Le traitement automatisé des courriels pour les services aux investisseurs: une approche par la question-réponse

Luc Bélanger
Université de Montréal
luc.belanger@acm.org

15 janvier 2003

Table des matières

Chapitre 1

Introduction

1.1 Description de la problématique de la réponse au courriel

La problématique que je compte étudier est la réponse automatisée aux courriels. La mise au point de solutions pratiques pour traiter ce problème est motivée par l'augmentation du nombre de communications effectuées par l'entremise du courriel. Le courriel est un médium de communication qui combine les avantages du courrier traditionnel à ceux du téléphone. La livraison du contenu est presque instantanée, comme avec le téléphone, et le traitement peut se faire en différé, comme avec le courrier traditionnel. Cette méthode de communication possède par contre certains inconvénients. L'utilisateur ne peut pas savoir si son courriel a été reçu par le destinataire et il s'attend à ce que le destinataire lui réponde à l'intérieur d'un délai de 24 heures. Les communications par courriel sont devenues très importantes depuis que le fait d'être connecté à l'Internet est devenu presque aussi courant que d'avoir le service téléphonique.

Dans les entreprises, le courriel est devenu un outil indispensable à leur succès. C'est un moyen simple et peu coûteux de rester en contact avec les clients. L'entreprise, en tant qu'émettrice du message se sert principalement du courriel pour faire de la publicité et annoncer ses résultats financiers. L'entreprise utilise ainsi le courriel de façon impersonnelle, sans communiquer avec une personne en particulier. Cette pratique n'est pas problématique pour l'entreprise car les coûts rattachés à cette pratique sont minimes et les économies réalisées sont substantielles par rapport à

l'utilisation du courrier traditionnel, de la télécopie ou du téléphone (télémarketing). L'inconvénient de cette utilisation du courriel incombe au client qui se retrouve parfois avec un très grand nombre de courriel qu'il ne désire pas.

Les problèmes de l'utilisation du courriel par une entreprise surviennent lorsque celle-ci devient réceptrice du message. Les messages reçus proviennent des utilisateurs et ils ont généralement un but interrogatif. Les clients qui envoient des messages veulent avoir rapidement des réponses à leurs questions. Puisque chaque service à l'intérieur des entreprises possède une ou des adresses de courriel, les gens communiquent directement avec eux sans rencontrer les inconvénients des médiums traditionnels. Cette pratique entraîne une augmentation des communications électroniques. Le courrier traditionnel a l'inconvénient d'avoir un temps de traitement trop long pour réaliser une communication efficace entre la compagnie et ses usagers. Le temps nécessaire pour écrire la lettre et effectuer les manipulations pour l'envoi postal ainsi que les contraintes sur l'horaire de levée du courrier rendent cette méthode de communication désuète pour les communications qui peuvent se faire par courriel. Un avantage considérable en faveur de l'utilisation du courriel est la possibilité d'insérer des documents dans le message sans en augmenter le coût contrairement à la messagerie traditionnelle. Le téléphone est le médium de communication le plus utilisé dans les entreprises pour le service à la clientèle et la façon la plus rapide pour les utilisateurs d'entrer en contact avec une compagnie. Il est par contre très coûteux pour une entreprise d'offrir un service par téléphone qui rencontre les exigences des clients. Pour satisfaire la clientèle, une compagnie devrait offrir un service où le client n'attend pas pour parler à un préposé et où ce même préposé peut répondre à toutes les questions du client. Un tel service n'est pas réalisable en pratique car il demanderait des ressources considérables. Les approches qui ont été proposées et qui sont présentement utilisées offrent toutes leurs lots d'inconvénients. La solution la plus simple est la mise en attente, elle permet à l'utilisateur de parler avec un préposé, mais au prix d'une attente parfois interminable. Un autre type de système est le menu téléphonique, celui-ci possède l'avantage de diminuer l'attente du client et donne généralement des résultats satisfaisants. L'inconvénient d'un système de menu est que l'utilisateur peut facilement s'y perdre et il est possible qu'il doive attendre pour parler à un préposé si sa question ne fait pas partie des choix qui lui sont offerts. Le dernier type de système de communication par téléphone est celui où l'utilisateur enregistre un message sur une boîte vocale pour qu'un préposé le rappelle. Cette façon de faire ne fait pas attendre l'utilisateur, mais lorsque le préposé retourne l'appel, l'utilisateur devra être disponible pour la conversation. L'utilisation du

courriel n'implique aucun de ces inconvénients, l'utilisateur envoie son message et un préposé se charge ensuite de lui répondre. L'utilisateur peut consulter la réponse lorsqu'il est disponible.

Les courriels reçus par une entreprise nécessitent habituellement une réponse. Le traitement de ces courriels prend beaucoup de temps de la part des préposés, ceux-ci doivent prendre le temps de lire et comprendre le courriel pour ensuite répondre à l'utilisateur. Ce traitement est réalisé en supposant que le préposé connaisse la réponse ou qu'il est assez compétent pour en faire la recherche. Par contre lorsque le préposé doit chercher la réponse, les coûts pour l'entreprise augmentent car le temps de travail nécessaire pour répondre au courriel augmente aussi. La réponse automatisée aux courriels est un moyen pour réduire le temps de traitement des courriels et un outil qui sera utilisé pour mettre en oeuvre les politiques de gestion des communications des entreprises avec les consommateurs[1]. Ces politiques ont pour but d'améliorer la satisfaction des consommateurs, cultiver leur loyauté envers l'entreprise, réduire les coûts et personnaliser le service. L'utilisation d'un système automatisé de réponse permet l'uniformisation du discours de l'entreprise, un service de qualité égale pour tous les départements et la possibilité de contrôler l'information confidentielle.

Les systèmes de réponses automatiques aux courriels présentement offerts sur le marché sont coûteux et ne répondent pas aux attentes des compagnies. Les systèmes de traitement automatique du courriel les plus simples sont de type *Majordomo* et leur mécanisme est très primaire. Le système identifie des mots-clés dans le sujet ou dans le corps du message et exécute l'action déterminée par la commande. Ce type de traitement est utilisé pour réaliser la gestion des listes d'envoi. Dans le courriel, tout ce qui n'est pas un mot-clé est ignoré par un tel système. Un autre type de fonctionnement est l'utilisation de plusieurs adresses de courriel, chacune retournant un message différent au destinataire, peu importe le contenu des messages. Les systèmes plus évolués utilisent des techniques plus sophistiquées, coûtent assez cher et semblent être d'une utilité plutôt limitée, étant donné qu'ils sont généralement développés spécifiquement pour un client. Les techniques utilisées dans ces systèmes sont principalement : des modules de classifications, des patrons d'expressions régulières, des extracteurs d'information et des patrons à trous pour la génération des réponses. Présentement, la majorité des systèmes de traitement n'utilisent pas toutes les possibilités de traitement du courriel. Ils ne servent qu'à faire l'aiguillage des messages vers les intervenants les plus aptes à répondre.

[1] « customer relation management » (CRM)

L'étude des systèmes automatisés de réponse au courriel n'est pas qu'un simple exercice théorique. Cette problématique est présente dans toutes les entreprises qui veulent améliorer leurs relations avec leurs clients à une époque où l'utilisation du courriel est omniprésente. Le projet de recherche que je présente dans ce document est réalisé à l'intérieur d'un projet développé dans le cadre des Laboratoires Universitaires Bell (LUB). La problématique des LUB est la gestion des courriers électroniques, dans le contexte des relations avec les investisseurs de la compagnie mère Bell Canada Entreprise (BCE). Le service de relation avec les investisseurs de BCE reçoit un nombre important de courriel d'investisseurs ou d'investisseurs potentiels qui veulent obtenir de l'information sur le cours de leurs titres, les assemblées d'actionnaires, les possibilités d'investissements dans la compagnie et les résultats (financier, commercial) de la compagnie. L'utilisation d'un système de réponse automatique aux courriels efficace, appliqué au service de relation avec les investisseurs permettra d'améliorer la qualité de ce service. Les employés de ce service pourront alors prendre le temps nécessaire pour donner un meilleur service aux courriels nécessitant un traitement plus spécialisé.

La réalisation de ce projet de recherche a été attribuée à une équipe du laboratoire RALI dont je fais partie. Une analyse préliminaire du projet a d'abord été réalisée par Leila Kosseim [6], cette analyse avait pour but de faire l'étude des solutions disponibles. Je ferai un retour sur cette étude dans le prochain chapitre. Puisque ce projet est d'une envergure considérable, il a été divisé en modules et chacun de ces modules a été assigné à un étudiant. Stephen Beauregard [1] a été le premier étudiant à travailler sur le projet de recherche, il a étudié la problématique de la génération de réponses. Julien Dubois [3] a travaillé sur la problématique de la classification des courriels. Dans son travail, Julien Dubois a étudié l'apport des méthodes de classification en usage présentement et des méthodes de prétraitement à la problématique de la réponse automatisée aux courriels. La classification des courriels nous aide à déterminer les caractéristiques prédominantes de chaque message pour pouvoir effectuer le traitement de ceux-ci par le module le plus approprié.

Dans le cadre de ce projet, je m'intéresse à l'utilisation des systèmes de question-réponse pour le traitement des courriels qui ont un caractère interrogatif. L'approche par question-réponse est très bien adaptée au contexte du service à la clientèle, les clients ont toujours des demandes et des questions. Les travaux reliés à cette approche ont déjà été entrepris par Luc Plamondon [11], ce qui a permis de créer le système de question-réponse QUANTUM. Le système QUANTUM a été

réalisé comme un système de question-réponse pur. Il ne possède aucune caractéristique adaptée au traitement du courriel, particulièrement pour la prise en compte du contexte du courriel et l'extraction de la question. De plus, QUANTUM est construit pour répondre à des questions générales, il n'est pas spécialisé pour le domaine des relations aux investisseurs que j'ai décidé de traiter dans ce projet. Il me sera tout de même très utile de m'inspirer des différentes facettes de ce système pour traiter les courriels, comme nous le verrons dans les prochains chapitres.

1.2 L'approche par Question-Réponse

L'utilisation de l'approche par question-réponse est motivée par la nature interrogative des courriels reçus par les personnes assurant le service auprès des clients. Le procédé qui consiste à répondre à un courriel peut être assimilé au processus cognitif de répondre à une question, à la différence que dans les courriels, la question est interprétée dans un contexte spécifique, défini par le courriel. La procédure de recherche de l'information pour formuler la réponse s'apparente en tout point à ce qui est réalisé dans les systèmes de question-réponse. Même lorsque le courriel n'est pas de nature interrogative, le préposé devra généralement répondre à la personne. Dans les cas où le courriel est un message de félicitation, de recommandation ou encore pire lorsque c'est une plainte, l'expéditeur du courriel s'attend tout de même à recevoir un accusé de réception, sous peine de se sentir ignoré.

Les travaux réalisés présentement dans le domaine de la question-réponse ne correspondent pas exactement à la problématique de la réponse automatique aux courriels. Par contre ces deux domaines possèdent assez de points en commun pour pouvoir espérer exploiter les réalisations effectuées dans le domaine de la question-réponse à la gestion des courriels. La problématique étudiée à l'heure actuelle dans le domaine de la question-réponse est bien encadrée et elle s'effectue en continuité avec les travaux réalisés dans le domaine de la recherche d'information. Les travaux en question-réponse sont présentement réalisés sous certaines conditions dans un « environnement contrôlé ». La question qu'un système de question-réponse a à traiter est toujours énoncée explicitement et dans un seule phrase.

Le système de question-réponse QUANTUM est issu de cette branche de la recherche en question-

réponse. Il reçoit une question en entrée et tente de trouver la réponse à celle-ci dans un grand corpus de données (environ 3 Go.) Le système a été conçu pour participer à l'évaluation annuelle « Text REtrieval Conference » (TREC), organisé par le « National Institute of Standards and Technology » américain (NIST). Le type de question auquel le système est censé répondre est : *What is caffeine ?*, *In Poland, where do most people live ?*, *What is the conversion rate between dollars and pounds ?*. La réponse retournée est seulement un mot ou une expression, aucune génération de texte n'est effectuée.

La tâche la plus complexe pour ce type de système est l'identification de la réponse [8], la question n'a pas besoin d'être analysée aussi profondément qu'un courriel doit l'être. Dans la réponse automatique aux courriels, la tâche d'extraire la question, s'il y a lieu, comporte de grands défis. Dans le chapitre 3, je fais une description des particularités des courriels et des caractéristiques du corpus de courriels utilisés jusqu'à présent pour ce projet de recherche. L'analyse des données met en évidence les problèmes que je vais rencontrer au cours de la réalisation de ce projet. Les deux problèmes majeurs à la mise en oeuvre d'un système automatisé de réponse aux courriels par une approche question-réponse sont l'analyse du courriel et la recherche de la réponse à partir des données disponibles. La recherche de la réponse est une tâche tout aussi difficile que l'analyse de la question. Cette tâche se complique davantage avec un système de type de QUANTUM car la question de la granularité de la réponse doit être considérée, ce qui n'est pas le cas. On peut répondre à une question de plusieurs manières, selon l'utilisateur qui soumet la question. Dans le cas d'un client qui veut savoir comment investir dans la compagnie, la réponse peut être un message référant le client à un courtier, mais si ce client est lui-même courtier cette réponse ne satisfera pas son besoin d'avoir une information plus précise.

Dans le chapitre 2, je fais un survol des systèmes de réponse automatisés aux courriels et je décris la structure générale des systèmes de question-réponse tel qu'ils sont étudiés présentement. Au chapitre 3, je fais une description des données disponibles pour la réalisation du projet. Je présente ensuite, au chapitre 4, mon plan de recherche pour le développement de solutions à la problématique de réponse aux courriels par l'approche question-réponse.

Chapitre 2

Revue de littérature

Ce chapitre est un résumé des différentes facettes des problèmes qui ont été analysés dans la littérature et dans la réalisation des systèmes de réponse automatisés aux courriels. La première section est un tour d'horizon des différents types de systèmes de réponse automatiques disponibles sur le marché. La deuxième section est une description des problématiques à solutionner pour la réalisation du système de question-réponse. Dans cette seconde section j'explique les avenues qui ont été étudiées pour trouver des solutions aux problèmes de la question-réponse.

2.1 Systèmes de réponse automatique

L'utilisation du courriel pour le service à la clientèle est un moyen efficace pour réduire les coûts associés à ce service comparativement au téléphone. Le traitement des requêtes n'est pas obligatoirement linéaire, l'information retournée en réponse à une requête peut être conservée pour y référer dans le futur, le client peut prendre le temps d'exprimer clairement ses idées, le préposé peut mieux organiser son temps en connaissant le nombre de requêtes auxquels il a à répondre ; préposé et client peuvent donc traiter l'information à leur rythme. Tous ces avantages font en sorte que le nombre de communications par courriel augmente rapidement. Ainsi, il est primordial d'avoir des solutions pour diminuer les coûts associés à cette augmentation des communications.

Les systèmes de gestion automatisée des courriels peuvent être catégorisés selon les opérations

qu'ils réalisent. La première catégorie de systèmes sont les systèmes d'auto-réponse qui ne sont que des robots programmés pour exécuter une action selon les mots-clés présents dans le message. Dans une catégorie différente, il y a les systèmes de gestion des courriels qui sont conçus pour faciliter la tâche de ceux qui ont à répondre aux courriels. Les systèmes auxquels je vais m'intéresser sont ceux capables d'analyser les courriels et de formuler une réponse. Ces trois types de systèmes sont décrits dans les sections qui suivent.

2.1.1 Systèmes d'auto-réponse

Les systèmes d'auto-réponse sont les premiers à avoir vu le jour, sous la forme de gestionnaire de listes de discussions. Ces programmes sont les plus simples, ils exécutent une action selon la présence de mots-clés dans le sujet ou le corps du message. Les logiciels *Majordomo* et *Mailman* sont deux systèmes bien connus de gestion de listes de discussions fonctionnant sur le principe d'activation par mots-clés. Ce type de système peut gérer les abonnements, le transfert des messages aux bonnes listes de discussions ainsi que les requêtes pour avoir accès aux anciens messages et traiter certaines tâches administratives. Les logiciels de type *Procmail* peuvent aussi être considérés comme faisant partie de cette catégorie, en autant qu'on se limite à utiliser les fonctionnalités fournies par le logiciel. Il existe aussi une autre catégorie de systèmes qui sont encore plus primitifs d'un point de vue technique, fondés sur ce que je pourrais appeler « l'ingénierie sociale », et dont une variante est utilisée par BCE. Dans ces systèmes seulement le champs *destinaire* du message est traité pour déterminer le traitement que le courriel recevra. Chez BCE, chaque département possède une adresse électronique et il incombe donc à l'utilisateur la tâche d'envoyer son courriel au bon département pour espérer avoir une réponse satisfaisante. Ainsi, un utilisateur voulant contacter le service des relations avec les investisseurs enverra un courriel à l'adresse `relations.investisseurs@bce.ca` s'il rédige son courriel en français ; s'il le rédige en anglais il devra l'envoyer à l'adresse traitant les courriels en anglais, `investors.relations@bce.ca`. Il en est de même pour la plupart des départements. Dans le cas de BCE, les courriels sont ensuite traités manuellement, dans le cadre d'un système automatisé, une réponse pré-établie est envoyée, sans avoir effectué de traitement sur le courriel envoyé. La caractéristique commune à ces systèmes est l'absence de traitement sur le contenu du courriel, toute l'information qui n'est pas dans la liste des mots-clés est complètement ignorée.

2.1.2 Systèmes de gestion de courriels

Les systèmes de gestion de courriels permettent d'augmenter la productivité des préposés affectés au traitement du courriel en leur offrant un meilleur environnement pour répondre aux courriels. Ces systèmes offrent plusieurs fonctionnalités, les plus importantes sont : la catégorisation des messages, l'envoi d'accusés de réception, le routage des messages, les suggestions de réponses, l'intégration du système dans l'environnement de travail du préposé, l'archivage des courriels, la production de rapports statistiques et historiques.

Parmi les systèmes de ce type, certains sont capables de générer des réponses. La première façon de faire est de suggérer une réponse à partir d'un patron pré-établi manuellement. Selon cette méthode, le préposé reçoit un message, possiblement classifié automatiquement, qu'il lit et dans lequel il annote l'information pertinente. Par la suite, le préposé peut alors sélectionner un patron approprié dans lequel l'information annotée sera insérée. Ce traitement peut se faire à deux niveaux, le premier pour générer un accusé de réception en attendant la réponse au courriel par un préposé et le deuxième pour générer la réponse complète d'un seul trait. L'inconvénient d'un tel mode de génération de réponse est que l'essentiel du travail est fait manuellement par un préposé. Malgré tout, une automatisation minimale est possible pour extraire le nom de l'expéditeur et les autres informations qui sont bien balisées.

Une autre façon de générer des réponses aux requêtes des clients est d'utiliser une méthode par formulaire. Cette approche est utilisée sur le site web de BCE (Figure 2.1) pour les demandes en ligne de documents tels que le rapport annuel, les rapports trimestriels, les prospectus, etc. L'entrée des données est balisée par des boîtes de dialogues que l'utilisateur est contraint de remplir de la façon dictée par le récepteur du message. L'avantage de ce modèle est que l'information nécessaire à la génération de la réponse est indiquée clairement par l'annotation automatique de l'information reçu dans les champs du formulaires. En ce qui concerne BCE, le traitement ne semble pas être automatisé comme je le décris, pour la seule raison qu'aucune réponse électronique ne sera envoyé car l'adresse de courriel ne figure pas dans les champs d'information à fournir.

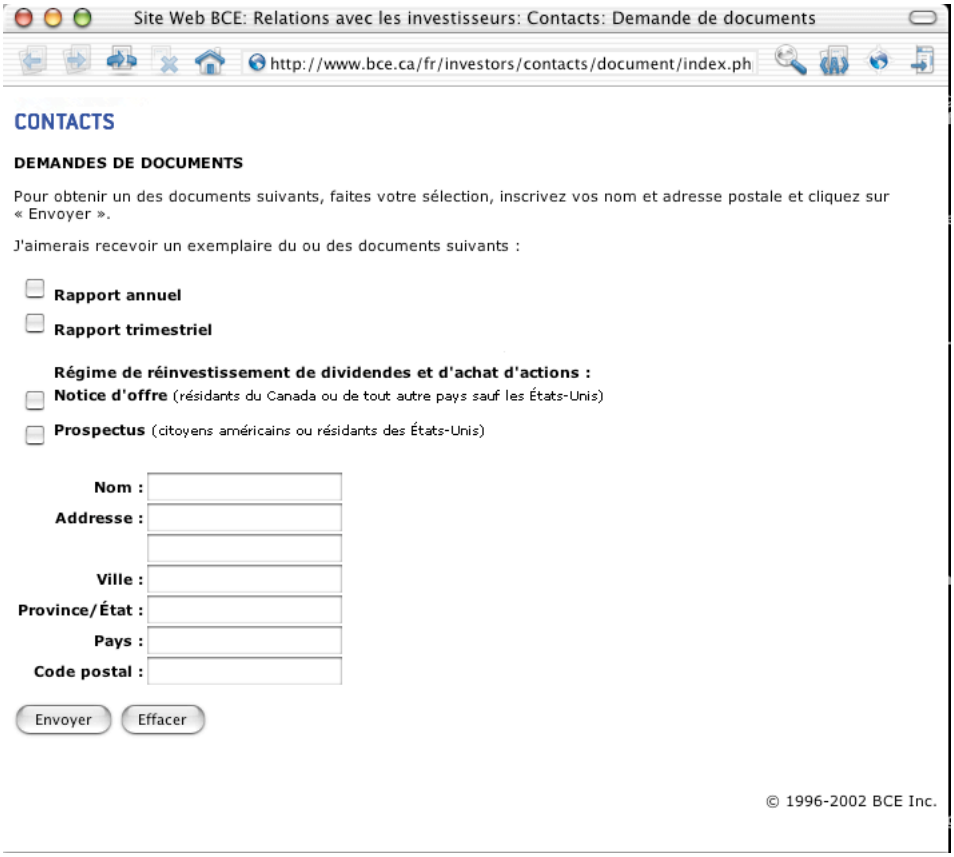

FIG. 2.1 – Formulaire DEMANDES DE DOCUMENTS de BCE

2.1.3 Réponse aux courriels non-balisés

Les systèmes de réponse aux courriels dont le contenu n'est pas balisé sont ceux qui essaient de répondre en se basant sur le contenu du message. Ce type de système est celui qui coûte le plus cher, il est généralement développé sur mesure pour l'entreprise qui en fait l'acquisition en fonction de ses besoins. La présente section s'inspire des travaux réalisés par Leila Kosseim [6], dont je me charge de faire un résumé. Je présente l'architecture des systèmes sans en cibler un précisément, mais en tentant de décortiquer leur fonctionnement général ainsi que les principales technologies utilisées.

La première étape est une analyse préliminaire du courriel par extraction d'information ou classification. Pour effectuer l'extraction d'information, les différentes techniques utilisées sont : des dictionnaires du champ d'application, des patrons de textes reliés aux domaine d'application

et des expressions régulières. Lorsqu'il y a classification, les principaux schémas utilisés sont la classification bayesienne, la classification par patrons (mots-clés, expression régulière, expressions à trous) et le raisonnement à base de cas.

Entre cette première étape et la génération de la réponse, les systèmes réalisent des traitements sur les courriels. Par contre, dû au fait que ces systèmes sont commerciaux, les avantages technologiques de chacun ne sont pas divulgués, ce qui m'empêche donc d'obtenir tous les détails de fonctionnement des systèmes. Après ces traitements spécifiques, l'étape de génération de la réponse est possiblement la plus importante. Cette étape du traitement a été étudiée par Stephen Beauregard dans le cadre de son mémoire de maîtrise [1]. Les méthodes de génération de réponses sont assez variées. La première est l'utilisation de documents sélectionnés à partir d'une banque de documents fixes, selon l'analyse du courriel, et envoyés intégralement comme réponse au courriel. Une variante de cette méthode est d'effectuer une personnalisation très simple des documents, où certaines phrases de politesse et règles d'usage sont instanciées selon l'information fournie par l'analyse du courriel. Les systèmes qui semblent être les plus évolués pour rédiger la réponse sont ceux qui utilisent des patrons à trous où l'information nécessaire pour remplir les trous demande une analyse beaucoup plus profonde des courriels.

Dans la problématique de la réponse automatisée au courriel dans le cadre du projet des LUB, la génération de la réponse est un problème qui sera étudié lorsque les modules d'analyses de courriels et d'extraction d'informations seront à un stade plus avancé. La question de la génération ne doit tout de même pas être ignorée car la qualité de l'information extraite et analysée des courriels dicte en quelque sorte la façon dont la réponse sera générée. Cette attention particulière à la qualité de l'information se manifeste dans les systèmes de question-réponse où la granularité de la réponse doit être bien déterminée pour satisfaire l'expéditeur du courriel.

2.2 Problématique de la Question-Réponse

Les systèmes de question-réponse ont pour tâche d'analyser une question formulée en langue naturelle et de lui trouver une ou plusieurs réponses à partir d'une banque de données, d'un texte ou d'une base documentaire. Le besoin pour des systèmes capables de répondre précisément et suc-

cinctement à des questions provient de l'inaptitude des engins de recherche à donner des réponses aux requêtes des utilisateurs ; les listes de documents retournées par ceux-ci sont rarement satisfaisantes. La problématique de la question-réponse est une variante simplifiée du « test de Turing » qui définit des critères pour déterminer l'intelligence d'une machine, à la différence que les systèmes de question-réponse ne font pas la conversation et que leur capacité de faire des raisonnements est très limitée. De plus, lors d'une conversation, une machine doit être capable d'interpréter l'information reçue dans un contexte, ce qui n'est pas le cas dans la problématique de la question-réponse. Pour réaliser le traitement automatique des courriels, le contexte doit être considéré afin que le système s'acquitte de sa tâche convenablement. Il y a plusieurs façons de situer la recherche effectuée dans le domaine de la question-réponse, la première étant l'approche par la compréhension d'histoire [9]. Celle-ci pourrait être assimilée à la génération de résumé car c'est dans ce domaine que se fait présentement les travaux les plus importants par rapport à la compréhension d'histoire. La deuxième approche pour aborder la question-réponse est l'extraction d'information à partir de grand corpus, cette piste de recherche peut être considérée comme la continuation des travaux réalisés dans le domaine de la recherche d'information.

Depuis la conférence TREC-8 (Text REtrieval Conference) [13] en 1999, le NIST organise une session dont le but est d'évaluer la performance des systèmes de question-réponse dans un environnement contrôlé, afin de comparer efficacement les systèmes entre eux. Cette tâche d'évaluation est réalisée à partir d'un ensemble de questions factuelles où les réponses sont extraites d'un corpus de nouvelles contenant plusieurs milliers d'articles. La nécessité d'étudier les systèmes de question-réponse provient du fait que les performances des systèmes de recherche d'information n'évoluent plus aussi rapidement qu'auparavant. Le but de ces recherches est de développer de nouvelles idées, de nouveaux modèles et de nouvelles méthodes pour faciliter la tâche des employés du monde de l'« intelligence », les analystes [2] oeuvrant pour les bureaux d'enquête, d'espionnage et de sécurité intérieure.

Les types de questions auxquelles les systèmes de QR peuvent être mis à contribution lors de la conférence TREC sont les questions factuelles et les questions de synthèse. Les questions d'opinions où le point de vue du narrateur est important sont aussi considérées[1]. Par contre je n'ai pas étudié cette problématique car elle ne se présente pas dans les données dont je dispose. Les questions

[1]Multiple Perspectives Workshop, Janyce Wiebe. NRRC été 2002

factuelles sont similaires aux questions que l'on retrouve dans les jeux questionnaires télévisés où la réponse est un mot ou une expression. Par exemple, à la question *What is the democratic party symbol ?*, la réponse est *éléphant*. Depuis la première conférence (TREC-8), la difficulté de la tâche a toujours augmenté même s'il est difficile d'établir une mesure de la complexité d'une question. Les questions font maintenant référence à des concepts dont les relations sont plus complexes, *What is the appropriate gift for a 10th anniversary ?*, et les réponses plus complexes *What two European countries are connected by the St. Gotthard Tunnel under the Alps ?*

Le laboratoire RALI participe à cette compétition depuis l'année 2000, soit lors de la tenue de TREC-9. Dans le cadre de cette compétition, le système de QR présenté fût XR^3 développé par Michael Laszlo et Leila Kosseim [7]. Le système XR^3 utilise une approche simple qui exécute une analyse de surface de la question et des fenêtres de résultats. L'analyse est réalisée à partir d'expressions régulières encodées manuellement. Ce système a été la source d'inspiration menant à la création, par Luc Plamondon, de QUANTUM [11]. Je vais, dans cette section, situer les différentes composantes de QUANTUM par rapport à un cadre générique de développement des systèmes de QR.

Les systèmes de QR présentement développés, principalement dans le cadre de la conférence TREC, sont presque tous construits autour d'une même architecture héritée des systèmes de recherche d'information. Je vais donc présenter le cadre global de cette architecture et ensuite les approches utilisées pour chacune des composantes. Les interactions entre les composantes d'un système de QR général sont illustrées dans la figure 2.2. L'architecture présentée correspond à l'utilisation de la question-réponse dans un cadre général qui ne reflète pas exactement celle des systèmes participant à la conférence TREC, mais qui peut être utilisée pour des problèmes comme le traitement automatisé du courriel.

1. **Analyse de la question** L'étape nécessaire qui doit analyser la question de l'utilisateur, formulée dans la langue naturelle, pour ensuite la traiter et l'envoyer à une autre composante du système de QR.
2. **Prétraitement des documents de références** Pour répondre à la question, le système doit avoir accès à des connaissances. Le prétraitement des documents permet de transformer les données disponibles en connaissances utilisables pour répondre aux questions.

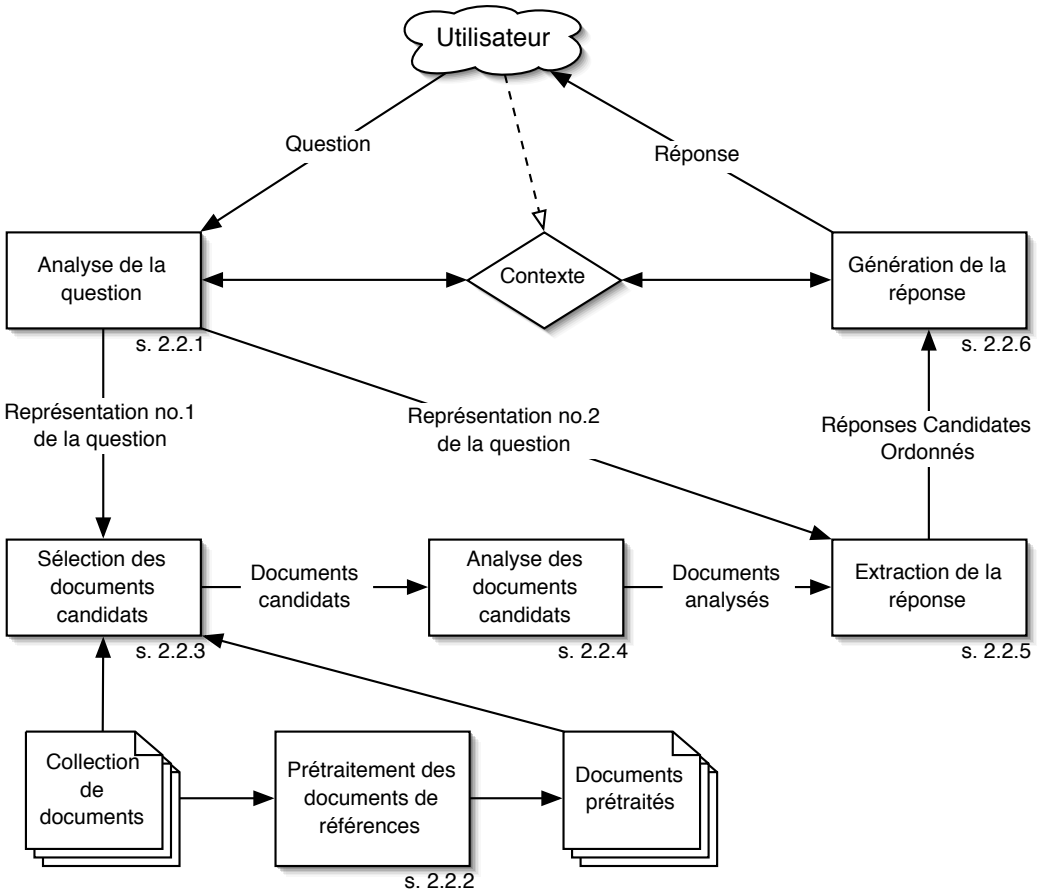

FIG. 2.2 – Architecture générale d'un système de Question-Réponse

3. **Sélection des documents candidats** Sélection d'un sous-ensemble de documents contenant possiblement une réponse à la question ou de l'information pertinente devant être analysée.
4. **Analyse des documents candidats** Analyse plus approfondie du sous-ensemble extrait précédemment dans le but d'extraire de l'information plus précise en rapport avec la question.
5. **Extraction de la réponse** Par l'utilisation d'une représentation adéquate de la question, un mécanisme d'appariement est réalisé pour extraire des candidats à la réponse qui sont ensuite ordonnés selon une mesure d'exactitude de la réponse.
6. **Génération de la réponse** Une réponse est retournée à l'utilisateur, celle-ci étant générée en fonction du contexte de la question définie par l'utilisateur ayant soumis la question.

2.2.1 Analyse de la question

L'analyse de la question est un élément primordial d'un système de QR. Si cette analyse est erronée, les chances de trouver une bonne réponse à la question sont sérieusement compromises. La tâche que l'on veut accomplir influence la façon dont la question est fournie au système. La question peut être dépendante du contexte, comme dans le cas d'une discussion, d'une analyse ou d'une application dans un domaine spécifique, ou indépendante comme c'est souvent le cas des utilisateurs du web qui veulent avoir une réponse précise à une question non spécialisée, mais généralement dépendante du temps.

Cette étape fournit en sortie une ou plusieurs représentations de la question qui seront utilisées lors des étapes subséquentes, elles seront utiles pour la sélection de l'ensemble de documents pertinents et pour l'extraction de la réponse. Une première représentation de la question est souvent réalisée par une réécriture sous la forme d'une requête (vectorielle, booléenne, probabiliste) à un engin de recherche. La requête sera créée en supprimant et en ajoutant certains mots qui permettront d'encadrer les mots-clés de la question pour exprimer le sens de la requête. Cette représentation servira à restreindre le nombre de documents à fouiller, mais sera pratiquement inutile lorsque viendra le temps d'identifier la réponse car elle est beaucoup trop vague.

Pour obtenir une représentation plus fine de la question, une analyse plus approfondie est nécessaire. La première chose à faire est de déterminer le type sémantique de la question, c'est-à-dire ce que l'on cherche. Ceci est facile lorsque la question contient des mots comme *where* (lieu), *when* (expression temporelle), *who* (personne) ou des expressions du type *what country*, *which president*. Cette facilité à identifier les principales expressions donnant un type à une question a incité plusieurs équipes de recherches à privilégier l'approche par construction de hiérarchies de types de questions pour leur système. Ces typologies de questions sont toutes construites manuellement soit d'un corpus de questions, des classes d'entités nommées ou de modèles linguistiques plus complexes.

Il faut ensuite déterminer les contraintes qui existent entre les mots-clés de la question qui seront utilisés pour trouver le paragraphe ou la phrase contenant la réponse. Ceci peut être aussi banal que d'extraire les mots-clés de la phrase qui seront comparés aux phrases candidates, ou complexe par l'utilisation de ressources lexicales sophistiquées et d'analyses syntaxiques complètes. S'il y a plusieurs entités nommées dans une question, il faut déterminer les relations syntaxiques

et sémantiques qui les unissent pour répondre à la bonne question. Cette analyse plus fine est nécessaire lorsque la question en contient une implicite comme dans *Who was the ruler of Egypt when Jesus Christ was born ?*, où la date de naissance de Jésus Christ doit être connue avant de pouvoir déterminer qui était le souverain de l'Égypte à ce moment. Cette analyse donnera comme résultat une représentation sémantique de la question qui sera utilisée pour établir les contraintes sur les réponses possibles en vue de ne conserver que les plus probables.

Dans le cas du système QUANTUM, l'étape d'analyse de la question a pour but de la classifier pour pouvoir déterminer quel module d'extraction utiliser pour trouver les réponses candidates. Au-delà de la classification, l'analyse de la question tente de déterminer le *type nominal* de la réponse à trouver. Cette analyse est réalisée par un ensemble d'expressions régulières appariées sur les mots et les étiquettes grammaticales, assignées au préalable sur chacune des questions. Le résultat de cette étape est une représentation de la question pour l'extraction de la réponse, il n'y a pas de représentation spécifiquement créée pour la sélection des documents candidats.

2.2.2 Prétraitement des documents de référence

L'étape de prétraitement des documents de référence est liée à la quantité de données à traiter. Elle est absolument nécessaire lorsque la tâche de répondre à une question doit se faire en temps réel, surtout si la banque de données est assez volumineuse, comme c'est le cas pour les systèmes de QR participant à TREC. L'approche la plus commune est d'utiliser une indexation de type *classique* tel qu'utilisée en recherche d'information, fondée sur le modèle booléen (MG), vectoriel (SMART) ou probabiliste (OKAPI).

Même si les documents candidats sont sélectionnés par un système de recherche d'information conventionnel, le prétraitement des documents peut accélérer la tâche, entre autre lors de l'analyse des documents candidats. L'information extraite (ou ajoutée) par le prétraitement est conservée en parallèle car l'ajout de cette information à la base d'indexation des documents peut biaiser la sélection des documents candidats. L'utilisation d'une étape de prétraitement évoluée augmente le temps global de calcul, mais permet d'obtenir une représentation logique ou sémantique du texte beaucoup plus riche que si elle avait été réalisée simultanément à l'analyse des documents candidats. Les prétraitements possibles des documents sont :

– l'analyse de surface qui extrait des patrons identifiant des éléments importants du texte ;
– l'étiquetage grammatical (*POS*) qui permet d'obtenir le rôle grammatical de chaque mot sans faire une évaluation syntaxique complète ;
– la reconnaissance d'entités nommées qui identifie les composantes intéressantes du texte comme les noms de personne, de ville, de compagnie, etc ;
– le découpage (*chunking*) qui réalise des groupements de mots selon leur relation fonctionnelle dans la phrase ;
– la dérivation de représentations logiques qui servira à réaliser des inférences sur les représentations ;
– l'annotation des termes par des marqueurs sémantiques selon l'information que l'on veut aller chercher ;
– l'extraction de relations sémantiques entre les entités.

Le système QUANTUM n'applique pas de traitement linguistique élaboré sur le corpus de données, la seule manipulation effectuée lors de cette étape est l'indexation du corpus par OKAPI pour la réalisation de l'étape suivante, la sélection des documents candidats.

2.2.3 Sélection des documents candidats

Comme il a été mentionné précédemment, la sélection des documents se fait généralement par l'utilisation d'un système de recherche d'information. Le but de cette étape est de restreindre le nombre de documents dans lesquels nous allons chercher la réponse. La première question à résoudre est de déterminer le nombre de documents qui seront sélectionnés, ou quel sera le seuil sur la mesure de similarité avec la requête pour considérer qu'un document ait des bonnes chances de contenir la réponse.

Un document est généralement une entité trop grosse pour l'extraction de la réponse, c'est pourquoi l'extraction de passages est un atout important lorsque vient le moment de choisir l'engin de recherche à utiliser. L'inconvénient d'utiliser un engin de recherche avec cette fonctionnalité est la nécessité de faire la mise au point des paramètres (longueur de la fenêtre, nombre de fenêtres par documents, distance entre les fenêtres). Pour pallier à ce problème, une fois les documents sélectionnés, on peut utiliser un segmenteur orienté vers un sujet pour identifier les passages de textes qui pourront ensuite être réévalués.

L'étape de sélection des documents candidats est exécutée de deux manières différentes dans QUANTUM. La première façon de procéder est d'utiliser la liste de documents pertinents provenant des organisateurs de la conférence TREC, générée à partir de la question et formulée comme une requête à un engin de recherche. Cette méthode ne semble pas être la plus efficace [11]. L'autre méthode qui utilise l'engin de recherche probabiliste OKAPI donne des résultats beaucoup plus intéressants dans le cadre de la compétition TREC-X. Cette méthode envoie en requête la question originale à l'engin de recherche OKAPI, dont les paramètres ont été établis de telle sorte que le résultat de la requête est une liste de passages extraits des documents pertinents. Les avantages de cette façon de procéder sont de pouvoir avoir plus d'un passage pertinent par document et de réduire la quantité de données à analyser lors de la prochaine étape.

2.2.4 Analyse des documents candidats

Une fois les documents (ou les passages) candidats sélectionnés, une analyse plus approfondie des textes restants est de mise. Si la collection de documents a complètement été traitée préalablement, cette étape ne sera pas absolument nécessaire. Présentement, les systèmes ne font pas un prétraitement systématique de toute la collection car elle est trop volumineuse pour être traitée efficacement avec les algorithmes connus, c'est pourquoi il est nécessaire pour les systèmes de réaliser une analyse plus détaillées des documents candidats.

La tâche minimale à exécuter lors de cette étape, pour qu'un système fonctionne bien, est l'identification des entités nommées, c'est-à-dire extraire et classifier les noms de lieux, de personnes, de compagnies, les adresses, les numéros de téléphones, les liens URL et les mesures. D'autres classes et sous-classes peuvent être ajoutées, celles-ci n'étant que les plus communes. Les autres traitements pouvant être exécutés sur les documents sont sensiblement identiques à ceux qui peuvent être appliqués lors du prétraitement de la collection, soit le découpage en phrase, l'annotation POS, le chunking, etc.

Outre ces analyses exécutées par des outils linguistiques *standard*, une approche utilisée régulièrement est l'analyse en surface. L'analyse en surface a pour but d'identifier les termes pertinents ainsi que ses dérivés (synonymes, hypernymes), pour ensuite modifier le rang des documents. Ces analyses retournent des représentations sous forme de contraintes ou de relations reliant les entités qui sont

identifiées dans le document. Certaines équipes vont même écrire cette représentation dans le langage de la logique du premier ordre. Le but de ces méthodes est de pouvoir déterminer si les entités susceptibles d'être une réponse sont dans l'ordre de la relation déterminée par la question.

L'étape d'analyse des documents candidats dans QUANTUM n'est pas isolée comme dans la description que je fais ici des systèmes de QR. Ceci est relié au fait que l'étape d'extraction des documents candidats retourne des passages et non pas des documents complets. L'analyse est réalisée sur un passage candidat, de longueur variant de un à trois paragraphes, par l'extraction des entités nommées et par une analyse syntaxique partielle, visant à mettre en valeur l'information pertinente pour l'extraction de la réponse.

2.2.5 Extraction de la réponse

À l'étape de l'extraction de la réponse, le système a déjà créé une représentation pour la question ainsi qu'une représentation des segments de textes susceptibles de contenir la réponse à la question. La tâche d'extraire la réponse est alors de faire un appariement entre la représentation de la question et les représentations des segments de textes, pour ensuite ordonner les réponses candidates selon leur « chance » d'être bonnes.

Pour faire l'appariement de la question avec les réponses candidates, le concept général est de créer des contraintes positives ou négatives. La première contrainte, utilisée par presque tous les systèmes de QR, est le type (catégorie sémantique) présumé de la réponse, donné par l'analyse de la question. Ainsi, la première étape est d'extraire les segments de texte qui contiennent des entités sémantiques correspondant au type attendu de la réponse ou à des types sémantiquement près, extraits de banques de données linguistiques (WordNet est régulièrement utilisé). Lorsque les segments pertinents ont été extraits, il faut leur appliquer d'autres contraintes provenant de la représentation de la question. L'application des contraintes peut se faire de façon absolue ou pondérée selon l'importance des contraintes.

Les systèmes qui créent des représentations de documents et des questions *riches de connaissances* (prédicats logiques, annotations sémantiques, relations de type <sujet,verbe,objet>) appliquent généralement les contraintes de façon stricte. Ces représentations ont l'avantage (ou l'in-

convénient) de pouvoir créer un nombre de contraintes plus grand que les systèmes utilisant des méthodes statistiques (*knowledge poor*). L'utilisation de *représentations riches* permet d'identifier les rôles dans la question et dans le texte, ce qui est pertinent lorsque la question fait intervenir un sujet (spécifié dans la question) et un objet (la réponse) du même type sémantique. L'appariement des contraintes est alors réalisé par un mécanisme de construction de preuve ou un système d'inférence. Par contre, cette approche n'est pas celle qui donne les systèmes les plus performants car l'élimination brutale des segments de texte entraîne une chute du taux de rappel que le gain en précision ne justifie pas. La conclusion à laquelle arrivent ceux qui utilisent cette approche est de ne pas traiter les contraintes de façon stricte, mais de seulement leur accorder une préférence.

La pondération des contraintes est possiblement moins précise, mais elle établit un ordre sur la pertinence des réponses. Cette approche est celle qui entraîne la plus grande diversité, elle est la plupart du temps établie à partir d'une heuristique. Une approche est de considérer les expressions retenues à la première étape de l'extraction de la réponse, d'établir une fenêtre autour de l'entité considérée et d'effectuer des mesures sur différents paramètres. Un paramètre calculé par plusieurs systèmes est le nombre de mots qui apparaissent simultanément dans la question, ou dans sa représentation, et dans la fenêtre (chevauchement). D'autres paramètres peuvent aussi être utilisés : le nombre de mots de la question qui ne se retrouvent pas dans la fenêtre, l'ordre des mots de la question par rapport aux mots de la fenêtre, la similarité des verbes apparaissant dans la question et la fenêtre. Ces paramètres sont ensuite combinés, souvent par une équation linéaire établie de façon heuristique, pour donner un poids global à chacun des passages. Une variante du calcul du score global est d'utiliser les techniques d'apprentissage automatique pour découvrir de bons paramètres à l'équation linéaire de pondération. L'application de ces contraintes ne se fait pas nécessairement dans cet ordre et sur une seule phrase ou fenêtre. La contrainte sur le type de la réponse peut être appliquée seulement à la toute fin après que les passages aient été pondérés. Le score d'une réponse candidate peut être obtenu à partir d'une combinaison de plusieurs fenêtres de longueurs et dispositions variables. Ces dernières techniques donnent des résultats intéressants, mais la justification de leur bon fonctionnement est difficile à réaliser.

Cette étape représente la composante la plus évoluée de QUANTUM, à partir de l'analyse de la question, le système sélectionne une fonction d'extraction définie dans ce module. Les fonctions d'extractions sont définies dans une hiérarchie orientée vers le type de réponse attendue, chacune

d'elle retourne un score permettant de déterminer la probabilité que la réponse soit bonne, pour ensuite les classer en ordre décroissant de probabilité. La hiérarchie des onze fonctions d'extractions est divisée en cinq catégories :

1. **Fonctions de hiérarchie** Ces fonctions sont la *définition* et la *spécialisation*, elles sont basées sur la relation d'hyponymie et d'hyperonymie. Pour réaliser leur tâche, ces fonctions d'extraction utilisent l'extraction d'entités nommées et les relations trouvées dans WordNet.
2. **Fonctions de quantification** Les fonctions *cardinalité* et *mesure* sont les deux fonctions de la catégorie de fonctions pouvant retourner des nombres. La fonction *cardinalité* est utilisée lorsque la question porte sur une quantité que l'on peut compter comme lorsque l'on compte sur nos doigts. La fonction *mesure* s'occupe d'extraire les nombres qui expriment une quantité, mais dans d'autres unités de mesure que le nombre d'entités de l'entité visée par la question.
3. **Fonction de caractérisation** La fonction de caractérisation *attribut* a pour tâche d'extraire des informations du même type que la fonction *mesure*, mais lorsque la caractéristique demandée n'est pas clairement définie dans la question.
4. **Fonctions de complétion de concept** Ces fonctions sont : *personne*, *temps*, *lieu* et *objet*. Les questions faisant appel à ces fonctions utilisent un pronom interrogatif pour faire référence à l'entité recherchée. Pour identifier les réponses, QUANTUM utilise l'extracteur d'entités nommées qui est présent dans le logiciel Alembic WorkBench.
5. **Autres fonctions** Les fonctions *manière* et *raison* n'appartiennent pas à une catégorie particulière. En fait, elles ne sont pas implémentées dans QUANTUM, mais elles correspondent aux questions de type *comment* et *pourquoi*.

Les fonctions d'extractions sont, pour la plupart, des expressions régulières sur les mots et leur étiquette grammaticale. Certaines fonctions nécessitent des paramètres provenant de l'analyse de la question, ceci permet d'avoir de l'information supplémentaire pour trouver la réponse. Le score retourné par l'étape d'analyse de la question de QUANTUM est une combinaison du score de la fonction d'extraction, du score donné par l'engin de recherche sur la pertinence du passage et d'un score de proximité. Le score de la fonction d'extraction est basé sur le niveau de confiance des développeurs envers la performance de la fonction.

2.2.6 Génération de la réponse

Puisque la plupart des systèmes de QR développés dernièrement le sont dans le but de bien performer à la conférence TREC, la génération de réponse n'a pas été un sujet de recherche dans les travaux récents sur les systèmes de QR. Depuis la conférence TREC-11, les réponses aux questions sont seulement la réponse, rien de plus, contrairement aux conférences précédentes où les réponses aux questions étaient des fenêtres de caractères[2], celles-ci n'étaient pas des réponses envisageables pour l'interaction avec les usagers. Pour que la réponse satisfasse l'usager, elle doit être intelligible et adaptée au niveau d'explications que l'utilisateur souhaite. Cette étape peut nécessiter une réécriture de la phrase en utilisant de l'information provenant d'autres phrases. La justification de la réponse peut être assez complexe et la réponse peut être la combinaison de plusieurs réponses.

2.3 Text Mining et Data Mining

Le but du *data mining* est de découvrir de l'information nouvelle à partir de données, trouver des régularités à partir de plusieurs ensembles de données et ainsi séparer l'information pertinente de celle qui ne l'est pas. Le *text mining* possède les mêmes buts, par contre les données utilisées pour y arriver sont différentes. L'utilisation de données textuelles modifie le problème, l'analyse des données ne se fait pas aussi facilement que lorsque les données sont numériques. Par exemple, la recherche d'information n'est pas à strictement parler du *text mining*, un système de RI retourne l'information que l'utilisateur demande. Ceci ne signifie pas que le système a découvert l'information, car elle était déjà connue de l'auteur puisqu'il l'a écrite dans le document.

Dans le cas du problème de la question-réponse, l'approche du problème par le *text mining* peut être pertinente. Les questions n'ont pas toutes le même degré de complexité, certaines questions se répondent facilement, comme les questions de type géopolitique, car l'information est souvent facilement identifiable dans les textes et elle est présente sous forme condensé dans des banques d'information (*CIA World Fact Book* par exemple.) Les questions que je considère complexes, sont celles qui demandent généralement un raisonnement. Les systèmes de question-réponse, dans leur état actuel, sont incapables de répondre correctement à ce type de question, même si la réponse

[2] 50 caractères lors de TREC-X

peut se retrouver dans la collection de documents. L'utilisation du *text mining* pourrait aider un système de question-réponse en lui fournissant un ensemble de connaissances qui serait extrait d'un ou plusieurs documents du corpus.

Chapitre 3

Étude des données

Les données du problème du traitement automatisé des courriels dans la problématique du service des relations avec les investisseurs sont différentes de celles utilisée dans les conférences TREC. La première différence étant que les courriels ne sont pas toujours des questions, et lorsqu'ils en sont, la tâche de retrouver le sens de la question est beaucoup plus complexe. La base de documentation est beaucoup plus restreinte, ce qui permet d'appliquer un pré-traitement qui est peu coûteux sur le corpus de recherche. Elle est aussi beaucoup plus spécialisée, ce qui implique que les connaissances doivent être ciblées vers le domaine des relations avec les investisseurs. La longueur des réponses n'est pas uniforme, parfois la réponse est un seul mot ou un seul nombre, d'autres fois la seule bonne réponse acceptable sera une page web au complet. Dans TREC, les bonnes réponses aux questions sont toujours bonnes, alors que dans le corpus BCE-4, une bonne réponse sera parfois mauvaise ou inappropriée car elle pourra être trop ou pas assez détaillée pour les besoins de l'utilisateur. La problématique de la granularité des réponses sera évidente lorsque je présenterai des extraits du corpus dans la section suivante.

La description de corpus que j'ai réalisée porte sur le corpus nommé BCE-4, d'autres analyses de corpus ont déjà été réalisées sur les corpus BCE-{1,2,3}. Le premier corpus reçu est BCE-1, il s'agit de 141 messages envoyés entre avril et septembre 2000. Les messages ont été produits à l'aide d'un formulaire de commentaire accessible à partir du site web de BCE. Ce corpus est composé de messages de nature générale et variée, à propos du site web, des contacts, des relations aux investisseurs et des plaintes sur la qualité du service. Ce corpus a permis de faire une analyse préliminaire

du domaine des messages reçus. Le corpus BCE-2 est particulier car il a été reçu en format imprimé, donc pratiquement impossible à analyser avec des outils informatiques sans avoir procédé à un traitement de numérisation. Ce corpus est constitué de 865 messages avec le suivi réalisé par les préposés de BCE pour chacun d'eux. L'avantage de BCE-2 est qu'il contient un plus grand nombre de messages que BCE-1, il a été utilisé pour vérifier les hypothèses émises lors de l'analyse de BCE-1 ainsi que pour établir la proposition du projet de réponse automatisée aux courriels. BCE-3 est un corpus de 1568 paires de messages avec leur suivi envoyés à `investor.relations@bce.ca` entre juin 1999 et novembre 2000. Ce corpus est le premier qui soit représentatif du domaine des relations aux investisseurs, il a été étudié en profondeur par Julien Dubois [3]. Dans la prochaine section, j'expose les observations que j'ai réalisées lors de l'analyse du corpus BCE-4, celle-ci a pour but de bien situer les problématiques reliées au traitement des courriels dans le cadre d'un champ d'application précis comme les relations avec les investisseurs.

3.1 Description du corpus BCE-4

Le corpus BCE-4 est composé d'environ 1200 courriels envoyés à `investor.relations@bce.ca` et à `relations.investisseurs@bce.ca`, rédigés en anglais et en français. J'ai réalisé l'analyse du corpus BCE-4 en examinant autant les courriels rédigés en français que ceux rédigés en anglais. Suite à cette analyse, j'ai décidé de ne traiter que les courriels rédigés en anglais. Cette décision est motivée par le fait que 80% des courriels sont rédigés en anglais, ce qui ne donne pas assez de données pour pouvoir traiter efficacement l'utilisation du français dans les courriels des relations aux investisseurs de BCE. Les courriels de ce corpus ont une longueur variant de une à cinq phrases, il y a bien sûr des exceptions, mais règle générale cette approximation est suffisante pour avoir un aperçu du corpus.

Pour avoir une meilleure idée des problèmes auxquels je dois faire face, j'ai classifié les courriels selon qu'ils demandent une réponse ou non. J'ai donc considéré les courriels qui nécessitaient une réponse de la part d'un préposé comme une question. Selon cette méthode de classement, j'ai déterminé qu'environ 25% des courriels sont des questions et que 35% sont pertinents mais ne peuvent être considérés comme des questions. Dans cette dernière catégorie, nous retrouvons les messages faisant état de changements d'adresse, de demandes de documents et d'inscriptions à

différents événements organisés par le groupe « investors relation ». Les autres courriels (40%) sont des messages non sollicités (pourriels), des virus ou n'étaient pas destinés à BCE.

Dans un deuxième temps, j'ai extrait les courriels jugés intéressants, c'est-à-dire ceux qui peuvent être vus comme des questions et je les ai reclassés en sous-catégories. Même si ces courriels sont classés dans les questions, cela ne veut pas dire pour autant qu'ils peuvent être répondus aisément. Cette seconde étape de classement est fondée sur les observations réalisées au cours de la première classification. Cette classification des messages est réalisée dans le but d'établir des hypothèses de travail solides et orientées spécifiquement vers notre problème. Les thèmes que j'ai retenus pour les catégories sont énumérés dans le tableau suivant (TAB. 3.1) :

Catégories	Section	Nombre	Nb. répondable
contact	3.1.1	17	7
dossiers personnels	3.1.2	31	0
date d'événements	3.1.3	25	23
finance et corporations	3.1.4	66	31
comment investir	3.1.5	36	16
prix des actions	3.1.6	62	40
divers		19	9
Total		256	126

TAB. 3.1 – Distribution de l'ensemble des messages

Les courriels ont été classifiés selon le but de la question telle qu'elle peut être interprétée par un humain. Certains courriels auraient pu être placés dans deux catégories, soit parce qu'ils contiennent plusieurs questions dans le même message, soit parce que leur catégorisation était ambiguë. L'ambiguïté des messages provient souvent des courriels dont le contenu est ponctuel. J'ai ajouté un paramètre à la catégorisation, le nombre de questions jugées répondables. Cette catégorie est subjective selon la personne qui lit le courriel. J'ai tenté de répondre à la requête pour chaque courriel. Les critères que j'ai utilisés pour déterminer si une question est répondable sont : la facilité d'interprétation du courriel, l'ambiguïté de la requête, la présence de l'information nécessaire dans le courriel, l'hypothétique accessibilité à l'information.

3.1.1 Contact

La catégorie *contact* est constituée des courriels demandant comment contacter (par courriel ou par téléphone) une personne précise dont le nom est mentionné, le responsable d'un département, la personne en charge d'exécuter une tâche précisée dans le courriel ou obtenir l'adresse d'une succursale. Ce genre de questions semble être facile à répondre automatiquement, pour peu que la question ne soit pas perdue dans un long texte indéchiffrable.

Ci-dessous, j'ai extrait les phrases clés de certains courriels de cette classe. La plupart de ces courriels possèdent une réponse connue et qui se retrouve avec une certaine facilité. De plus, les questions sont posées clairement et l'information nécessaire pour trouver la réponse est contenue dans le courriel. Dans les exemples (1.1) et (1.3) qui suivent, le nom de la personne est clairement indiqué, respectivement John Doe et Foo Bar, pour chacun d'eux. Dans l'exemple (1.2), le nom n'est pas mentionné, mais la fonction de la personne est par contre clairement indiqué.

(1.1) Kindly provide Mr. John Doe's e-mail address. I am a former Vice-President of CBRS ...
(1.2) Please advise the name and phone number of your head of Investor Relations.
(1.3) Could you please send me an email address and phone number for Foo Bar ?

Par contre, certains courriels sont plus difficiles à analyser et/ou l'information pour y répondre n'est peut-être pas facilement accessible pour un système automatisé. En voici donc quelques exemples :

(2.1) Please could you let me know if you have any offices in the Caribbean, and if so, where are they and could I possibly have some contact details/emails and addresses ?
(2.2) Is there a phone number for the Analyst Meeting on the 12th, instead of watching over the webcast ? Please forward the telephone numbers to listen in to the meeting.
(2.3) We are about to send out a survey via email regarding your DRP/DSPP plan and would like to make sure that it goes to the correct person within your department. ... Can you kindly supply us with the name and email address of this person.

Ces exemples demandent des traitements plus délicats et dans le cas de l'exemple (2.2), la question est ambigüe si le complément d'information à propos du *webcast* est absent. Dans les exemples (2.1) et (2.3), la problématique est plus complexe, dans l'exemple (2.1) il n'y a qu'une seule réponse à

trouver, mais l'équivalent de deux questions à traiter. De même dans l'exemple (2.3) il faut être capable de découvrir qui est celui qui s'occupe d'une certaine tâche dans *your department*.

Pour répondre aux requêtes répondables facilement, il faut pouvoir compter sur un annuaire des personnes susceptibles d'être contactées directement par les clients. Cet annuaire doit être assez descriptif par rapport aux tâches et aux responsabilités de chacune des personnes qui en font partie. En plus de contenir de l'information à propos des individus de l'entreprise, l'annuaire devra contenir de l'information relative à chacun des départements, bureaux et compagnies que l'entreprise possède. L'ensemble de ces connaissances peut être encodé selon un modèle énumérant les différentes propriétés d'une entité ou un modèle définissant les relations entre les différentes entités représentées dans le corpus où la réponse est probablement située. Un tel annuaire peut possiblement être construit à partir de l'analyse des documents publiés par BCE, par des méthodes de forage d'information appliquées aux données textuelles. L'utilisation de lexiques adaptés au domaine de la finance, et particulièrement au monde des relations aux investisseurs, est une approche qui peut aider à la génération de représentation des connaissances pour situer les entités auxquelles les courriels font référence. La représentation de ce type de connaissances devra aussi tenir compte du temps car les différentes actions que l'entreprise exécute rendent certaines informations désuètes. Par exemple, les nominations à des postes clés de l'entreprise entraînent une modification de la structure organisationnelle.

3.1.2 Dossiers personnels

Dans la catégorie dossiers personnels, j'ai regroupé tous les courriels qui demandaient de l'information à propos de cas de particuliers. Parmi ces courriels, il y a les cas de successions, la recherche d'actions perdues, la valeur des actions que le client possède et ceux qui demandent s'ils ont reçu leurs papiers. Les courriels de cette catégorie sont ceux qui sont le moins susceptibles d'être traités de façon automatique, chaque cas demandant un traitement particulier. Cette constatation provient de deux facteurs, le premier est l'imprécision des demandes, le second est un problème évident de sécurité de l'information. L'imprécision des demandes est attribuable au fait que les auteurs de ces courriels ne donnent pas assez d'information à propos de leur dossier et qu'ils ne connaissent pas assez le domaine de la finance pour pouvoir poser une question précise.

Cette catégorie met en évidence les différences qui peuvent exister entre les utilisateurs. Un utilisateur expérimenté, connaissant bien son domaine, pourra exprimer sa question clairement et donner assez d'information pour qu'un préposé (homme ou machine) puisse lui répondre clairement et simplement. Par contre, lorsque l'utilisateur semble néophyte, un système de réponse automatique devra pouvoir décider d'engager une conversation avec cet utilisateur pour qu'il précise sa requête et/ou qu'il donne de l'information supplémentaire pour son traitement. La majorité des personnes qui ont envoyé des courriels à propos de dossiers personnels sont des petits investisseurs, des exécuteurs testamentaux et des courtiers. Les deux premiers types de personne ne connaissent pas beaucoup le domaine de l'investissement, leurs questions sont imprécises et il apparaît un manque flagrant d'information pour leur répondre. Dans la plupart de ces cas la réponse doit être simple, au sens où les menus détails ne sont pas de la plus grande importance, mais tout de même complète pour satisfaire le besoin d'information du client. L'autre catégorie de personne est celle des *investisseurs éduqués*, qui fournissent généralement l'information nécessaire pour que le préposé puisse répondre à leurs requêtes. La problématique de cette catégorie est lié au traitement précis de l'information, l'analyse du message ne peut pas se permettre d'être approximative, car la demande d'information est généralement très bien ciblée.

Ci-dessous, je présente quelques exemples. Le premier (3.1) est un cas où l'information est incomplète, il est donc impossible d'y répondre précisément. De plus, dans cet exemple, on remarque que l'auteur du courriel ne semble pas être un expert du monde de la finance, dans ce cas-ci la réponse par courriel ne pourra être très précise. L'exemple suivant (3.2) est répondable puisque l'information est complète et la question est clairement formulée. Il est intéressant parce qu'il met en évidence que l'analyse du courriel ne doit pas être faite seulement sur le texte, mais aussi sur l'information complémentaire, comme le sujet du courriel où l'information à propos du certificat en question se localise. De plus, l'information demandée par l'auteur du message devra être retournée de façon précise et correcte car l'auteur semble connaître le domaine et savoir ce qu'il recherche, mais il n'a pas nécessairement accès aux outils disponibles pour aller chercher lui-même l'information.

(3.1) I am the executor for my father's estate.
I have not record that at his death Sept. 6th, of this year he held any BCE stock.
Could you please check your records and determine if any are held at the moment.

(3.2) **Subject :** Share Certificate # DC RR928007
...
Can you please advise the history, current status and value of this certificate ?

D'autres exemples pourraient être extraits, mais la principale caractéristique de ces courriels

est l'uniformité de la formulation. Dans des cas comme ceux-ci, ce que je crois possible de réaliser est de répondre à l'utilisateur par une formule préétablie lui signifiant quelle démarche il doit faire pour avoir les réponses à ses interrogations.

3.1.3 Dates d'événements

Les dates d'événements sont des courriels possiblement faciles à répondre automatiquement à condition d'avoir deux choses, une chronologie des événements passés et un agenda des événements à venir. Les types d'événements que le service de relations aux investisseurs semble traiter sont : les dates d'émission et de division d'action, les dates de remise de dividendes, les dates d'émission et d'encaissement des obligations ainsi que les dates des différentes conférences où les investisseurs sont invités à prendre part. Ces courriels font généralement référence au temps de façon indirecte par l'utilisation des marqueurs temporels de type *next*, *last*, *since*. Les exemples ci-dessous proviennent des messages que j'ai classifiés dans la catégorie des courriels faisant référence au temps ou aux dates d'événements.

(4.1) Could you please let me know the date and location of your 2002 annual meeting. Also, do you have a list of the release dates for your 2002 quarterly results ?

(4.2) Would you be able to supply me with the date and location of BCE's next AGM ?

(4.3) Could you please advise as to when the rates will be posted ?

(4.4) **Subject :** When will BCE be reporting its Q4 2001 results ?

(4.5) Would you advise me of the BCE stock splits since February 1994.

(4.6) According to our records, we are expecting a dividend announcement from you. Please could you let me know when this announcement is scheduled to happen (approximately), i.e. :

1. The expected ANNOUCEMENT/DECLARATION date for the dividend to be paid in JANUARY 2002.
2. Could you also confirm the PAY and RECORD date for this dividend yet to be announce ?
3. In addition could you also tell me when your next Annual Shareholders Meeting is ?

(4.7) Can you please confirm what time the first presentation will begin at for the meeting on December 12th.

L'exemple (4.1) est typique de ceux qui demandent de l'information sur des événements à venir. Ce type de message est généralement court, c'est-à-dire de une à trois phrases courtes, et la demande d'information est formulée de façon très précise. L'exemple (4.2) est du même type

sauf qu'il utilise l'abréviation *AGM* dont la signification est *Annual General Meeting*, ceci met en évidence la nécessité de traiter les collocations. Il sera donc très important d'avoir un système qui puisse reconnaître les abréviations et les collocations pour pouvoir bien cibler la date et le lieu de l'événement dont l'utilisateur s'enquérit. Une problématique similaire est identifiable dans l'exemple (4.3) où l'objet de la question est mal défini. Lorsque l'utilisateur mentionne *rates*, il est impossible (même avec le contenu complet du message) de déterminer de quel taux il est question, de plus l'expression *to post the rates* ne semble pas s'interpréter de la même façon dans le domaine des investisseurs que dans le domaine des connaissances générales. L'exemple (4.4) démontre, comme l'exemple (3.2) d'ailleurs, que toute l'information disponible doit être utilisée, car dans ce cas le message est vide et le seul contenu est la question qui se retrouve dans le champ **Sujet** du courriel. L'extrait de l'exemple (4.5) portant sur les *stock splits* est fréquent dans le corpus, j'ai classé seulementt trois courriels comme celui-ci dans la catégorie de *dates d'événements*. Par contre, je peux en retrouver douze autres dans la catégorie *prix des actions* qui pourrait être considérée comme une sous-catégorie des événements. Ce type de question se formule de différentes façons, certaines questions peuvent se répondre directement alors que d'autres demandent un processus de raisonnement, comme dans le cas présent, où un compte doit être réalisé. Dans l'exemple (4.6), il y a trois questions, parmi ces questions on retrouve un cas de réutilisation de réponse avec les deux premières questions. Ces deux questions, tout comme la troisième, concernent des événements futurs dont la réponse n'est possiblement pas encore connue. Il faut alors être capable de répondre à l'utilisateur que la date n'est pas encore établie ou encore lui retourner la date à laquelle celle-ci sera déterminée. Le dernier exemple (4.7) est très précis, il demande un programme des présentations de la journée de la rencontre du 12 décembre.

Dans la plupart des exemples précédents, les questions sont bien formulées et leur traitement ne semble pas poser de problème pour un préposé humain car presque tous les courriels de cette catégorie peuvent être caractérisés comme répondables (23 sur 25). Les problèmes pour le traitement automatique des courriels de cette catégorie se situent au niveau de la recherche de la réponse parce que l'information recherchée pour retrouver la réponse ne s'extrait pas facilement, particulièrement à partir de données textuelles. Le prétraitement des documents de référence par un étiquetage automatique de l'information à propos des événements dans les publications de la compagnie est une étape qui rendrait l'extraction de réponse plus facile pour un module traitant le temps et les événements.

3.1.4 Finance et corporations

Cette catégorie est composée de messages concernant les finances de BCE, c'est-à-dire le fonctionnement même de l'entreprise, la façon de faire les calculs comptables et les différentes questions que se posent les investisseurs par rapport aux publications de BCE, entre autres, le rapport annuel. J'ai aussi inséré dans cette catégorie tous les courriels concernant la structure de BCE, le nombre d'employés, la répartition des investisseurs ainsi que ceux demandant de l'information « commerciale » à propos des produits de BCE. Plusieurs de ces questions pourraient être répondues automatiquement, mais l'information demandée est possiblement confidentielle, ce qui pose un problème supplémentaire à un système automatique.

Les courriels concernant la finance sont parfois très difficiles à comprendre, même pour les utilisateurs humains. C'est pourquoi je n'ai pas beaucoup d'espoir pour qu'un système automatique puisse répondre à ce type de question. Les principaux problèmes que j'y vois sont la technicité du langage et l'interprétation, faite par les différents intervenants, de ces termes techniques. De plus, l'information utilisée pour répondre à ce type de requête est généralement dépendante du temps ; comme dans l'exemple (5.1) ci-dessous, où le pourcentage d'actions détenues par BCE à titre d'actionnaire dans Nortel Networks est variable selon la date à laquelle l'information est demandée.

(5.1) Please tell me if BCE Inc. still owns any shares of Nortel Networks Corp., and if so, what percentage of Nortel's outstanding shares does BCE Inc.'s holdings represent.

(5.2) Hi, I am a fourth year student at Wilfrid Laurier University and have an assignment about BCE for my Investments Management class.
I am hoping that you will be able to give me the following information.
The P/E Ratio for 1995-2000.
The average stock price for each year 1995-2000
The market and book value of the common shares.

(5.3) Would you please be able to let me know what were your EBITDA for 1998, 1999 and 2000 and how they were calculated (actual earnings, interests, taxes, depreciation and amortization figures).

L'exemple (5.2) ci-dessus est un courriel qui contient trois questions, mais en fait le besoin d'information de cet étudiant ne se limite pas à trois réponses. La réponse aux questions contenant des intervalles de temps comme 1995-2000 doivent être traités avec attention car dans ce courriel la réponse devra contenir l'information pour l'ensemble des années demandées. En procédant à l'analyse de chacune des questions, je peux voir le type de problème qui m'attend pour réaliser

l'analyse automatique. La première requête de cet exemple peut être interprétée de deux façons, la première interprétation est que la personne veut avoir le *price per earnings ratio* (*P/E Ratio*) pour l'ensemble de la période 1995-2000. La deuxième interprétation qui peut être faite est influencée par le contexte (la deuxième question), celle-ci est que la personne demande en fait le *price per earnings ratio* pour chacune des années de 1995 à 2000. La deuxième question possède une interprétation claire, la personne établit spécifiquement par l'utilisation de *each year* qu'elle veut l'information pour chacunes des années. La troisième question laisse place à l'interprétation, tout comme la première.

Le dernier courriel (5.3) possède des caractéristiques intéressantes pour être jugé répondable. La première partie de cet exemple ne demande qu'à savoir ce qu'est un EBIDTA et d'avoir les données correspondants aux années voulues pour y répondre, l'analyse de cette phrase semble assez claire. Par contre, la deuxième partie fait appel à un raisonnement plus complexe, où on demande comment sont calculées les différentes données financières. Ce type de question peut poser des problèmes à deux niveaux : le premier problème étant au niveau de la représentation de cette connaissance, l'extraction de cette information à partir de données textuelles est une tâche qui demande une précision parfaite de la part d'un engin de forage d'information. Le deuxième problème est au niveau de la réponse à ce type de question, est-ce que la réponse devra être une formule mathématique avec des variables et des chiffres, ou bien seulement le nom d'une méthode connue du domaine de la finance ? Cette deuxième problématique nous ramène encore une fois au problème de la granularité de la réponse qui doit être établie lors de l'analyse du courriel pour pouvoir constituer une réponse satisfaisant l'utilisateur.

3.1.5 Comment investir

Ce sont les courriels des clients (ou futurs clients) qui demandent comment investir dans les différents produits financiers que BCE propose, auxquels j'ai ajouté les courriels de ceux qui veulent savoir comment réinvestir leurs dividendes. La plupart des questions de cette catégorie, une fois extraites, ont des réponses qui sont identifiables dans le site web de BCE. Dans certains cas, la question formulée dans un courriel est similaire à la description de certains documents mis à la disposition des investisseurs par BCE. Les réponses à ce type de courriels sont longue, c'est-à-dire qu'elle ne peuvent généralement pas être répondues simplement par une expression ou une phrase.

La réponse sera la description d'une procédure, quelle personne contacter ou même l'impossibilité de l'opération, ce qui sera difficile à déterminer automatiquement.

Hormis les courriels de la catégorie dossiers personnels, cette catégorie est sans doute la plus difficile à traiter automatiquement. La principale raison est la difficulté de la tâche de génération de réponse, les réponses à cette catégorie de questions peuvent avoir des répercussions très grandes sur un investisseur si celui-ci reçoit une information erronée ou une explication qui manque de précision. De plus, pour la plupart de ces courriels, même s'ils contiennent une question analysable et répondable, la suite logique est une discussion avec un préposé ou directement avec un courtier. Un autre facteur qui influence la réponse est le lieu de résidence de l'investisseur, une facette du problème qui m'était inconnu avant d'examiner certains documents provenant du site web de BCE, ainsi les procédures d'investissements au Canada, aux États-unis et ailleurs dans le monde ont toutes leurs particularités, ce qui complique la problématique d'extraction de la réponse.

3.1.6 Prix des actions

Cette dernière catégorie de courriels est celle qui contient le plus de courriels, ce sont principalement des actionnaires ou ex-actionnaires qui demandent le prix d'actions à différentes dates. Parmi ces courriels, il y a tous les investisseurs qui demandent comment calculer la formule d'allocation reliée au « spin-off » de Nortel, ceux qui veulent savoir comment calculer le coût de base ajusté (pour faire leur rapport d'impôt) et toutes les autres questions reliées au calcul du prix des actions. La difficulté de répondre à ces courriels se situe principalement au niveau de l'extraction de l'information contenue dans le courriel. Si cette information est complète et extraite convenablement, la tâche est alors d'exécuter des calculs bien définis à partir de données existantes sur le site web de BCE, plus particulièrement, les cours de fermeture quotidiens des actions ordinaires de BCE depuis janvier 1983 présentés dans un fichier de chiffrier électronique (MS Excel.)

Les questions qui causeront des problèmes seront celles qui feront appel à des données qui ne sont pas facilement accessibles ou bien à des calculs spécifiques, comme dans l'exemple (6.1), où le prix moyen des actions (*average trading price*) n'est pas nécessairement connu et où il faut possiblement le calculer selon les formules comptables de BCE.

(6.1) Could you provide me with the average trading price for BCE for the years 1970 and 1979.

(6.2) Please advise the closing price of BCE common shares on April 27, 1998.

(6.3) COULD you please let me know what the annual dividend rate has been established at for the Series T Preferred Shares.

L'exemple (6.2) est le type de question qui se retrouve le plus souvent dans cette catégorie, soit une demande du prix de fermeture de l'action de BCE à une date précise. L'exemple (6.3) est un cas qui peut amener un système de réponse automatisé à avoir des problèmes, la raison étant que les actions privilégiées de série T ont été converties en actions d'une autre série. Ainsi, pour pouvoir répondre à cette question, il faut être capable d'aller extraire toute l'information possible. Dans ce cas ci, le taux demandé n'est pas indiqué sur le site web, c'est un exemple où le *text mining* peut possiblement nous être utile, et où la réutilisation des connaissances peut grandement améliorer les performances d'un système de réponse automatique. Dans ces exemples, la question de la temporalité de l'information est primordiale. Pour les deux premiers, les dates sont indiquées de façon très claire, mais pour le dernier la temporalité apparaît de façon plus subtile. La temporalité dans cet exemple réfère à la chronologie des événements se rapportant à un titre, les actions de cet exemple ont une durée de vie. Si le message est reçu dans l'intervalle de l'existence de ce titre, le problème de l'existence du titre ne se présente pas, dans le cas contraire, il faut être capable de déterminer si l'information que nous retrouvons dans les données est encore pertinente au moment où la question se pose.

3.2 Conclusion de l'étude du corpus BCE-4

La première chose à remarquer dans ce corpus est l'utilisation du jargon de la finance qui contribue à l'interprétation ambiguë de certains messages, certaines expressions ont une interprétation qui diffère selon qu'elles sont utilisées dans le domaine de la culture générale ou dans le domaine des relations aux investisseurs. Une avenue de solution à cette problématique est l'utilisation de ressources lexicales spécialisées pour le domaine de la finance et des relations aux investisseurs. En plus de ces ressources *linguistiques*, l'utilisation d'autres ressources spécialisées donnant la *sémantique* ou la formule mathématique de certains termes et expressions serait de mise. Ce genre de connaissances est essentiel pour répondre à des courriels comme celui de l'exemple (5.2) où le *P/E Ratio* est l'expression raccourcie de *price per earnings ratio* qui se calcule par l'équation

$$\text{Price per earnings ratio (P/E)} = \frac{\text{Adjusted trading price at the end of the period}}{\text{Earnings per share}}$$

L'extraction, la représentation et l'utilisation de ces connaissances est une difficulté particulière du problème de réponse aux courriels dans le cadre du service aux utilisateurs qui ne se retrouve pas dans les recherches effectuées présentement dans les systèmes de QR.

L'ensemble des données qui vont être utilisées pour solutionner ce problème est de petite taille en comparaison avec ce qui est utilisé pour les systèmes de questions-réponses de la conférence TREC. En faisant cet énoncé, je suppose que je n'ai accès qu'aux données disponibles sur le site web de BCE. Cet état de fait peut amener à la fois des avantages et des inconvénients pour le traitement automatisé des courriels, les avantages étant plus nombreux. Le premier avantage à noter par l'utilisation de ce petit corpus de référence est le peu d'information inutile que contient le corpus, ceci rend la tâche de sélection de documents pertinents beaucoup plus aisée car la première étape d'analyse des données peut se faire plus rapidement. L'information contenue dans les données est précise et ciblée pour le domaine que nous traitons, c'est-à-dire les investisseurs. Cette précision de l'information permet d'utiliser ou de développer des algorithmes qui sont capables d'extraire avec précision l'information ciblée par les courriels au profit de la robustesse. De plus, une partie de l'information qui nous sera utile pour répondre aux questions « numérique » est déjà dans une représentation utilisable présentement sans avoir à utiliser d'algorithmes d'extractions particuliers.

Comme je l'ai mentionné précédemment, l'information que nous avons à traiter est dépendante du temps, une question n'a possiblement pas la même réponse d'une journée à l'autre. Il est alors très important d'avoir des représentations de l'information qui tiennent compte de la problématique de la modification de l'information à travers le temps. Au delà de l'aspect strictement temporel, il y a l'ensemble des événements de la vie corporative qui modifie constamment les données. Dans le cas présent, l'événement entourant la vente des actifs de BCE dans la compagnie Nortel Networks a entraîné une modification de la valeur de certaines actions et ceci a aussi modifié la structure organisationnelle de l'entreprise BCE. Ce problème ainsi que les autres mentionnés précédemment font en sorte que la problématique de la réponse automatisée aux courriels est intéressante à étudier dans le but de développer de nouvelles méthodes dans le traitement de l'information comme je le présente dans le prochain chapitre.

Chapitre 4

Plan de recherche

4.1 Objectifs de la recherche

Le but de la recherche que je vais effectuer est de développer de nouvelles méthodes pour améliorer les systèmes de question-réponse. Le développement de ces méthodes se fera dans le contexte de la réponse automatisée aux courriels. La problématique que j'étudie diffère de celle étudiée dans les systèmes de QR. Le domaine d'application, le service de relations aux investisseurs, est beaucoup plus restreint que celui des conférences TREC ou des systèmes de QR développés pour le web. Cette caractéristique du problème permet d'envisager des solutions faisant appel à des connaissances provenant du domaine d'application, soit par l'utilisation de lexiques ou d'ontologies spécifiques au domaine des relations aux investisseurs. L'avantage d'avoir un contexte précis pour effectuer la recherche introduit des contraintes pour l'élaboration de solutions et la principale est liée à des facteurs humains. L'auteur du courriel envoie sa requête dans le but d'avoir une information précise. Le fait de délimiter le domaine d'application doit permettre d'atteindre un niveau de précision élevé, les réponses approximatives ne pourront jamais être considérées comme satisfaisantes.

La nature des données est un autre aspect du problème qui diffère de celui abordé par les systèmes de QR ordinaires. Une partie des données est composée des questions qui proviennent de courriels rédigés par des utilisateurs qui ont un besoin d'information ; l'autre partie est l'ensemble

des documents de références dont la fonction originale est de fournir l'information aux investisseurs. Les courriels ont la particularité d'être rédigés dans un langage qui contient des fautes d'orthographe, des abréviations, des expressions utilisées exclusivement dans les courriels et l'utilisation abusive de certains signes de ponctuation comme le point d'exclamation peut souvent porter à confusion. Les questions qui sont contenues dans les courriels sont aussi beaucoup plus spécialisées que celles des systèmes de QR traditionnels et sont exprimées dans un contexte dont il faut tenir compte pour les analyser correctement. L'ensemble des documents de références est composé de différents types de documents, dans cet ensemble on retrouve des rapports annuels et trimestriels, des communiqués de presse et des tableaux de données. La diversité de l'information retrouvée dans ce corpus de référence est une problématique nouvelle pour les systèmes de QR, les systèmes traditionnels utilisent pour la plupart un corpus d'articles de journaux , de revues ou de nouvelles provenant du fil de presse.

Les différentes particularités de notre problème rendent mon projet de recherche différent de celui abordé dans l'étude des systèmes de QR tel que réalisé présentement dans le cadre des conférences TREC. La création d'un système de réponse automatique aux courriels est une problématique qui se divise en plusieurs tâches tout comme la création d'un système de QR. Dans l'optique d'utiliser l'approche par question-réponse pour réaliser ce projet, je compte développer un système de QR, possiblement dérivé de QUANTUM, qui puisse analyser le courriel et retourner l'information pertinente à la rédaction d'une réponse. En comparaison à l'architecture présentée à la figure 2.2 (p.16), je ne compte pas étudier la problématique de la génération de la réponse. Suite à la description de la problématique que j'ai fait précédemment, j'ai extrait certaines caractéristiques qui serviront de balises pour la création et l'évaluation des composantes du système. La principale caractéristique pour qu'un système de réponse automatique aux courriels, appliqué à notre problématique, donne des résultats satisfaisants est d'avoir une mesure de précision de nos divers modules qui soit toujours très élevée. Cette mesure de précision peut, en fait, être reliée à un taux de confiance du système, celui-ci doit être capable de déterminer systématiquement si la réponse est bonne ou non. Ceci aura pour effet de diminuer le taux de rappel au profit de la précision. Par contre, dans le problème que je présente cette façon de faire est essentielle car les conséquences d'une réponse erronée peuvent être beaucoup plus graves que celles d'avoir à confirmer manuellement la validité de la réponse.

Les grands objectifs de la recherche sont de développer de nouvelles idées pour l'analyse des

courriels et pour le forage et l'extraction d'information pertinente à la problématique du service aux investisseurs. La problématique qui sera abordée en premier dans mes travaux sera l'analyse des courriels. Le développement de ce module d'analyse ne sera pas réalisé de façon à seulement solutionner le problème, j'ai l'intention d'explorer de nouvelles avenues particulièrement en ce qui a trait à la représentation sémantique de l'information que l'on retrouve dans les courriels et à explorer des méthodes de classification différentes de ce qui a été réalisé par Julien Dubois [3]. Le taux de satisfaction que je compte atteindre pour l'étape d'analyse des courriels est difficilement quantifiable, parce qu'il n'existe pas de mesure pour déterminer si un courriel est analysé correctement. Les résultats que j'utiliserai pour évaluer si l'analyse est bien effectuée proviendra d'une évaluation réalisée par des humains qui quantifieront la qualité de l'information retournée aux différents autres modules qui constitueront le système de réponse automatisé. Le second objectif que je crois réalisable se situe au niveau du forage et de l'extraction de l'information à partir du corpus de données. Les travaux reliés à cette étape du développement d'un système de traitement automatisé des courriels sont importants car leur utilité ne se limite pas simplement à notre problème, le but est de rendre les solutions proposées utilisables pour les autres domaines de la recherche en linguistique informatique. De plus, le résultat de cet objectif peut être utilisé pour améliorer le système d'analyse des courriels en fournissant de l'information spécifique au problème. Le dernier objectif que je me fixe pour la réalisation de ce projet est d'utiliser les résultats du module d'analyse de courriels et du traitement des données de références pour fournir des réponses aux courriels contenant une question ou une requête. Ce dernier objectif n'est atteignable que si les deux premiers donnent d'excellents résultats, mais si tel n'était pas le cas, des données traitées manuellement peuvent être utilisées pour effectuer le processus d'identification de la réponse.

4.2 Pistes de solution

Le problème de traitement automatique du courrier électronique se divise en trois parties : le traitement des courriels des utilisateurs, l'analyse du corpus de documents de références et l'extraction de la réponse. Ces trois parties font appel à des domaines différents du traitement des langues naturelles. Dans cette section, je vais décrire les pistes de solutions possibles pour la réalisation des objectifs décrits précédemment. Cette description de pistes de solutions servira à justifier les travaux que je compte exécuter pour arriver au but que j'ai fixé dans la section précédente.

Dans la première partie, l'objectif de traiter les courriels provenant des investisseurs présente plusieurs pistes de solution. Le but de cette étape est de déterminer la nature du message. Le processus de traitement peut se détailler de la façon suivante. En premier lieu, il faut déterminer si le courriel est une question ou une requête qui nécessite une réponse. Si c'est le cas alors le traitement de celui-ci peut être réalisé par notre système. L'étape suivante est la détermination d'un type sémantique de la question contenue dans le courriel, ce type est déterminé en fonction du type de réponse attendue. La détermination du type de la question se fera de façon similaire à ce qui est fait dans le système QUANTUM à la différence que je compte explorer l'avenue des méthodes empiriques. La réalisation de ce traitement nécessite l'exploration de plusieurs méthodes pour solutionner les sous-problèmes du traitement des courriels. La découverte de solutions à ces problèmes pourra s'appliquer aux autres étapes du traitement automatisé, entre autre pour le traitement du corpus de documents de références. Les problèmes les plus susceptibles d'être traités dans mes travaux et qui vont possiblement avoir le plus d'impact sur mes résultats sont : le traitement des collocations, le traitement des anaphores et l'extraction de relations sémantiques. Je décris ces problématiques plus loin dans cette section.

La deuxième partie du problème est de rendre les données utilisables pour réaliser le processus de question-réponse. Pour ce faire, la solution que je préconise est d'utiliser des méthodes de forage et d'extraction d'informations. Cette étape de traitement doit me permettre d'extraire les entités qui sont susceptibles d'être utilisées comme réponse aux courriels, c'est-à-dire les noms de personnes, les noms de compagnies, les références temporelles, les transactions et les numéros (téléphone, adresse, émission d'action). Le but de cette extraction est de pouvoir ensuite découvrir des relations entre certaines de ces entités, celles-ci seront utiles pour l'identification de la réponse. Les entités et les relations qui auront été découvertes devront ensuite être conservées dans une représentation sémantique. Deux types de représentation seront explorés pour conserver les connaissances, la première est la *discourse representation theory* (DRT) qui est très courante mais que je connais très peu, la seconde est l'utilisation de formalismes de représentation adaptés spécifiquement aux données que j'ai à représenter. L'avantage de ce dernier type de représentation est d'offrir une grande flexibilité d'utilisation pour l'encodage de l'information.

Le traitement des collocations est une problématique étudiée dans le traitement des langues naturelles et qui devra être explorée pour analyser et extraire l'information des courriels et des

données. Les collocations sont des expressions idiomatiques ou des groupes de mots agencés de façon à former un groupement sémantique [10]. Le traitement des collocations permettra de désambiguïser le sens des phrases et d'extraire certaines connaissances qui auraient pu être ignorées. J'ai fait référence à cette problématique dans la section 3.1.3, particulièrement dans l'exemple 4.2 de la page 32 avec l'abréviation AGM en lien avec la collocation *annual general meeting*. Le traitement des abréviations n'est pas le même problème que celui des collocations, je me permettrai quand même de traiter les deux simultanément. Les expressions *annual dividend rate* et *Series T Preferred Shares* sont des collocations qui apparaissent dans le corpus, le traitement des collocations sera réalisé par l'utilisation de ressources lexicales adaptées au domaine et de méthodes empiriques pour offrir une couverture plus étendue.

Le traitement des anaphores est une problématique qui est étudiée tant sur le plan théorique que pratique. Les théories sur le traitement des anaphores ne feront pas l'objet d'étude dans ce projet de recherche, mais j'explorerai les solutions aux problèmes pratiques. Dans les deux phrases qui suivent, j'ai mis en gras les expressions qui doivent être résolues avant de pouvoir comprendre le sens des questions.

(7.1) My folks inherited some BCE shares on April 30th 1983. Could you please tell what the share value was **that day** and also how does that relate to **today's** share price.

(7.2) Could you please check your records and determine if **any** are held at the moment.

Parmi les solutions que j'envisage pour faire l'analyse et la résolution des anaphores, l'utilisation des représentations sémantiques comme la DRT ou le lambda calcul sont des avenues qui sont intéressantes et qui s'harmonisent avec le processus d'analyse de la question.

Les relations sémantiques jouent un rôle important dans le problème de la question-réponse. Les questions posées peuvent porter sur des relations entre les entités ou sur une caractéristique liée à une entité. Dans le domaine des relations aux investisseurs, les relations susceptibles d'être traitées sont : les relations temporelles, les relations de possession ou des caractéristiques spécifiques à certaines entités. Le but d'extraire des relations sémantiques est de pouvoir facilement identifier les propriétés des objets lorsque vient le temps d'identifier les réponses. L'utilisation de relations peut aussi permettre d'effectuer un raisonnement, qui pourra être utilisé pour avoir un plus grand nombre de réponses et un taux de confiance possiblement plus élevé.

4.3 Travaux envisagés

L'objectif du travail est de traiter automatiquement les courriels dans le contexte du service aux investisseurs de BCE. Je décris les travaux qui seront réalisés pour élaborer de nouvelles solutions aux problèmes de traitement de la langue pour la réalisation du projet. À l'intérieur de cette section, j'explique ce que j'ai fait jusqu'à présent et ce qui devra être fait. La structure des travaux futurs est inspirée de la description de la problématique, chacun des modules mentionnés précédemment fera l'objet d'études à l'intérieur de ce projet.

Depuis le début de mes études doctorales, je me suis familiarisé avec le domaine de recherche du traitement des langues naturelles et j'ai approfondi mes connaissances du domaine de la recherche d'information. Ces activités d'apprentissage ont été réalisées dans le cadre de cours et par des lectures personnelles me permettant d'identifier les questions susceptibles de faire l'objet de recherches. Ma première année d'étude a été consacrée aux activités scolaires mentionnées précédemment et à la préparation des examens pré-doctoraux écrits. Je me suis aussi impliqué dans la vie départementale en étant auxiliaire d'enseignement pour le professeur J.-Y. Nie.

J'ai débuté les travaux par l'analyse du corpus BCE-4 que je présente à la section 3.1 de ce document. L'analyse a nécessité beaucoup de temps, six ou sept semaines de travail. J'ai dû travailler avec les particularités techniques des courriels : messages en plusieurs parties, messages au format HTML, pièces attachées au courriel, etc. J'ai donc développé des outils de traitement pour ces petites difficultés techniques, l'utilisation de ceux-ci permettant de traiter les courriels plus facilement. L'étape d'analyse m'a permis de bien établir les bases du problème à solutionner, de faire ressortir les caractéristiques importantes qui font que ce problème est différent des problèmes traités dans les systèmes de question-réponse, mais en même temps assez similaires pour pouvoir les traiter avec les méthodes de la question-réponse.

Après l'analyse du corpus, j'ai été sélectionné, par voie d'un concours, pour participer à un atelier de travail (*workshop*) portant sur la reconnaissance du temps et des événements pour les systèmes de question-réponse (TERQAS). L'atelier de travail s'est tenu de janvier à juillet 2002, il était organisé par James Pustejovsky de l'université Brandeis et il avait lieu au centre de recherche MITRE, en banlieue de Boston. Ma participation active à cet atelier a débuté au mois d'avril

et s'est terminée à la fin du mois de juillet. Durant l'atelier, j'ai exploré l'aspect sémantique des relations de temporalités des textes journalistiques et j'ai participé au développement du langage d'annotation TimeML. Ma contribution principale dans ces travaux est la création d'un outil de représentation graphique de documents annotés au format TimeML. Cet outil est utilisé pour aider les annotateurs à vérifier le résultat de leurs annotations et il pourrait être éventuellement utilisé comme base à un logiciel d'annotation plus complet. Les travaux réalisés au cours de cet atelier de travail serviront probablement pour le traitement des aspects temporels de mon problème de recherche, particulièrement lorsqu'il sera question de l'extraction de l'information.

Les travaux prévus pour la réalisation de mon projet de thèse vont être échelonnés sur une période d'un peu plus de deux ans, afin d'effectuer la soutenance de ma thèse au printemps 2005. Dans le reste de la section, je donne un échéancier des travaux prévus ainsi que la contribution de chacune des étapes, les grandes lignes de celui-ci se retrouvent dans le tableau 4.1. Les années 2001 et 2002 de l'échéancier concernent les activités que j'ai réalisé depuis mon inscription au doctorat. Les années 2003 et 2004 se rapportent aux deux années de travail que je crois nécessaire pour réaliser le projet de recherche.

Année	Début	Activité
2001	Janvier	Cours et examens de synthèse prédoc
2002	Janvier	Analyse du corpus BCE-4
	Mars	Atelier de travail TERQAS
	Août	Préparation du projet de recherche
2003	Janvier	Module d'analyse des courriels
	Juillet	Recherche de la solution
2004	Janvier	Analyse de résultats
	Février	Retour de l'évaluation et modifications
	Juin	Analyse finale
	Juillet	Rédaction de la thèse

TAB. 4.1 – Échéancier des travaux

La première étape de mon travail sera l'élaboration d'un module d'analyse des courriels. Le module d'analyse des courriels sera élaboré à partir du module d'analyse de questions de QUANTUM, du module de classification de Julien Dubois et des outils que j'ai développés lors de mon analyse de

corpus. Le développement de cette partie du système sera l'occasion d'explorer les différents modèles de représentations de connaissances et d'évaluer la pertinence de l'utilisation des représentations adaptées spécifiquement à mon problème. L'élaboration du module d'analyse des courriels sera probablement l'étape qui nécessitera le plus d'énergie car les problématiques sont nombreuses, la découverte de solutions pour ces problématiques pourra être utilisée aux étapes subséquentes de la réponse automatisée aux courriels. L'évaluation des résultats de ce module sera difficile à faire sans utiliser les modules de recherche de solutions car ce sont les modules qui auront à utiliser les résultats de l'analyse du courriel.

La deuxième étape du travail consiste à étudier les catégories de questions pour élaborer des schémas de raisonnement en fonction de l'extraction de la réponse. L'élaboration de ces schémas de raisonnement me permettra de déterminer l'information que je dois extraire des données du corpus et m'aidera à choisir une représentation des connaissances qui soit efficace pour effectuer des raisonnements afin d'extraire les réponses. Les solutions possibles pour l'extraction de la réponse ont déjà été esquissées précédemment lorsque j'ai décrit les fonctions d'extractions de QUANTUM. Je compte utiliser le même principe, mais en utilisant des fonctions plus spécifiques aux problèmes de relations aux investisseurs, en lien avec la classification des courriels du corpus BCE-4. Cette étape sera l'occasion de développer des méthodes d'extraction d'information pour le corpus de données de BCE. Je compte utiliser les méthodes de *text mining* pour extraire l'information pertinente à la recherche des réponses. Ces méthodes sont maintenant utilisées de façon courante lorsque les données sont numériques, la modification de celles-ci pour le traitement du texte pourrait donner des résultats intéressants, entre autre pour suivre l'évolution des entreprises, des titres financiers et des personnes faisant partie de l'environnement de BCE.

La dernière étape des travaux sera l'évaluation des résultats, puisque je ne traite pas la génération des courriels, l'évaluation que je propose sera réalisée en considérant le système de traitement automatisé comme un système d'aide à la réponse suggérant au préposé une réponse. Ce type d'évaluation pourra se réaliser avec des sujets humains qui détermineront si le système est utile ou non. Cette tâche d'évaluation sera ensuite utilisée pour modifier le système de sorte qu'il puisse déterminer automatiquement la qualité des réponses trouvées.

Bibliographie

[1] Beauregard, S. Génération de texte dans le cadre d'un système de réponse automatique à des courriels. Master's thesis, Université de Montréal, March 2001.

[2] Carbonell, J., Harman, D., Hovy, E., Maiorano, S., Prange, J., and Sparck-Jones, K. Vision statement to guide research in question & answering (q&a) and text summarization. Tech. rep., NIST, 2000.

[3] Dubois, J. Classification automatique de courrier électronique. Master's thesis, Université de Montréal, June 2002.

[4] Filatova, E., and Hovy, E. Assigning time-stamps to event-clauses. In *Proceedings of the Workshop on Temporal and Spatial Reasoning at the Conference of the ACL* (Toulouse, France, 2001).

[5] Hirschman, L., and Gaizauskas, R. Natural language question answering : the view from here. *Natural Language Engineering 7*, 4 (December 2001), 275–300.

[6] Kosseim, L. Systèmes de réponse automatique : état de l'art, March 2000.

[7] Laszlo, M., Kosseim, L., and Lapalme, G. Goal-driven answer extraction. In *Proceedings of the ninth text retrieval conference (TREC-9)* (2000).

[8] Lavenus, K., and Lapalme, G. Evaluation des systèmes de question réponse, aspects méthodologiques. *accepté à Traitement automatique des langues* (Août 2002), 30 pages.

[9] Lenhert, W. *The Process of question answering : a computer simulation of cognition.* Erlbaum Associates, New York, 1978.

[10] McKeown, K. R., and Radev, D. R. Collocations. In *A Handbook of Natural Language Processing*, R. Dale, H. Moisl, and H. Somers, Eds. Marcel Dekker, 2000.

[11] Plamondon, L. Le système de question-réponse quantum. Master's thesis, Université de Montréal, March 2002.

[12] Schilder, F., and Habel, C. From temporal expressions to temporal information : semantic tagging of news messages. In *Proceedings of the Workshop on Temporal and Spatial Reasoning at the Conference of the ACL* (Toulouse, France, 2001).

[13] Voorhees, E. M., and donald K. Harman, Eds. *Proceedings of the Eight Text Retrieval Conference (TREC-8)* (2000), NIST.