Catégorisation automatique du courriel en fonction de catégories pragmatiques

Inge Alberts, Dominic Forest, Suzanne Bertrand-Gastaldy

École de bibliothéconomie et des sciences de l'information, Université de Montréal – C.P. 6128, succursale Centre-ville, Montréal, QC, H3C 3J7 – Canada

Résumé

Cette communication expose les résultats d'une expérimentation visant la catégorisation automatique du courriel de cadres gouvernementaux et municipaux. Dans un premier temps, une typologie des patrons de catégorisation du courriel des cadres fut développée, grâce à une étude qualitative réalisée auprès de 34 participants. Inspirée de la théorie des actes de langage, cette typologie imite les comportements de lecture et de tri du courriel. Dans un deuxième temps, un corpus de 1703 messages fut élaboré, à partir d'un échantillon remis par deux cadres. Les traitements statistiques, qui ont fait appel à l'algorithme des KNN, démontrent que la catégorisation automatique basée sur le lexique des messages est beaucoup plus performante, lorsque K=2 et N=2000. Pour quatre catégories, le taux de rappel moyen est de 94,32%, le taux de précision moyen est de 94,50% et le taux d'exactitude (nombre de documents correctement catégorisés / nombre total de documents dans le corpus) est de 94,54%. Pour treize catégories, le taux de rappel moyen est de 91,09%, le taux de précision moyen est de 84,18% et le taux d'exactitude est de 88,70%. Également, il apparaît que les qualités non lexicales des messages sont profondément influencées par la pragmatique du courriel. Parmi les critères de caractérisation les plus pertinents, ce sont les indices relatifs au destinataire et à l'expéditeur qui arrivent en tête.

Abstract

This communication presents the results of an experiment aiming for the automatic categorization of email pertaining to municipal and governmental managers. This experiment first developed a typology based on email categorization patterns derived from a qualitative study of 34 participants. Inspired by the Speech Act Theory, this typology simulated the managers' practices while reading and sorting their email. A corpus of 1703 messages was then developed from a sample provided by two managers. A statistical analysis, which employed the KNN algorithm, revealed that automatic categorization based on the lexicon of messages is much more efficient when K = 2 and N = 2000. Four categories had an average recall rate of 94.32%, an average precision rate of 94.50% and an accuracy rate of 94.54%. Of the thirteen categories, the average recall rate was 91.09%, the average precision rate is 84.18% and the accuracy rate is 88.70%. Results also showed that the non lexical qualities of email messages are deeply influenced by the pragmatics of the email. Among the characterizing criteria, the most relevant are the cues related to the recipient and the sender.

Keywords: speech act, automatic categorization, email, KNN, supervised approach, pragmatic theories

1. Introduction

La surcharge induite par le courriel est un phénomène attesté. Les utilisateurs se plaignent du volume de messages qu'ils reçoivent, de la présence de messages non sollicités et du manque de fonctionnalités offertes par les systèmes de messagerie commerciaux (Alberts, 2009). Ces problèmes découlent principalement du fait que les systèmes ne mettent pas à disposition de

leurs utilisateurs des options élaborées de tri. Suivant l'objectif de proposer une solution de catégorisation automatique permettant à des cadres gouvernementaux et municipaux de filtrer plus efficacement leur courriel, cette communication explore une approche pragmatique faisant appel à la théorie des actes de langage.

2. Revue de la littérature pertinente

2.1. Pragmatique et actes de langage

Au cœur de la pragmatique, la théorie des actes de langage stipule que le texte permet de véhiculer un message porteur de sens tout en comportant des indications relatives au rôle adopté par le locuteur et les actes que ce locuteur tente d'accomplir. On distingue généralement trois types d'actes de langage permettant de distinguer ce qui est accompli au cours de l'énonciation : l'acte locutoire, l'acte illocutoire et l'acte perlocutoire (Austin, 1970).

L'acte locutoire est défini comme un acte de langage qui consiste simplement à produire des sons, organisés selon une certaine grammaire et possédant une certaine signification (Austin, 1970 : 181). Par essence, l'acte locutoire est également un acte illocutoire qui « produit quelque chose EN disant » et un acte perlocutoire qui « produit quelque chose 'PAR le fait' de dire » (Austin, 1970 : 181). Pour qu'un acte illocutoire soit accompli avec succès, un auditeur doit reconnaître la valeur que le locuteur entend donner à son énonciation ; c'est donc en vertu de conventions que les actes illocutoires sont reconnus comme tels, contrairement aux actes perlocutoires qui ne résultent pas de conventions, mais relèvent plutôt du domaine de l'implicite (Cervoni, 1987 : 118).

Parce que l'acte illocutoire est un acte conventionnel relativement prévisible, Searle (1972) a popularisé une typologie des valeurs illocutoires du langage. Cette typologie comporte cinq familles d'actes illocutoires : (1) déclaratifs ; (2) directifs ; (3) promissifs ; (4) expressifs ; (5) assertifs. Cette typologie a inspiré la recherche sur la catégorisation automatique du courriel.

2.2. Pragmatique et catégorisation automatique du courriel

Étant donné que le courriel est principalement utilisé pour négocier et déléguer des tâches, des chercheurs proposent d'utiliser la typologie des actes illocutoires afin de catégoriser automatiquement les messages selon les intentions de l'expéditeur (Cohen et al., 2004 ; Carvalho and Cohen, 2005 ; 2006 ; Goldstein and Sabin, 2006 ; Scerri et al., 2007). À cet effet, les auteurs s'inspirent généralement de la taxonomie des actes du courriel de Cohen et al. (2004).

Développée à partir des travaux de Winograd (1988) qui avait adapté la typologie des actes illocutoires de Searle au développement d'un outil de communication asynchrone, la typologie des actes du courriel inventorie les différentes fonctions ('purposes') du courriel. Elle consiste en deux principaux types d'actes (commissifs et directifs) auxquels sont associés des verbes (tels que demander ou proposer) et des objets (tels qu'un événement ou une opinion). Si la plupart des études qui adoptent cette approche visent le développement d'algorithmes performants pour détecter ou catégoriser des messages en fonction de leurs actes de langage, certaines études se démarquent par des apports théoriques novateurs.

Par exemple, Goldstein et Sabin (2006) proposent d'enrichir la typologie des actes du courriel de Cohen et al. (2004) en tenant compte de la dimension conversationnelle propre aux échanges asynchrones. Ils raffinent cette typologie en ajoutant des catégories telles que « répondre » et « répondre avec transmission de documents ». Dans un ordre d'idées similaire, Scerri et al. (2007)

proposent également d'enrichir la typologie des actes de courriel par une approche conversationnelle qui tiendrait compte de la dimension perlocutoire des actes de langage. Cette typologie modélise le flot conversationnel, en indiquant par exemple si l'acte de langage à l'origine d'une conversation initie ou continue une discussion existante. On y tient également compte de l'objet de l'acte de langage, qui peut porter sur une activité à l'extérieur des frontières du courriel (un événement ou une tâche) ou sur un objet propre à un message (une information ou une ressource documentaire).

Dans le cadre de cette présente recherche, ces avancées permettent d'envisager une catégorisation automatique du courriel plus performante, non seulement en fonction de l'intention de l'émetteur d'un message (acte illocutoire), mais aussi suivant d'autres facteurs contextuels susceptibles d'influencer les pratiques réelles de lecture et de tri.

3. Phase 1 : Étude des pratiques d'interaction avec le courriel

Avant de procéder à l'exercice de catégorisation automatique du courriel (phase 2), une étude qualitative fut menée (phase 1), afin de déceler l'ensemble des dimensions pragmatiques propres à la lecture et au tri des messages en contexte professionnel. La méthodologie ainsi que la typologie des actes du courriel résultant de cette analyse sont brièvement décrites.

3.1. Méthodologie

La phase qualitative s'est déroulée dans le contexte de deux projets de recherche ainsi que d'une thèse de doctorat. L'échantillon comprenait dix-sept cadres et de dix-sept secrétaires provenant de deux administrations publiques. Trois modes de collecte des données furent employés : l'entrevue semi-dirigée, le journal de bord et l'enquête cognitive. Il s'agissait d'étudier les différentes pratiques d'interaction avec les textes dans les environnements numériques de travail, particulièrement celles exécutées par l'entremise du courriel.

Trente-quatre entrevues semi-dirigées, trente journaux de bord, trente entrevues de clôture réalisées lors de la remise des journaux de bord ainsi que cinq enquêtes cognitives (verbalisation des pratiques de lecture et de tri du courriel) ont été colligées. L'analyse a fait appel à une stratégie de codage alliant les modes inductif et déductif.

3.2. Typologie des patrons de catégorisation du courriel

Au terme de l'analyse, les résultats ont révélé que certains facteurs contextuels (le délai, l'expéditeur, l'action à entreprendre, la position dans une conversation, l'objet et le sujet) soutiennent une catégorisation implicite des messages par le destinataire, qui lui associe différents niveaux de priorité et d'engagement (Alberts and Bertrand-Gastaldy, 2008). Cette catégorisation, qui facilite le processus de tri, permet à l'employé d'organiser plus efficacement ses activités de travail. En vue de préparer les traitements statistiques visant une catégorisation automatique du courriel des cadres, les catégories décrites au cours des entrevues et des enquêtes cognitives ont été précisées lors de l'analyse du corpus qui sera décrite à la section 4.1. Cette analyse a permis d'affiner les catégories en soutenant l'élaboration d'une typologie des patrons de catégorisation du courriel (Fig. 1).

La typologie présente quatre catégories de premier niveau et treize sous-catégories qui résument les patrons de catégorisation des cadres. Soulignons que ces patrons sont également adoptés par les secrétaires qui trient le courriel de leurs superviseurs. La typologie privilégie le point de vue du destinataire plutôt que celui de l'émetteur, avec des catégories qui tiennent compte

des facteurs contextuels importants lors du tri des messages. Elle soutient ainsi l'évaluation de l'engagement engendré par un message et l'attribution d'un niveau de priorité.

Les quatre catégories de premier niveau aident le destinataire à évaluer le degré d'engagement d'un message. La catégorie *action*, qui regroupe les courriels engageant le destinataire à effectuer une tâche, nécessite le plus haut degré d'engagement. La catégorie *réaction*, qui regroupe les réponses d'un expéditeur à une demande du destinataire, nécessite un niveau d'engagement moyen, une réponse pouvant engendrer une nouvelle action pour le destinataire. Les catégories *suivi de l'environnement direct* et *suivi de l'environnement indirect* regroupent les courriels à caractère informationnel, ne nécessitant pas d'action pour le destinataire. La distinction entre ces deux catégories se fonde sur une modélisation du réseau social du destinataire (le nom des employés sous sa supervision ainsi que ses supérieurs hiérarchiques). Le suivi direct rassemble les messages pour lesquels l'expéditeur appartient au réseau social du destinataire. Le suivi indirect rassemble les informations non ciblées et les communications générales (« spam »).

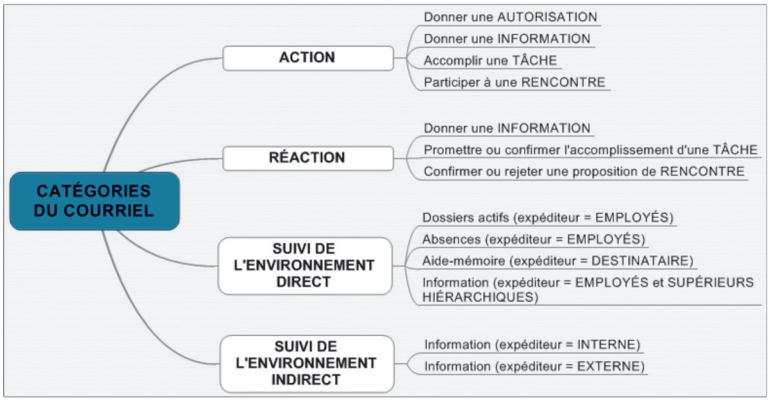

Figure 1 : Typologie des patrons de catégorisation du courriel

4. Phase 2 : Catégorisation automatique du courriel

Consécutive à la phase qualitative, la deuxième phase de la recherche a impliqué la mise au point de traitements statistiques visant une catégorisation automatique du courriel de cadres. Il s'agissait d'abord d'évaluer s'il est possible de prédire une catégorisation *a priori* des courriels en fonction d'indices lexicaux, mais aussi d'indices non lexicaux propres à la typologie des patrons de catégorisation élaborée durant la première phase. Il s'agissait également de déterminer parmi les indices non lexicaux qui caractérisent chaque catégorie de cette typologie, quels sont les plus discriminants. Dans le présent article, nous nous sommes limités à discuter principalement les résultats obtenus lors des expérimentations qui furent menées en utilisant les indices de catégorisation (c'est-à-dire les traits discriminants) de nature non lexicale. Nous les comparons très sommairement aux résultats obtenus lors des expérimentations employant les traits discriminants de nature lexicale. Une présentation détaillée des expérimentations utilisant les traits discriminants lexicaux pourra être consultée prochainement dans Alberts et Forest (en préparation).

4.1. Constitution du corpus et prétraitement

Un corpus fut élaboré à l'aide du courriel de deux cadres gouvernementaux. Le premier échantillon (cadre 1) comprenait originellement 591 messages et le deuxième (cadre 2), 1615 messages.

4.1.1. Analyse et catégorisation manuelle

La première étape a consisté à analyser l'ensemble des messages remis par les cadres afin de les catégoriser en fonction de la typologie (Fig. 1). Au total, 503 messages furent retirés du corpus final : 313 messages de langue anglaise, 117 messages pouvant appartenir à plus d'une catégorie, 15 messages personnels (peu représentés dans le corpus), 16 confirmations de lecture automatiques, 25 messages d'avertissement que la taille limite de stockage est dépassée et 17 messages reçus en double. Même si la présence de ces messages reflètent la réalité et que leur suppression peut constituer une source de biais, la décision fut néanmoins prise de les retirer du corpus afin de limiter le risque d'erreurs lors de cette première expérimentation. À cette étape, les 1703 messages restants ont été renommés avec l'étiquette de leur catégorie d'attribution suivie d'un numéro séquentiel et du code du répondant.

4.1.2. Profilage manuel (constitution du corpus de référence)

Afin que l'algorithme de catégorisation puisse tenir compte des indices non lexicaux lui permettant de prédire à quelle catégorie un message appartient, chaque message a dû être décrit manuellement dans une *matrice non lexicale*. La matrice utilisée à cet effet comportait donc des indices non lexicaux, choisis en fonction des résultats de la phase qualitative, ainsi que la catégorie manuellement attribuée à chaque message (à des fins d'apprentissage).

Par exemple, on retrouve dans la matrice une description de chaque courriel en fonction d'indices visuels (utilisation de caractères gras, de majuscules et d'images insérées), d'informations structurelles (nombre de destinataires dans le champ À, emplacement du destinataire dans le champ À ou CC, présence des abréviations RE, FW et TR dans le champ Objet), de caractéristiques des pièces jointes (pièce jointe directement au message ou présente dans le fil de discussion) et d'informations contextuelles (appartenance de l'expéditeur au réseau social du destinataire, présence d'une signature officielle ou utilisation du prénom).

4.1.3. Filtrage en vue de la catégorisation non lexicale

Une fois la matrice complétée, les messages en format propriétaire ont été enregistrés en format texte. Afin de limiter les sources de bruit au cours du processus de catégorisation automatique, les décisions suivantes ont été prises à cette étape : suppression des fils de discussion, suppression de la portion anglaise des messages, suppression des pièces jointes et suppression des signatures. Des attributs relatifs à la présence ou à l'absence d'un fil de discussion, à la présence de messages bilingues, à la présence ou à l'absence de pièce jointe, à la position de la pièce jointe dans un fil de discussion (externe au message ou imbriquée dans un message) ainsi que des attributs relatifs à la présence et à la nature d'une signature (signature officielle par opposition à une signature informelle) ont été consignés dans la matrice.

4.1.4. Filtrage en vue de la catégorisation lexicale

L'étape du nettoyage automatique a porté sur le lexique du corpus à analyser. Il s'agissait ici de lui appliquer différents filtres statistiques et linguistiques dans le but d'augmenter la qualité générale des analyses (Forest, 2006 : 37). Le filtrage du lexique a été effectué à l'aide de l'application commerciale de fouille de textes Wordstat de Provalis Research (version 5.1.10). À cette étape, trois opérations ont été effectuées sur le lexique : suppression des mots fonctionnels, lemmatisation et application d'un filtre statistique sur les mots dont la fréquence dans chacun des messages était inférieure à 2 et ceux figurant dans plus de 25% des messages (seuils déterminés empiriquement).

Au terme des étapes de prétraitement, nous disposions d'un corpus de 1.703 messages individuels en format texte qui ont servi à la mise au point des traitements automatiques. Le lexique filtré comporte 5.051 mots, ce qui constitue environ 55% du lexique d'origine (c'est-à-dire, 160.410 occurrences et 9.262 formes).

4.2. Format des données soumises aux différents traitements

Dans le cadre de cette expérimentation, les messages soumis aux différents traitements statistiques se présentent sous forme de matrices de vecteurs. Puisque les analyses statistiques visent non seulement la catégorisation automatique des messages, mais également l'identification des indices non lexicaux les plus discriminants pour chaque catégorie de messages, deux matrices furent utilisées au cours des traitements :

a) Une *matrice non lexicale* comportant 22 indices ("features") choisis sur la base des résultats de la phase qualitative. Cette matrice est construite en indiquant l'absence (valeur = 0) ou la présence (valeur = 1) de chacun des traits caractéristiques dans chacun des 1703 messages. Certaines valeurs (par exemple, le nombre de destinataires dans le champ CC) ont également été pondérées (telles que 1, pour un seul destinataire, 2 pour deux destinataires et 3 pour plus de trois destinataires). La matrice non lexicale comprend donc 1703 X 22 vecteurs.

b) Une *matrice lexicale* comportant les 5051 mots ("features") du lexique filtré, pondérée en fonction de la fréquence d'apparition du lexique dans le corpus à l'étude. La taille de la matrice créée est donc de 1703 (courriels) x 5051 (nombre maximal de traits discriminants).

4.3. Paramètres des expérimentations

4.3.1. Catégorisation automatique

La première série d'expérimentations a porté sur les indices non lexicaux propres au courriel, tels que décrits dans la *matrice non lexicale*. La catégorisation automatique et le repérage des indices les plus discriminants furent effectués à l'aide du logiciel de fouille de données Tanagra (version 1.4.28) développé par l'Equipe de Recherche en Ingénierie des Connaissances de l'Université Lumière Lyon 2.

Après plusieurs explorations statistiques, il s'est avéré ici que l'algorithme des KNN ("k nearest neighbours", ou k plus proches voisins) est le plus performant pour catégoriser les messages en fonction des treize catégories de la taxonomie. Cette méthode d'apprentissage supervisé repose sur l'hypothèse théorique que des messages semblables devraient appartenir à des catégories semblables. Il s'agit ainsi de définir un « voisinage » (nombre de K) autour du message à classer et estimer localement les probabilités d'appartenance à une catégorie, K=1 étant le plus proche voisin.

Lors de nos expérimentations, nous avons fait varier à la fois le nombre de voisins K autour du message à classer et le nombre de traits à considérer (ici, les 22 traits non lexicaux). Ainsi, plusieurs expérimentations ont été effectuées à l'aide de ces paramètres afin d'obtenir une catégorisation automatique optimale.

La deuxième série d'expérimentations a porté sur le lexique des messages. La catégorisation automatique fut effectuée à l'aide du logiciel Wordstat de Provalis Research (version 5.1.10). Après plusieurs explorations statistiques, il s'est également avéré que l'algorithme des KNN est le plus performant pour catégoriser automatiquement les messages sur une base lexicale. Afin d'assurer la qualité de la procédure de catégorisation, nous avons adopté une méthode d'échantillonnage reposant sur le principe de la validation croisée. Dans notre expérimentation,

nous avons eu recours à 10 groupes ; l'apprentissage s'est effectué sur 9 groupes, le taux d'erreur étant calculé sur le groupe restant ("10-fold cross-validation").

4.3.2. Sélection des indices non lexicaux les plus discriminants

Afin de déterminer quels sont les indices non lexicaux les plus pertinents lorsqu'il s'agit de caractériser les catégories dans leur ensemble, nous avons utilisé une fonction de « feature ranking », laquelle permet d'identifier les traits discriminants les plus influents lors du processus de catégorisation. Cette fonction, qui mesure le degré d'association entre deux variables nominales, permet ici d'évaluer quels indices non lexicaux sont davantage discriminants lorsqu'il s'agit de caractériser les messages appartenant aux quatre catégories de premier niveau, puis aux treize catégories de sous-niveaux.

5. Résultats

Cette section présente les résultats des analyses statistiques du courriel des cadres. Tout d'abord, les résultats de la catégorisation automatique sont exposés. Ensuite, les indices non lexicaux qui s'avèrent les plus discriminants sont présentés.

5.1. Catégorisation automatique

La catégorisation automatique en fonction des indices non lexicaux et lexicaux fut effectuée à l'aide de l'algorithme KNN, où K=2. Pour la catégorisation lexicale, il s'est avéré que les performances étaient généralement plus élevées lorsque K=2 à partir de N=2000 (i.e. lorsque 2000 mots sont employés pour décrire les documents); à N=2000, les performances atteignent un plateau pour n'augmenter que très minimalement par rapport à l'augmentation rapide du nombre de mots ("features"). Afin de permettre une comparaison des performances de la catégorisation lexicale par rapport à la catégorisation non lexicale, nous avons retenu les résultats où K=2 et N=2000, pour quatre et treize catégories. Tab. 1 résume les résultats obtenus par l'opération de catégorisation, selon les quatre catégories principales de la typologie, puis les treize sous-catégories.

Catégorie	Mesure	**Catégorisation non lexicale** Score (K=2)	**Catégorisation lexicale** Score (K=2)
4 catégories	Rappel	56.01	94.32
	Précision	40.91	94.50
	Exactitude	59.71	94.54
13 catégories	Rappel	35.98	91.09
	Précision	59.63	84.18
	Exactitude	51.82	88.70

Tableau 1 : Résultat de la catégorisation automatique non lexicale et lexicale

En consultant les résultats figurant dans ce tableau, on constate que le processus de catégorisation automatique en fonction d'indices non lexicaux propres au courriel donne des résultats beaucoup moins performants. Sur 1700 messages classés dans le corpus (3 messages ont été écartés par le système), les performances moyennes du système sont les suivantes : pour quatre catégories, le taux de rappel moyen est de 56.01%, le taux de précision moyen est de 40.91% et le taux d'exactitude (nombre de documents correctement catégorisés / nombre total de documents dans

le corpus) est de 59.71%. Pour les treize catégories, le taux de rappel moyen est de 35.98%, le taux de précision moyen est de 59.63% et le taux d'exactitude est de 51.82%. Cependant, on constate que le processus de catégorisation automatique en fonction d'indices lexicaux propres au courriel donne des résultats très satisfaisants. Sur les 1703 messages du corpus, les performances moyennes du système sont les suivantes : pour quatre catégories, le taux de rappel moyen est de 94.32%, le taux de précision moyen est de 94.5% et le taux d'exactitude est de 94.54%. Pour les treize catégories, le taux de rappel moyen est de 91.09%, le taux de précision moyen est de 84.18% et le taux d'exactitude est de 88.70%.

5.2. Sélection des indices non lexicaux les plus discriminants

La figure 2 révèle la pertinence des caractéristiques liées à l'émetteur et au destinataire des messages lorsqu'il s'agit de catégoriser le courriel en fonction des quatre catégories de premier niveau. Spécifions que, contrairement au seuil de signification traditionnellement admis en statistiques (soit 0.05), la valeur du *p* correspond ici à la probabilité que le degré d'association soit dû au hasard. Ainsi, on peut conclure, si p=0, que le hasard n'est pas la cause d'explication du lien entre les variables.

N°	Trait non lexical	Statistiques	Histogramme	p
1	De=Groupe social	0.493090		0.000000
2	À=Liste d'envoi	0.319338		0.000000
3	Nbre Cc (4 valeurs)	0.268207		0.000000
4	Caractères gras	0.263694		0.000000
5	Bilingue	0.240622		0.000000
6	De=Liste d'envoi	0.202207		0.000000
7	Image	0.147928		0.000000
8	Majuscule	0.132210		0.000000
9	Fil Discussion	0.117114		0.000000
10	Couleur	0.115434		0.000000
11	Signature prénom	0.106738		0.000000
12	Objet=RE	0.105822		0.000000
13	Cc=Liste d'envoi	0.105345		0.000000
14	Importance haute	0.104924		0.000000
15	Objet=TR	0.104212		0.000001
16	Signature absence	0.102243		0.000001
17	Signature officielle	0.101462		0.000001
18	Nbre À (3 valeurs)	0.092501		0.000003
19	Objet=FW	0.079915		0.000295
20	De=À	0.062954		0.008523
21	PJ directe	0.060340		0.018843
22	PJ indirecte	0.054274		0.033729

Figure 2 : Pertinence des indices non lexicaux pour quatre catégories

L'appartenance de l'émetteur au groupe social du destinataire constitue ici l'indice qui arrive en tête (De=Groupe social). En deuxième et troisième position, c'est le destinataire qui est en cause, avec des indices portant sur la présence d'une liste d'envoi dans le champ principal À (À=Liste d'envoi) et le nombre de destinataires en copie conforme (Nbre CC). En sixième position, le fait que l'émetteur est une liste d'envoi est également représenté (De=Liste d'envoi). Fait intéressant, il semble également que certains indices liés à l'aspect visuel des messages s'avèrent pertinents lorsqu'il s'agit de caractériser les messages suivant quatre catégories. On retrouve en quatrième position l'indice lié à l'usage du caractère gras, en septième position l'indice lié à l'usage d'images et en huitième position le recours aux majuscules. L'indicateur

relatif à l'usage alterné du français et de l'anglais (Bilingue) est aussi un trait pertinent, puisqu'il arrive en cinquième position.

Fig. 3, qui permet de comparer quels indices non lexicaux sont davantage pertinents pour caractériser les messages appartenant aux treize catégories de sous-niveaux, présente des résultats similaires à ceux obtenus pour quatre catégories. Notamment, on retrouve en deuxième position l'indice lié à l'appartenance de l'émetteur au groupe social du destinataire (De=Groupe social), ce qui confirme sa pertinence au cours du processus de catégorisation. Le nombre de destinataires en copie conforme (Nbre CC), le bilinguisme des messages (Bilingue), le recours aux listes d'envoi (À=Liste d'envoi, De=Liste d'envoi) ainsi que l'utilisation des caractères gras sont également des indices importants.

N°	Trait non lexical	Statistiques	Histogramme	p
1	De=À	0.446520		0.000000
2	De=Groupe social	0.372082		0.000000
3	Nbre Cc (4 valeurs)	0.306914		0.000000
4	Bilingue	0.299111		0.000000
5	À=Liste d'envoi	0.275591		0.000000
6	Objet=FW	0.251415		0.000000
7	Caractères gras	0.249228		0.000000
8	De=Liste d'envoi	0.249075		0.000000
9	Fil Discussion	0.216368		0.000000
10	Objet=RE	0.201733		0.000000
11	Image	0.190165		0.000000
12	Nbre À (3 valeurs)	0.179601		0.000000
13	Signature absence	0.173091		0.000000
14	Couleur	0.169493		0.000000
15	Signature officielle	0.139327		0.000000
16	Objet=TR	0.131083		0.000000
17	PJ indirecte	0.124613		0.000000
18	Majuscule	0.116428		0.000000
19	Signature prénom	0.109529		0.000000
20	PJ directe	0.106701		0.000000
21	Importance haute	0.097343		0.000000
22	Cc=Liste d'envoi	0.081559		0.000096

Figure 3 : Pertinence des indices non lexicaux pour treize catégories

6. Discussion et conclusion

Dans notre étude, nous avons obtenu des résultats performants avec l'algorithme des KNN (où K=2 et N=2000) pour catégoriser automatiquement les courriels en fonction de leur lexique. Pour la catégorisation lexicale à quatre catégories, le taux de rappel moyen fut de 94.32%, le taux de précision moyen fut de 94.50% et le taux d'exactitude fut de 94.54%. Pour treize catégories, le taux de rappel moyen fut de 91.09%, le taux de précision moyen fut de 84.18% et le taux d'exactitude fut de 88.70%. Dans le cadre de méthodes supervisées, Carvalho et Cohen (2005) ont obtenu des résultats allant jusqu'à 82% pour huit catégories selon la mesure F, avec un algorithme ('Dependency Network') tenant compte du lexique des messages. Goldstein et Sabin (2006) ont obtenu jusqu'à 63% d'exactitude pour cinq catégories avec un algorithme ('Random Forest') tenant compte du lexique et de la longueur des messages. Nous avons démontré que les indices non lexicaux s'avèrent moins performants que les indices lexicaux comme traits discriminants pour catégoriser automatiquement de courriels en fonction de catégories pragmatiques. Même si les résultats obtenus à l'aide des indices non lexicaux

sont moins élevés, ils demeurent malgré tout acceptables compte tenu de la complexité de la tâche. Par ailleurs, au-delà des performances, ces expérimentations ont permis d'identifier quels indices non lexicaux sont les plus discriminants pour décrire les documents que l'on souhaite catégoriser automatiquement.

Cette recherche valide la pertinence d'avoir recours à une approche inspirée de la théorie des actes de langage pour l'étude du courriel, puisque ce sont les actions véhiculées dans les messages qui intéressent davantage les cadres lors du tri des messages. Contrairement à la plupart des études qui s'inspirent de la typologie de Searle (1972) pour catégoriser le courriel, nous proposons une approche basée sur la présence de déclencheurs cognitifs (la priorité et l'engagement perçus par le destinataire) à l'origine des pratiques de tri. Ici, c'est plutôt la dimension perlocutoire de l'acte de langage (effet sur le destinataire) que sa dimension illocutoire (intention de l'émetteur) qui prévaut. Selon nos résultats, cette approche concorde davantage avec le processus d'interprétation des messages en contexte de travail réel.

Références

Alberts I. (2009). Exploitation des genres de textes pour assister les pratiques textuelles dans les environnements numériques de travail : le cas du courriel chez des cadres et des secrétaires dans une municipalité et une administration fédérale canadiennes. Thèse de doctorat. Université de Montréal.

Alberts I. and Bertrand-Gastaldy S. (2008). A pragmatic perspective of e-mail management practices in two Canadian public administrations. In *Culture and identity in knowledge organization. Proceedings of the tenth international ISKO conference*, pp. 347-353.

Alberts I. and Forest D. (en préparation). Catégorisation automatique du courriel en fonction de catégories pragmatiques fondées sur l'utilisation de traits discriminants lexicaux.

Austin J.L. (1970). *Quand dire, c'est faire*. Paris : Éditions du Seuil.

Carvalho V. and Cohen W. (2005). On the collective classification of email "speech acts". In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 345-352.

Carvalho V. and Cohen W. (2006). Improving email speech acts: analysis via N-gram selection. In *Proceedings of the analyzing conversations in text and speech (ACTS) workshop*, pp. 35-41.

Cervoni J. (1987). *L'énonciation*. Paris : PUF.

Cohen W., Carvalho V.R and Mitchell T.M. (2004). Learning to classify email into "speech acts". In *International conference on empirical methods in natural language processing*, pp. 309-316.

Forest D. (2006). Application de techniques de forage de textes de nature prédictive et exploratoire à des fins de gestion et d'analyse thématique de documents textuels non structurés. Thèse de doctorat. Université du Québec à Montréal.

Goldstein J. and Sabin R. (2006). Using speech acts to categorize email and identify e-mail genres. In *Proceedings of the 39th annual HICSS*.

Scerri S., Davis. B. and Handschuh S. (2007). Improving e-mail conversation efficiency through semantically enhanced e-mail. In *18th international conference on database and expert systems applications*, pp. 490-494.

Searle J.R. (1972). *Les actes de langage ; essai de philosophie du langage*. Paris : Hermann.

Winograd T. (1988). A language-action perspective on the design of cooperative work. *Human computer interaction*, vol 3 : 30.