricité du contenu du message qualifie jusqu'à quel point le message porte sur une situation générale (ex. *Quelle est votre politique de remboursement?*) ou sur un cas particulier mettant en jeu des entités spécifiques (ex. *Quel était l'état du compte no 123 le 1 janvier 1999?*). Les messages génériques peuvent être traités par toutes les approches. Les messages spécifiques, plus délicats à traités ont fait l'objet de peu de travaux. Ces messages peuvent difficilement être traités par une approche dynamique car la réponse ne se trouvera généralement pas directement dans une base documentaire existante. Les approches statiques, elles, peuvent être utilisées pour les messages spécifiques si les paires message-suivi d'entraînement sont paramétrisées et si les entités nécessaires sont extraites du message.

Critère 2: La longueur et la clarté des messages influencent le choix d'une approche. En effet, la majorité des approches n'effectuent pas d'analyse linguistique du message, mais considèrent plutôt le texte comme un vecteur d'attributs (souvent la présence de termes). Cette technique permet une analyse plus robuste lorsque le message est complexe ou contient du bruit (grammaire de variantes, fautes d'orthographes . . .) (ex (LF00; BSA00)), mais en revanche, elle est plus sensible lorsque le message est court. De leur côté, les systèmes de question-réponse tentent une analyse sémantique pour identifier le but de la question (ex. *quand?* $\Rightarrow$ *une date*) et la relation entre les termes de la question. À l'heure actuelle, les systèmes de QR ne sont donc appropriés que lorsque le message est court et clair (TRE99; TRE00).

Critère 3: Le type de suivi qualifie *comment* un courriel doit être traité: par une réponse, par un classement (incluant un transfert à autrui) ou par une action physique[2] (ex. envoyer un document par la poste). Les actions physiques nécessite généralement une intervention humaine et ne seront donc plus discutés ici. Une approche statique peut tout aussi bien classer que répondre, alors qu'une approche dynamique n'est pas naturellement adaptée à classer des textes car elle est basé sur une recherche textuelle.

Critère 4: La fréquence des suivis tient compte du nombre de suivis différents envoyés par rapport au nombre de messages reçus. Puisque les approches statiques utilisent des paires message-suivi pré-établies, si chaque suivi n'est utilisé que rarement, ce corpus d'entraînement sera difficilement complet et le taux de succès sera assez faible. Dans ce cas, seul un système dynamique est envisageable, car les questions ne doivent pas être planifiées.

Critère 5: La nature de la réponse s'intéresse au degré d'argumentation de la réponse: de la réponse argumentative (ex. une explication, une définition) à la réponse factuelle. La question-réponse, par exemple, est basée sur l'identification d'entités sémantiques précises (lieux, dates, . . .); elle permet donc d'identifier des faits, mais actuellement n'est pas en mesure de composer un texte argumentatif. Les autres approches, par contre, sont adaptées à toutes sortes de réponses, car celles-ci sont pré-rédigées.

Critère 6: L'ergonomie de l'interaction détermine jusqu'à quel point l'envoi d'une réponse décousue (comme un extrait d'une FAQ) est acceptable ou si une réponse bien rédigée est nécessaire. Dans ce dernier cas, la réponse peut être personnalisée: *à l'envoyeur* – le contenu du message est fixe, mais les salutations sont personnalisées; *à la situation* – le contenu du message est personnalisé en utilisant les entités et les relations retrouvées dans le message; *à l'historique* – le contenu de la réponse est influencé par les échanges précédents avec ce même envoyeur.

Critère 7: La stabilité temporelle des paires message-suivi mesure si les paires message-suivi sont variables au fil du temps (ex. *Pourquoi votre site Web est-il inaccessible?*). Si les

[2] Notons, que ces trois types de suivi ne sont pas mutuellement exclusifs.

paires sont stables, un système fixe est suffisant; sinon, le système devra pouvoir s'adapter à de nouvelles paires message-suivi. Les approches dynamiques sont adaptables par définition car seule la base textuelle indépendante doit être mise-à-jour, et ce de façon transparente au système. En revanche, les systèmes utilisant une approche statique nécessitent des modifications (mise-à-jour de leur propre base ou ré-entraînement).

Critère 8: L'étendue du domaine de discours influence essentiellement l'extraction d'information. En effet, l'extraction de scénarios se base sur la recherche de termes et de constructions syntaxiques pré-identifiées et l'application de règles sémantiques et discursives particulières au domaine de discours (MUC98). Si les messages font référence à plusieurs domaines de discours, l'extraction de scénarios ne peut être utilisée.

3 Étude de cas

Nous avons étudié un corpus de 865 paires de message-suivi (anglais) fourni par Bell Canada Entreprises (BCE) afin de situer le corpus par rapport aux critères de la figure 1 et faire ressortir l'approche la plus appropriée pour traiter les courriels de BCE.

Critère 1: La généricité du contenu. Le langage utilisé dans les courriels a fait l'objet de nombreuses études linguistiques. En particulier, le travail de (CB96), basée sur le modèle MD-MF de (Bib88), a comparé différentes dimentions textuelles des courriels en comparant la présence de certains traits linguistiques dans ces textes et dans des textes de genres textuels différents (entrevus, lettre personnelles, ...). Inspirés de ce travail, nous avons tenté de mesurer la dimention de généricité de notre corpus en utilisant 3 traits linguistiques: 1) Le taux de pronoms et de déterminants à la première personne par rapport au nombre total de pronoms et déterminants. 2) Le taux de noms propres par rapport au nombre total de noms. 3) Le taux de numéraux par rapport au nombres total de mots. À titre de référence, nous avons comparé ces taux à ceux de deux groupes de news: un extrait de la liste corpora (COR) et dogfriendly.com (DOG)[3]. Le taux de pronoms personnels semble indiquer que les courriels de BCE sont plutôt spécifiques (BCE: 31.6%; DOG: 32.3%; COR: 11.2%), mais les autres mesures ne sont pas aussi déterminantes (BCE: 24.1% – 4.5%; DOG: 19.3% – 4.6%; COR: 36.4% – 6.0%). Nous avons donc situé notre corpus au centre de l'axe du critère 1 (*cf* la figure 1).

Critère 2a: La longueur des messages a été évaluée par le nombre de mots: 13% des messages comptent moins de 20 mots; 70%, de 20 à 100 mots; et 17%, plus de 100 mots. Toutes les approches semblent donc utilisables pour la majorité des messages, mais la question-réponse n'est pas du tout adaptée pour 17% des messages (*cf* la figure 1).

Critère 2b: La clarté des messages a été évaluée par deux mesures: la longueur moyenne des phrases (en faisant l'hypothèse que les messages a phrases courtes sont plus clairs) et la densité de conjonctions, d'adjectifs et d'adverbes (en faisant l'hypothèse que les messages a syntaxiquement complexes sont moins clairs). Le bruit textuel (ex. fautes d'orthographes, reprises, ...) fera l'objet d'une prochaine étude. Par référence, nous avons comparé ces mesures avec les questions de la piste QA de TREC-8 et TREC-9 (TRE99; TRE00), la liste corpora (COR) et dogfriendly.com (DOG). Dans BCE, les phrases sont en moyenne composées de 13 mots. Par contre, si on sépare les messages courts (de moins de 20 mots) des messages longs

[3]COR contient environ 260 Ko de messages de chercheurs en linguistique-informatique. Ces messages sont quelquesfois génériques (ex. demande d'information sur un thème) et quelquesfois bien spécifiques (ex. annonces de conférences). DOG contient environ 175 Ko de messages d'amateurs de chiens décrivant leur expériences avec leur animal. Ces messages sont plutôt spécifiques.

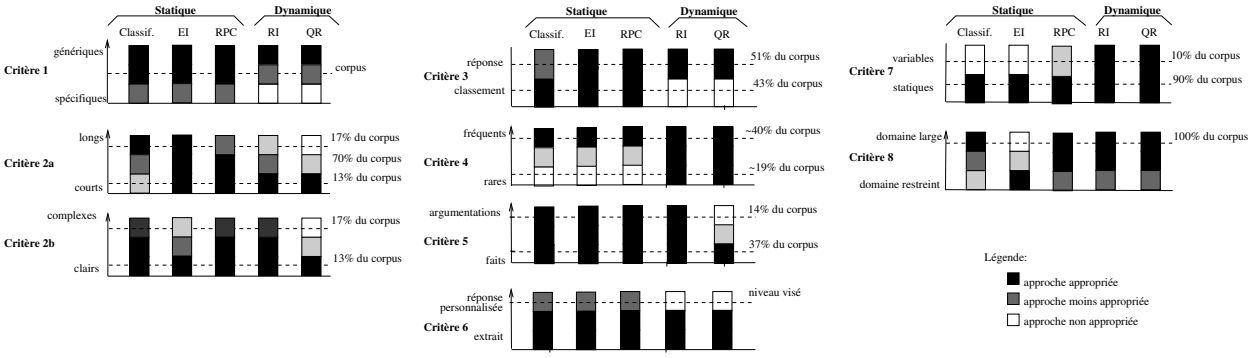


Figure 1: Critères de sélection et leur influence sur le corpus de BCE

(plus de 100 mots), on remarque que les messages courts ont des caractéristiques semblables aux questions de TREC (BCE-courts: 5,7 mots, densité = 4%; TREC: 7,6 mots, densité = 6%), mais très différents de COR (14,6 mots, densité = 10%) et de DOG (13,8 mots, densité = 12%); alors que les messages longs de BCE ont des caractéristiques inverses (BCE-longs: 15,6 mots, densité = 12%). Les messages courts de BCE utilisent donc des structures plus simples (d'où les axes pointillés de la figure 1).

Critère 3: Le type de suivi compte 43% de classements et 51% de réponses.

Critère 4: La fréquence des suivis est élevée dans le corpus de BCE. En effet, même s'il existe un grand nombre de suivis différents, seul un petit nombre sont utilisés fréquemment. Parmi les 342 classements, il y a 111 destinataires différents, mais 5 d'entre eux ont reçu 44% des messages. En ce qui concerne les réponses, un seul type de réponse compte pour 37% des cas[4] et 6 types de réponses comptent pour 58% des cas. Cependant, mis à part ces suivis très fréquents, la majorité des suivis sont rarement utilisés. En effet, 15% des réponses et 22% des classements sont uniques.

Critère 5: La nature des réponses a été évaluée par la longueur des réponses en faisant l'hypothèse qu'un fait est court et qu'une argumentation est plus longue. 37% des réponses sont composées d'une seule phrase, donc facilement assimilables à des faits; 31% des réponses sont composées de 2 phrases; 18%, de 3 phrases et 14% des réponses ont plus de 3 phrases, donc, plus près de réponses argumentatives. Ceci est illustré par les axes pointillés de la figure 1.

Critère 6: L'ergonomie de l'interaction telle qu'utilisé actuellement par BCE a peu d'importance. Nous nous fixons de faire de la personnalisation à la situation autant que possible.

Critère 7: La stabilité temporelle des paires message-suivi est variable dans le corpus de BCE. Certaines paires sont stables au fil du temps, alors que d'autres sont variables. Par exemple, l'entreprise a reçu en 5 jours seulement, près de 85 messages (10% du corpus) à propos de la division d'une de ses actions, et à l'extérieur de cette période, aucun message n'avait rapport à cet événement.

Critère 8: L'étendue du domaine de discours est assez variée. En effet, les messages proviennent principalement du site corporatif de BCE, qui présente de l'information relative à l'ensemble des entreprises du groupe BCE. Les messages portent sur 8 domaines différents et vastes (des

[4] la réponse du type *Voir <URL>*

questions financières, de l'information générale, le service de téléphone, le service Internet, . . .). Cette grande variété de domaines de discours fait en sorte qu'une approche par extraction de scénario n'est pas appropriée.

L'analyse précédente résumée à la figure 1) montre qu'aucune approche n'est parfaitement adaptée à notre corpus. L'extraction de scénarios n'est pas du tout applicable selon le critère 8, mais les autres approches peuvent être utilisées avec plus ou moins de succès selon le critère d'intérêt. Suite à cette observation, une approche complémentaire semble plus appropriée. Lors de la réception d'un message, un module d'extraction d'entités nommées étiquettera le message pour y identifier des expressions sémantiques particulières. Ensuite, les différentes méthodes traiteront le message étiqueté de façon indépendante. Finalement, un dernier module combinera les résultats des modules précédents en respectant les critères identifiés précédemment pour donner plus de poids à une approche plutôt qu'une autre et ainsi produire le suivi le plus approprié.

4 Conclusion

Pour faire le suivi automatique du courriel, plusieurs approches sont disponibles; chacune appropriée à des messages et des suivis différents. Dans cet article, nous avons défini huit critères de sélection permettant de choisir l'approche la plus appropriée. Nous avons ensuite appliqué ces critères de sélection sur un corpus de messages de Bell Canada Entreprises. Suite à cette étude, nous proposons l'utilisation d'approches complémentaires pour pouvoir profiter des forces de chaque approche et augmenter la confiance dans le suivi.

Remerciements Nous tenons à remercier les rapporteurs pour leurs commentaires et M. Robillard, Communications corporatives de BCE. Ce projet est financé par les Laboratoires universitaires Bell et le CRSNG.

Références

[Bib88] D. Biber (1988) *Variation across speech and writing*. Cambridge University Press.

[BHK+97] R. Burke et al. (1997) Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System. *AI Magazine*, 18(2):57–66.

[BSA00] S. Busemann, S. Schmeier, et R. Arens. (2000) Message Classification in the Call Center. Dans *Proceedings of ANLP-2000*, pp. 159–165, Seattle.

[CLM+99] F. Ciravegna, et al. (1999) Facile: Classifying texts integrating pattern matching and information extraction. Dans *Proceeding of IJCAI-99*, pp. 890–895, Stockholm, Sweden.

[Coh96] W. Cohen. Learning rules that classify e-mail. Dans *Proceeding of the 1996 AAAI Spring Symposium on Machine Learning in Information Access*, 1996.

[CB96] M. Collot et N. Belmore. (1998) Electronic Language: A New Variety of English. Dans S. Herring, editeur, *Computer-Mediated Communications: Linguistic, Social and Cross-Cultural Perspertives*, pp. 13–28, John Benjamins.

[CT98] L. Capus et N. Tourigny. (1998) Générer de façon automatique des résumés grâce à des expériences similaires. Dans *Actes de la Conférence internationale sur le traitement des langues naturelles et ses applications industrielles*, pp. 183–187, Moncton, Canada.

[Exp00] (2000) http://www.islanddata.com/ site visité en décembre.

[KBL00] L. Kosseim, S. Beauregard et G. Lapalme. (2000) Using information extraction and natural language generation to answer e-mail. Dans *Proceedings of NLDB'2000*, Versailles, France.

[LF00] Y. Lallement et M. Fox. (2000) Interact: A Staged Approach to Customer Service Automation. Dans H. Hamilton et Q. Yang, editeurs, *Canadian AI 2000*, LNAI 1822, pp. 164–175. Springler-Verlag.

[MUC98] (1998) *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/muc_7_toc.html.

[TRE99] (1999) *Proceedings of the Text Retrieval Conference (TREC-8)*, Gaithersburg, Maryland.

[TRE00] (2000) *Proceedings of the Text Retrieval Conference (TREC-9)*, Gaithersburg, Maryland.