

Recherche collaborative de documents : comparaison assistance humaine/automatique

Jean-Baptiste Louvet¹, Guillaume Dubuisson Duplessis²,
Nathalie Chaignaud¹, Jean-Philippe Kotowicz^{1,3}, Laurent Vercoüter¹

¹ Normandie Univ, INSA Rouen, LITIS, 76000 Rouen, France
{jeanbaptiste.louvet, nathalie.chaignaud, laurent.vercoüter}@insa-rouen.fr

² LIMSI-CNRS, Paris
gdubuisson@limsi.fr

³ UEMF, INSA Euro-Méditerranée, Fès, Maroc
j.kotowicz@insa.ueuromed.org

Résumé : Partant du postulat que l'étude de l'interaction humain-humain peut servir à modéliser l'interaction humain-machine, nous étudions un corpus de dialogues de recherche collaborative de documents. De cette analyse est né un modèle de la tâche dans le cadre d'une interaction humain-humain. L'utilité de ce modèle est discuté pour le cas d'une interaction humain-machine.

Mots-clés : Recherche de documents, interaction humain-humain, interaction humain-machine

1 Introduction

Nous nous intéressons au système d'indexation de connaissances médicales CISMéF (Daroni *et al.*, 2000). Partant de l'étude d'un corpus de dialogues d'assistance à la recherche de documents (RD), nous définissons un modèle de la tâche de recherche collaborative de documents dans le cadre d'une interaction humain-humain. Nous montrons en quoi ce modèle de collaboration humain-humain (h-h) peut servir de base de discussion pour le cas d'une interaction humain-machine (h-m).

2 Recherche d'information

Cette section présente quelques modèles du processus de recherche d'information (RI) réalisé par un individu isolé pouvant s'appliquer à la RD. Un état de l'art plus complet est disponible dans les travaux de Dubuisson Duplessis (Dubuisson Duplessis, 2014).

On conçoit généralement la RI comme un processus de résolution de problème (Marchionini, 1989) impliquant un *chercheur* ayant un *besoin d'information identifié*. Le problème est alors de combler ce manque d'information. Le besoin d'information spécifié, le chercheur choisit un *plan de recherche* qu'il exécute au travers de la recherche en elle-même. Il évalue les résultats trouvés pour éventuellement réitérer le processus entier. La RI est un *processus itératif* (Sutcliffe & Ennis, 1998; Broder, 2002; Marchionini & White, 2007).

Le modèle standard est limité sur deux aspects. D'un côté, le *besoin d'information* de ce processus est vu comme *statique*. De l'autre, le chercheur est vu comme raffinant successivement sa requête jusqu'à tomber sur un ensemble de documents répondant à son besoin d'information *initial*. Certaines études ont démontré au contraire que le besoin d'information n'est pas statique

et que l'objectif n'est pas de déterminer une requête unique retournant un ensemble de documents répondant au besoin d'information (Bates, 1989; O'Day & Jeffries, 1993).

3 Analyse du processus de recherche collaborative h-h de documents à partir du corpus

De manière à comprendre l'aspect collaboratif d'un processus de RD d'un utilisateur assisté par un expert et à analyser les stratégies mises en place par ce dernier, nous avons mené une étude sur une interaction h-h de RD médicale. Cette étude est fondée sur l'analyse du corpus h-h recueilli lors du projet COGNI-CISMEF (Loisel, 2008; Chaignaud *et al.*, 2010).

3.1 Présentation du corpus COGNI-CISMEF

Le corpus est constitué de dialogues d'assistance sur une tâche de RD entre un expert et un utilisateur : l'expert se retrouve en situation de co-présence avec l'utilisateur qui apporte une formulation de son besoin d'information. L'expert dispose d'un accès au portail CISMEF et est chargé de mener la recherche en coopération avec l'utilisateur. Le portail CISMEF propose une interface graphique et un langage de requête permettant de décomposer une requête en éléments issus d'un lexique contrôlé. La terminologie CISMEF comporte des mots-clés, des qualificatifs, des méta-termes et des types de ressources. Le système s'est étoffé en offrant la possibilité d'effectuer des requêtes étendues mais l'utilisateur est souvent peu enclin à l'utilisation de requêtes complexes pourtant plus efficaces.

L'expérimentation a été menée auprès de membres volontaires (enseignants-chercheurs, étudiants, administratifs...) qui soumettaient une requête à un expert. Deux chercheurs du projet (un informaticien et une psycholinguiste), s'étant formés à la terminologie CISMEF, jouaient le rôle d'expert. Le corpus est constitué des retranscriptions des 21 dialogues issus de cette expérimentation et contient environ 37 000 mots.

3.2 Phases de la recherche de documents

L'analyse du corpus a permis d'identifier et de caractériser les différentes phases des dialogues du corpus COGNI-CISMEF (Dubuisson Duplessis, 2014), qui jouent chacune un rôle dans l'avancement de la tâche. Cinq phases ont été distinguées :

- la verbalisation : c'est l'établissement du sujet de la recherche entre l'utilisateur et l'expert. Elle commence par une formulation de la demande de l'utilisateur et peut être suivie de précisions spontanées de la part de celui-ci. L'expert peut alors commencer la construction de la requête s'il considère que la formulation est suffisante, demander des précisions si elle lui semble insuffisante ou tenter de reformuler la demande de l'utilisateur ;
- la construction de la requête : c'est la recherche de termes dans la terminologie CISMEF correspondant à la verbalisation de l'utilisateur. Il s'agit pour les interlocuteurs de trouver de façon collaborative un alignement entre la terminologie métier et le vocabulaire usuel de l'utilisateur pour ensuite remplir le formulaire de recherche ;
- le lancement de la requête : c'est simplement l'exécution de la requête courante par l'expert. Cette phase est souvent réalisée de manière implicite ;
- l'évaluation des résultats : l'expert évalue les résultats retournés par la requête. Si ceux-ci ne lui semblent pas satisfaisants, il décide de directement réparer la requête. Sinon, il les

- présente à l'utilisateur. Si celui-ci les juge satisfaisants, le but est atteint et la recherche se termine ; s'il les juge partiellement satisfaisants ou insatisfaisants, il faut réparer la requête ; suite à une évaluation négative, il est aussi possible de terminer sur un abandon ;
- la réparation de la requête : l'expert et l'utilisateur tentent de mettre en place des tactiques pour modifier la requête tout en respectant le besoin d'information. Trois tactiques ont été observées : la précision, la reformulation et la généralisation. Cependant, ces tactiques ne sont pas mutuellement exclusives : il est possible de combiner une précision ou une généralisation avec une reformulation.

En plus de ces phases, une phase d'ouverture et de clôture ont été observés. La phase d'ouverture est facultative et consiste en de simples salutations, tandis que la phase de clôture fait apparaître des propositions d'une nouvelle recherche.

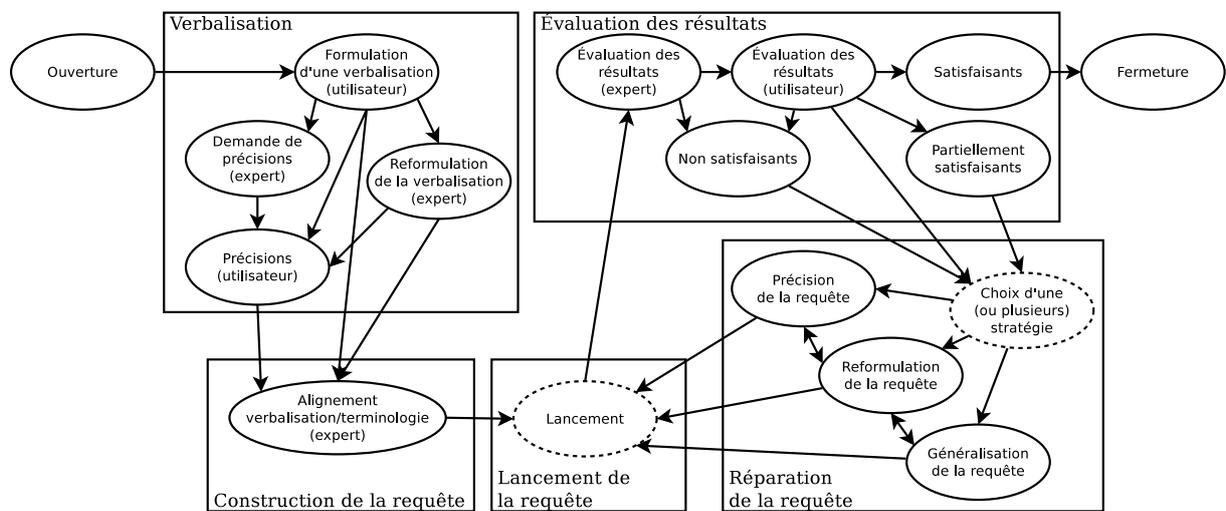


FIGURE 1 – Scénario issu de l'analyse du corpus représentant les enchaînements des phases de la tâche de RD.

3.3 Scénario issu de l'analyse du corpus

L'analyse de ce corpus a montré que cette tâche de RD est un processus itératif, opportuniste, stratégique et interactif (Dubuisson Duplessis, 2014; Bates, 1990). L'aspect itératif de ce processus est illustré par la répétition du motif lancement/évaluation/réparation de la requête.

Le scénario représenté par la Figure 1 synthétise les phases ainsi que les enchaînements possibles. Les états en pointillés sont les actions qui peuvent être réalisées de manière implicite par les participants à l'interaction. La boucle lancement/évaluation/réparation est bien présente.

4 Recherche collaborative de documents lors d'une interaction h-m

Nous nous intéressons maintenant aux différences entre une interaction h-h et une interaction h-m pour la RD collaborative. Nous comparons ici les capacités d'un expert humain et d'un système d'assistance dans une tâche d'assistance de formulation de requête.

4.1 Cadre de l'interaction h-m

Le cadre d'interaction h-m donne à l'utilisateur la possibilité de diriger la recherche et de modifier la requête sans l'accord du système. C'est un changement important car il donne l'ascendant à l'utilisateur sur l'interaction. Il implique une restriction des libertés du système par rapport à celles de l'expert humain. En effet, le système peut prendre des initiatives pour faire évoluer la requête en proposant des modifications qui ne seront effectives que si elles sont acceptées par l'utilisateur. Cette inversion des droits de modification de la requête permet cependant à chaque protagoniste de prendre part à l'interaction : l'initiative mixte est conservée.

Un autre changement est introduit : pour exposer son besoin d'information au système, l'utilisateur emploiera des mots libres ou éventuellement des phrases courtes avec utilisation d'outils TAL simples de la part du système (extraction de noms. . .).

4.2 Phase de verbalisation

Dans l'interaction h-h, c'est un moment d'interaction riche et complexe rendant explicite le besoin d'information de l'utilisateur. Cela nécessite de nombreuses compétences et connaissances de la part de l'expert : compréhension de l'expression verbale de l'utilisateur, connaissances générales du domaine d'application. . .

Dans le cadre h-m, cette phase sera forcément appauvrie. Un système d'assistance n'aura pas les capacités de TAL dont dispose l'expert humain et l'utilisateur devra donc entrer des mots libres (ou éventuellement des phrases très simples avec extraction de noms). Le système pourra toutefois demander des clarifications sur ces mots libres s'ils sont jugés trop peu nombreux.

L'expert humain est donc à son avantage dans la phase de verbalisation, celle-ci nécessitant des capacités cognitives difficiles à mettre en œuvre pour un système.

4.3 Phase de construction de la requête

Cette phase a pour objectif d'aligner la verbalisation de l'enquête avec la terminologie CISMEF. Dans le meilleur des cas, un ou plusieurs termes correspondent directement avec la terminologie. Pour le système, cet alignement sera fait instantanément et le système pourra proposer à l'utilisateur d'ajouter à la requête les termes reconnus. Cette tâche est plus difficile pour l'expert humain car il doit extraire de la verbalisation les « termes importants » puis les saisir dans l'interface de CISMEF pour vérifier qu'ils correspondent ou non à un terme de la terminologie. Cela peut s'avérer fastidieux s'il y a plusieurs « termes importants ».

Dans le cas où les termes de la verbalisation ne correspondent pas à la terminologie, l'expert humain et l'utilisateur collaborent en parcourant celle-ci afin de trouver une correspondance. De même, le système peut demander à l'utilisateur de parcourir la terminologie pour proposer des termes. Il s'agit donc, pour l'humain ou le système, de faire appel au jugement de l'utilisateur.

Cependant, le système dispose des dictionnaires de synonymes, hyponymes, hyperonymes et de liens « voir aussi » de CISMEF permettant de trouver des termes proches de la verbalisation de l'utilisateur. De plus, il est capable de conserver les sessions de RD précédentes et peut donc aussi tirer parti de cette connaissance (liens entre besoins d'information, requêtes et documents) pour trouver des termes connexes à la recherche courante (El Guedria & Vercouter, 2015).

4.4 Phase de lancement de la requête

Le lancement de la requête arrive après la phase de formulation ou de réparation de la requête. Quand il est à l'initiative de l'expert, celui-ci considère que la requête est suffisamment complète pour apporter des résultats pertinents au besoin d'information de l'utilisateur. Cependant, il arrive que les résultats de la requête soient rejetés après une évaluation sommaire, dans le cas où il y a trop ou trop peu de documents ou quand ceux-ci sont manifestement hors sujet.

De son côté le système d'assistance a l'avantage de pouvoir lancer « en privé » (c-à-d sans que l'utilisateur ne le sache) les requêtes. L'intérêt est que, avant de proposer une modification à l'utilisateur, il pourra tester celle-ci pour déterminer si elle apporte des résultats intéressants ou non.

4.5 Phase d'évaluation des résultats

Dans une interaction h-h, cette phase commence par une évaluation des résultats par l'expert avant de les présenter à l'utilisateur pour que celui-ci juge leur pertinence. Cette étape est supprimée dans le cadre d'une interaction h-m car cette évaluation aura été faite en privé par le système. C'est un avantage pour le système, lui permettant d'anticiper sur l'intérêt de la requête. Cependant, le système est limité par sa capacité à donner une évaluation correcte des résultats. Là où l'expert humain possède des compétences cognitives pour rejeter ou non des résultats, le système se limitera à quelques paramètres descriptifs : nombre, score, métadonnées. . .

Dans le cas h-h et dans le cas h-m, le jugement est réalisé en dernier ressort par l'utilisateur.

4.6 Phase de réparation de la requête

La première étape dans une réparation de requête est l'identification de la tactique à mettre en place pour améliorer la requête. Le choix d'une tactique paraît assez simple : si la requête retourne trop de résultats, il faut la préciser ; si elle n'en retourne pas assez, il faut la généraliser et si elle en retourne un nombre suffisant mais que les documents ne sont pas satisfaisants pour l'utilisateur, il faut la reformuler. Cependant, ces tactiques peuvent être combinées.

Une précision nécessite un ajout de termes à la requête et les moyens employés sont proches de ceux utilisés lors de la phase de construction de la première requête. Une généralisation nécessite la suppression de termes de la requête. Grâce à ses capacités combinatoires, le système peut tester différentes possibilités de suppression et ainsi déterminer laquelle sera la plus avantageuse.

Pour la reformulation, le système peut faire appel aux ressources lexicales de l'application sur les termes de la requête (Audeh *et al.*, 2013; Soualmia, 2004) et il pourra faire appel à l'utilisateur pour proposer des reformulation de certains termes de la requête .

5 Conclusion et perspectives

Dans cet article, nous avons étudié une situation collaborative h-h de RD. Un modèle de la tâche entre un expert humain et un utilisateur est décrit sous la forme d'un scénario. À partir de ce modèle, nous avons comparé les capacités mises en jeu par l'expert humain et celles d'un système d'assistance à la formulation de requête.

Notre objectif est de concevoir un système collaboratif pour la RD capable d'interagir avec l'utilisateur et de lui proposer des stratégies coopératives. Le système doit pouvoir analyser

l'objectif de l'utilisateur et lui proposer des solutions pour faire évoluer l'état de sa recherche. Il doit présenter des exemples, des aides, des corrections ou des clarifications. Il doit accompagner l'interlocuteur jusqu'à trouver une solution en élargissant son but initial si nécessaire.

Une étude des aspects dialogiques de ce corpus a été proposée (Dubuisson Duplessis *et al.*, 2013, 2015). Elle montre que le modèle conventionnel des jeux de dialogue est bien adapté pour modéliser les niveaux interprétatif et génératif du dialogue. Nous travaillons actuellement sur les liens entre le modèle du dialogue et le modèle de la tâche.

Références

- AUDEH B., BEAUNE P. & BEIGBEDER M. (2013). Expansion sémantique des requêtes pour un modèle de recherche d'information par proximité. In *INFORSID*, p. 83–90.
- BATES M. J. (1989). The Design of Browsing and Berrypicking Techniques for the Online Search Interface. *Online Review*, **13**(5), 407–424.
- BATES M. J. (1990). Where should the person stop and the information search interface start? *Inf. Process. Manage.*, **26**(5), 575–591.
- BRODER A. (2002). A taxonomy of Web search. *ACM SIGIR Forum*, **36**(2), 3–10.
- CHAIGNAUD N., DELAVIGNE V., HOLZEM M., KOTOWICZ J.-P. & LOISEL A. (2010). Étude cognitive des processus de construction d'une requête dans un système de gestion de connaissances médicales. *Revue Technique et Science Informatiques (TSI)*, **29**, 991–1021.
- DARMONI S., LEROY J., BAUDIC F., DOUYERE M., PIOT J. & THIRION B. (2000). CISMef : a structured health resource guide. *Methods of Information in Medicine*, **39**, 30–35.
- DUBUISSON DUPLESSIS G. (2014). *Modele de comportement communicatif conventionnel pour un agent en interaction avec des humains : Approche par jeux de dialogue*. PhD thesis, INSA de Rouen.
- DUBUISSON DUPLESSIS G., CHAIGNAUD N., KOTOWICZ J., PAUCHET A. & PÉCUCHE J. (2013). Empirical specification of dialogue games for an interactive agent. In *Advances on Practical Applications of Agents and Multi-Agent Systems (PAAMS) Salamanca, Spain*, p. 49–60.
- DUBUISSON DUPLESSIS G., PAUCHET A., CHAIGNAUD N. & KOTOWICZ J. (2015). A conventional dialogue model based on empirically specified dialogue games. In *International Conference on Tools with Artificial Intelligence (ICTAI), Vietri sul Mare, Italy*, p. 997–1004.
- EL GUEDRIA Z. & VERCOUTER L. (2015). Personnalisation par un système multi-agent de la navigation au sein d'un corpus documentaire. In *JFSMA*, p. 61–70 : Cépaduès Éditions.
- LOISEL A. (2008). *Modélisation du dialogue Homme-Machine pour la recherche d'informations : approche questions-réponses*. PhD thesis, INSA de Rouen.
- MARCHIONINI G. (1989). Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science*, **40**(1), 54.
- MARCHIONINI G. & WHITE R. (2007). Find What You Need, Understand What You Find. *Int. J. Hum. Comput. Interaction*, **23**(3), 205–237.
- O'DAY V. L. & JEFFRIES R. (1993). Orienteering in an information landscape : how information seekers get from here to there. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, p. 438–445 : ACM.
- SOUALMIA L. (2004). *Etude et évaluation d'approches multiples d'expansion de requêtes pour une recherche d'information intelligente : application au domaine de la santé sur l'Internet*. PhD thesis, INSA de Rouen.
- SUTCLIFFE A. & ENNIS M. (1998). Towards a cognitive theory of information retrieval. *Interacting with computers*, **10**, 321–351.