Développement d'un système de détection des infections associées aux soins à partir de l'analyse de comptes-rendus d'hospitalisation

Caroline Hagège1 Denys Proux1 Quentin Gicquel2 Stefan Darmoni3
Suzanne Pereira4 Frédérique Segond1 Marie-Helène Metzger2

(1) XRCE, 6 Chemin de Maupertuis, 38240 Meylan, France
(2) UCBL-CNRS, UMR 5558 Lyon, France
(3) CISMEF, Rouen, France
(4) VIDAL, Issy les Moulineaux, France
Caroline.Hagege@xrce.xerox.com, Denys.Proux@xrce.xerox.com, Quentin.Gicquel@chu-lyon.fr,
Stefan.Darmoni@cismef.fr, Suzanne.Pereira@vidal.fr, Frederique.Segond@xrce.xerox.com,
Marie-Helene.Metzger@chu-lyon.fr

Résumé

Cet article décrit la première version et les résultats de l'évaluation d'un système de détection des épisodes d'infections associées aux soins. Cette détection est basée sur l'analyse automatique de comptes-rendus d'hospitalisation provenant de différents hôpitaux et différents services. Ces comptes-rendus sont sous forme de texte libre. Le système de détection a été développé à partir d'un analyseur linguistique que nous avons adapté au domaine médical et extrait à partir des documents des indices pouvant conduire à une suspicion d'infection. Un traitement de la négation et un traitement temporel des textes sont effectués permettant de restreindre et de raffiner l'extraction d'indices. Nous décrivons dans cet article le système que nous avons développé et donnons les résultats d'une évaluation préliminaire.

Abstract

This paper describes the first version and the results obtained by a system which detects occurrences of healthcare-associated infections. The system automatically analyzes hospital discharge summaries coming from different hospitals and from different care units. The output of the system consists in stating for each document, if there is a case of healthcare-associated infection. The linguistic processor which analyzes hospital discharge summaries is a general purpose tool which has been adapted for the medical domain. It extracts textual elements that may lead to an infection suspicion. Jointly with the extraction of suspicious terms, the system performs a negation and temporal processing of texts in order to refine the extraction. We first describe the system that has been developed and give then the results of a preliminary evaluation.

Mots-clés : Extraction d'information médicale, compte-rendus d'hospitalisation, infection nosocomiale, analyse syntaxique
Keywords: Information extraction in medical domain, hospital discharge summaries, hospital acquired infections, parsing

1 Introduction

Les infections acquises en milieu hospitalier (infections nosocomiales) sont un enjeu important pour les établissements de soins. On estime en effet qu'en France ce type d'infections touche entre 5% et 10% des patients hospitalisés. Il a aussi été estimé qu'environ 30% de ces infections pourraient être évitées. On comprend donc sans peine que ce problème soit au cœur des préoccupations du milieu médical comme en témoignent les nombreuses études épidémiologiques et les actions de surveillances réalisées dans ce domaine (voir l'activité des CCLIN[1] par exemple). Le travail décrit ici est développé dans le cadre du projet ANR ALADIN DTH[2] (se référer à (Proux et. al., 2009) pour une description générale du projet) et a pour but de développer un outil qui, à partir de comptes-rendus d'hospitalisation rédigés par des médecins, est capable de déterminer si ce compte-rendu relate un épisode relevant d'une infection nosocomiale (IN). Cet outil spécifique a été développé à partir d'un outil plus général permettant d'extraire de l'information syntaxique et sémantique à partir de textes tout-venant (Aït-Mokhtar et al., 2002). Dans la suite de cet article, nous décrivons d'abord les adaptations de l'outil linguistique général pour notre domaine applicatif. Puis, dans un deuxième temps, nous détaillons les développements spécifiques réalisés. Enfin, nous concluons en indiquant quels sont les développements en cours et ceux envisagés dans un futur proche.

2 Description du système

Nous considérons deux étapes principales dans le développement du système. La première consiste en l'adaptation d'un outil de traitement linguistique relevant du domaine général au nouveau domaine des comptes-rendus d'hospitalisation. La seconde étape consiste en un ensemble de développements spécifiques pour la finalité applicative. Nous décrivons ci-dessous ces deux étapes.

2.1 Adaptation au domaine médical de l'outil de traitement linguistique général

XIP (Aït-Mokhtar et al., 2002) est un outil de traitement linguistique général qui procède à l'analyse syntaxique fine en dépendances de textes tout venant (aux formats txt ou xml). Un module de reconnaissance des entités nommées (EN) relevant du domaine général (presse d'information générale, encyclopédie générale etc.) est intégré dans cette analyse. Les textes que nous traitons dans le cadre du projet ALADIN relèvent cependant du domaine particulier des comptes-rendus d'hospitalisation (que nous désignerons désormais par CR) provenant de services de réanimation, de chirurgie digestive, de chirurgie orthopédique et de neurochirurgie de divers hôpitaux. Un enrichissement lexical a donc été nécessaire pour que notre outil puisse traiter correctement ces textes. Par ailleurs, les CR sont des textes libres, qui, selon les services ou les rédacteurs peuvent présenter des particularités comme des emplois de style télégraphique, des abréviations, certaines fautes d'orthographe récurrentes etc. Enfin, ces textes ont un type de format et d'organisation particuliers. L'adaptation au domaine a été organisée selon deux axes qui sont d'une part l'enrichissement lexical et terminologique et d'autre part la prise en compte des particularités de la collection de documents que nous traitons.

Le traitement des textes spécialisés a nécessité un apport important de ressources lexicales et terminologiques. Ces ressources ont été intégrées dans les lexiques à états-finis utilisés par l'analyseur. Nous avons choisi dans un premier temps cette approche par rapport à une approche qui aurait consisté à intégrer dans le flux des traitements l'utilisation d'outils d'indexation terminologique déjà existants pour deux raisons principales : La première est qu'il n'existe pas de terminologie spécifique aux problèmes d'infections acquises à l'hôpital[3] et que nous aurions du utiliser des terminologies vastes et plus générales. Les essais préliminaires que nous avons effectués dans ce sens ont montré que nous obtenions une couverture bien trop large pour nos besoins d'extraction d'information. Ces outils d'indexation sont en effet destinés à favoriser le rappel (l'idée sous-jacente est de permettre d'extraire TOUS les textes pertinents à partir d'une recherche par mots-clés) et nous souhaitons insister dans ce contexte sur la précision. Par

[1] http ://www.cclin-france.fr

[2] http://www.aladin-project.eu

[3] Une telle terminologie est actuellement en cours de développement au CISMEF dans le cadre du projet.

ailleurs, l'intégration directe des ressources lexicales et terminologiques dans nos automates lexicaux permet d'obtenir un temps de traitement plus court. Ces deux raisons ont fait que nous avons opté pour le codage des termes dans l'analyseur lexical[4]. Nous avons intégré environ 4000 termes et éléments du vocabulaire médical à notre système (correspondant approximativement à 11000 formes fléchies). Cet apport lexical et le fait de travailler dans un domaine particulier a une influence sur le traitement de l'ambigüité du système général. Par exemple, l'ajout du terme *Candida* dans le lexique a pour effet de créer une nouvelle ambigüité entre nom de personne (Candida est un prénom féminin) et un nom de germe (levure). Les nouveaux termes ont été analysés par le système général avant leur introduction dans le lexique et les nouvelles ambigüités dues à cet apport lexical ont été examinées et parfois résolues. Dans le cas précis de *Candida*, dans la mesure où nous travaillons sur les CR anonymisés (donc sans nom propre), nous avons systématiquement privilégié un étiquetage par un nouveau terme par rapport à un étiquetage en nom de personne.

Le deuxième élément important est l'adaptation linguistique de l'analyseur au traitement des documents particuliers que sont les CR. Ces documents sont rédigés de manière synthétique par des médecins et contiennent des coquilles, des manques de ponctuation et un emploi du style télégraphique qui viennent troubler le comportement habituel de l'analyseur. Ainsi, il est courant de trouver des phrases sans verbe comme *fracture jambe gauche.* Pour un système d'analyse linguistique général, le mot *fracture* sera désambigüisé comme Verbe à l'impératif, alors que dans le contexte des CR, il s'agit d'un nom. De plus, des erreurs orthographiques fréquentes sur des termes ont été relevées dans les corpus et codées dans les automates lexicaux (par exemple, graphie incorrecte *cytrobacter* au lieu de *citrobacter*).

2.2 Développements spécifiques à l'application

L'outil que nous avons développé prend en entrée un compte-rendu d'hospitalisation et produit en sortie ce même compte-rendu annoté. L'annotation consiste en l'extraction des indices potentiels pouvant conduire à une suspicion d'IN accompagnée de l'information finale concernant le statut de ce document par rapport à l'infection (suspicion ou non suspicion). L'outil est développé en JAVA et utilise l'API Java de notre analyseur ce qui nous permet d'avoir accès et de stocker l'information linguistique calculée par l'analyseur lors de l'exécution du programme. Nous avons développé le système selon les axes suivants :

- Extraction d'indices avec traitement de la négation
- Traitement temporel
- Heuristiques pour la détection des infections

2.2.1 Extraction d'indices

Le but de cette extraction est de détecter et de typer toute l'information qui relève du problème des IN dans les textes. Ce travail d'extraction se situe dans la lignée de travaux d'extraction d'information dans le domaine du médical (voir par exemple (Deléger et al., 2010)). Cette extraction est comparable à une reconnaissance d'entités nommées classique pour laquelle de nouvelles EN et de nouveaux types sémantiques sont considérés. Pour notre application, nous considérons les types suivants : BACTERIES, VIRUS et LEVURES qui constituent trois sous-types d'un type général GERME_INFECTIEUX. ANTISEPTIQUE correspond à des mentions de produits antiseptiques utilisés en chirurgie. TEMPERATURE marque des expressions simples ou phrastiques indiquant qu'un patient a de la fièvre. DISPOSITIF type les occurrences de dispositifs invasifs (ex. cathéters) qui sont souvent des points d'entrée pour les IN. INFECTION correspond à un diagnostic établi d'infection. Il peut s'agir d'une maladie infectieuse (comme une pneumopathie par exemple) ou de la présence de pus. ANTIBIOTIQUE désigne la classe des antibiotiques. Enfin, EXAMENS subsume les sous-types EXAMEN_BIOLOGIQUE, EXAMEN_BACTERIOLOGIQUE et EXAMEN_RADIOLOGIQUE. La reconnaissance de ces nouveaux types d'entités se base sur l'information lexicale et terminologique du domaine qui a été intégrée dans le lexique, mais aussi sur une série de règles (contextuelles ou plus larges utilisant des dépendances syntaxiques). Il est important d'avoir une très bonne précision pour la reconnaissance de ces types car ils

[4] Nous avons utilisé les termes des CR annotés manuellement par les médecins partenaires que nous avons enrichis par des ressources mis à notre disposition par les partenaires CISMEF et Vidal.

vont être utilisés par les heuristiques de détection des infections. Nous avons par ailleurs rajouté la reconnaissance des types DIAGNOSTIC (autres que maladie infectieuse) et INTERVENTION_CHIRURGICALE, qui interviennent de manière moins directe dans la détection des suspicions d'infection mais que les médecins souhaitent voir annotés dans les documents car ils vont aider à donner le contexte de l'histoire du patient et faciliter la relecture des CR. L'extraction de ces indices est accompagnée d'un traitement fin de la négation intégré à l'analyseur linguistique. Nous avons présenté et évalué (Hagège, 2011) une méthode pour le traitement de la négation dans le sous-domaine des CR d'hospitalisation qui nous montre qu'environ 8% des indices potentiels apparaissant dans les textes comportent une négation. Cette méthode permet de détecter ces négations avec une précision de 95,6% et un rappel de 96,6%.

2.2.2 Traitement temporel

Un module de traitement de la temporalité est intégré dans l'analyseur. On doit pouvoir associer à chaque indice extrait une coordonnée temporelle calculée en terme de distance par rapport à la date d'entrée dans le service (cas de la réanimation) et à la date de l'intervention (cas de la chirurgie). En effet, la prise en compte du facteur temps, date d'apparition de l'infection par rapport à la date de l'hospitalisation ou de l'intervention chirurgicale, est un paramètre capital pour la détection correcte des infections acquises à l'hôpital. Le traitement de la temporalité dans les comptes-rendus d'hospitalisation est décrit dans (Hagège et. al., 2010). Il permet de découper le compte-rendu initial en blocs temporels, chacun de ces blocs ayant un attribut temporel exprimé en termes de distance temporelle avec une date T0 (date d'hospitalisation ou d'intervention chirurgicale). Tout indice extrait à l'intérieur d'un bloc temporel aura une coordonnée temporelle correspondant à la valeur de l'attribut temporel du bloc d'où il est extrait.

2.2.3 Ensemble d'heuristiques pour la détection des infections

Une des principales difficultés vient du fait que la survenue d'une IN est rarement mentionnée de manière totalement explicite dans les CR[5]. De plus, nous constatons que nous trouvons rarement dans les CR la mention de toutes les directives médicales permettant de définir les IN et ISO (voir par exemple (Horan et al. 2008)). En effet, si nous étudions les CR dans lesquels les médecins ont considéré un risque d'infection, nous vérifions que seul un sous-ensemble de ces conditions définissant l'infection est annoté dans ces textes. Enfin, alors que nous avons souhaité favoriser la précision dans l'extraction des indices suspects pour l'infection, nous souhaitons pour l'extraction des CR, et à la demande des médecins, favoriser un bon rappel. Grâce à un ensemble de CR annotés par les médecins investigateurs du projet et mis à notre disposition nous avons développé et testé différentes heuristiques. Ces heuristiques utilisent l'extraction des indices et de l'information temporelle décrites aux points précédents. Elles ont été déterminées grâce à l'étude de CR annotés par les médecins investigateurs, à l'utilisation des algorithmes de détection formalisés par les partenaires médicaux, ainsi que par transmission des connaissances médicales nécessaires à la compréhension de la thématique par ces même partenaires. La détection des infections nosocomiales a été étudiée en réanimation et en chirurgie (neurochirurgie, chirurgie digestive, orthopédie) où l'objectif était la détection de l'infection du site opératoire (ISO). Voici différents exemples d'heuristiques :

Nous considérons qu'un CR provenant d'un service de réanimation comporte une suspicion d'IN si l'une des conditions ci-dessous est remplie :
- On trouve dans le texte une mention explicite à une infection nosocomiale (suite de caractère « infection nosocomiale » ou entité de type INFECTION modifiée par l'adjectif « nosocomial »
- On trouve dans le texte conjointement au moins deux entités INFECTION et ANTIBIOTIQUE et les ancrages temporels de ces entités sont supérieurs ou égal à deux jours après T0 (date d'hospitalisation) ET le patient ne présente pas d'infection a l'entrée dans le service (entité INFECTION, GERME_INFECTIEUX avec coordonnée temporelle égale à T0) ET le patient n'est pas décédé lors de l'hospitalisation[6]

[5] Nous n'avons trouvé sur 64 CR de réanimation contenant des épisodes d'IN qui ont été utilisés pour la mise au point de l'outil que 11 CR où une forme du mot « nosocomial » apparaît.

[6] En effet, en cas de décès du patient lors de l'hospitalisation, des éventuelles limitations thérapeutiques font que souvent moins d'indices pour la detection des IN apparaissent dans les textes (par exemple, une infection peut-etre détectée mais non soignée par des antibiotiques). Cette condition a pour résultat d'ajouter des contraintes supplémentaires aux heuristiques de detection des l'infection.

- On trouve dans le texte conjointement au moins deux entités GERME_INFECTIEUX et ANTIBIOTIQUE dont les ancrages temporels sont supérieurs ou égal à deux jours après T0 ET que le patient ne présente pas d'infection a l'entrée dans le service (entité INFECTION, GERME_INFECTIEUX avec coordonnée temporelle égale à T0) ET le patient n'est pas décédé lors de l'hospitalisation
- Le patient a une infection à l'entrée dans le service (INFECTION ou GERME_INFECTIEUX avec coordonnée temporelle égale à T0) ou il décède lors de l'hospitalisation ET l'on trouve au moins deux entités INFECTION, GERME_INFECTIEUX, ANTIBIOTIQUE, TEMPERATURE, DISPOSITIF avec une coordonnée temporelle supérieure ou égale à T0+2jours

Nous considérons qu'un CR provenant d'un service de chirurgie comporte une suspicion d'ISO si l'une des conditions ci-dessous est remplie :
- on trouve une mention explicite à une infection du site opératoire dans le texte
- on trouve au moins une des entités INFECTION, ANTIBIOTIQUE, ANTISEPTIQUE, GERME_INFECTIEUX, EXAMEN_BACTERIOLOGIQUE postérieure à la date d'intervention dans le document.

3 Evaluation et futurs développements

Une évaluation préliminaire des performances de détection a été effectuée par nos partenaires médicaux UCBL sur 205 CR du jeu d'apprentissage (avant leur utilisation pour développement) (Berrouane et al., 2011). Le but de cette évaluation est de déterminer dans quelle mesure le système est capable de classer les documents dans les deux catégories, infectés ou non infectés. Ces CR proviennent des différentes spécialités (Réanimation, chirurgie digestive, orthopédie et neurochirurgie) des différents établissements de santé (Rouen, Lyon et Nice). Sur ces 205 CR, 128 contiennent une occurrence d'infection. Cette répartition entre CR infectés et non infectés ne correspond pas à la réalité de la proportion des cas d'infection dans les services hospitaliers et il est important de le signaler car il explique pourquoi nous n'avons pas calculé la précision. Un des éléments importants dans le développement était d'estimer le rappel ainsi que nous l'avons indiqué au point 2.2.3. Cet échantillonnage, bien que non représentatif de la réalité dans les établissement de santé, nous permet de l'évaluer. Nous donnons également les résultats du calcul de la spécificité. Les formules utilisées pour les calculs sont rappelées dans le tableau où VP, VN, FP et FN représentent respectivement les vrais positifs, vrais négatifs, faux positifs et faux négatifs obtenus lors de la comparaison entre l'annotation manuelle et l'annotation automatique.

	VP	VN	FP	FN	total	Rappel VP/(VP+FN)	Spécificité VN/(VN+FP)
Tous	**113**	**74**	**2**	**16**	**205**	**87,6%**	**97.4%**
Réa.	10	12	1	6	29	62,5%	92.3%
Ch. Dig.	35	28	0	4	67	89,7%	100%
Ch. Orth.	14	4	1	2	21	87,5%	80%
Neurochir.	54	30	0	4	88	93,1%	100%

Tableau 1 : Résultats de l'évaluation par spécialités

Ces résultats ont montré que l'objectif d'un rappel élevé était atteint (87,6%]) avec toutefois des variations selon la spécialité médicale, le rappel étant plus faible en réanimation (62,5), et le plus élevé en neurochirurgie (93.1%). L'analyse des résultats montre que les erreurs du système proviennent de différentes sources. La première source d'erreur, la moins fréquente cependant, est due à du codage insuffisant de termes. La seconde source d'erreur (concernant la détection d'ISO) est la présence possible de cas d'IN dans des CR de chirurgie. Or, pour la chirurgie, à la demande des médecins, seules les occurrences d'ISO doivent être prises en considération. Enfin, la source d'erreur la plus importante est celle qui est liée au traitement temporel (pour les CR de réanimation). En effet, ainsi qu'on peut le constater dans les heuristiques présentées plus haut, dans le contexte de la réanimation, la détection des IN est très dépendante

du calcul correct des coordonnées temporelles des entités extraites du texte. La poursuite de l'analyse des CR d'apprentissage (800 CR mis à disposition au total) devrait permettre l'amélioration de ces performances, notamment en réanimation.

Par ailleurs, nous sommes actuellement en train de raffiner les règles de détection des ISO afin de déterminer si l'infection est superficielle ou profonde. Nous souhaitons également développer une interface qui présentera aux médecins les comptes-rendus annotés par le système ainsi que l'heuristique qui a été utilisée pour déterminer s'il s'agit d'un cas d'IN. La mise en œuvre de l'outil est actuellement prévue au centre hospitalier universitaire de Lyon. Nous souhaitons également utiliser les CR ayant servi à l'apprentissage de notre outil pour entrainer des arbres de décisions (en cours d'élaboration par nos partenaires UCBL), afin de les comparer à nos heuristiques et éventuellement les améliorer. Une fois tous ces développements réalisés, une évaluation de l'outil sera réalisée par nos partenaires médicaux UCBL avec un nouveau jeu de 800 CR.

Remerciements

Nous remercions l'ANR pour le financement du projet ANR-08-TecSan-001-01 ALADIN- DTH qui s'inscrit dans le cadre de la recherche partenariale en technologies pour la santé et l'autonomie (TecSan).

Références

AIT-MOKHTAR S., CHANOD J-P., ROUX C. (2002). Robustness beyond Shallowness : Incremental Deep Parsing. *Natural Language Engineering* 8, 121-144.

BERROUANE Y., HAGÈGE C., GICQUEL Q.,, KERGOULAY I., PEREIRA S., PROUX D., DARMONI S., SEGOND F., METZGER M-H. (2011) Preliminary evaluation of an automated detection tool for healthcare-associated infections, based on screening natural language medical reports. *21st European Congress of Clinical Microbiology and Infectious Diseases*, Milan, May 2011 - (Affiche topic 52).

DELEGER L., GROUIN C., ZWEIGENBAUM P. (2010). Extracting medication information from French clinical texts. Proceedings of *MEDINFO 2010,* Vol 160. 949-953.

HAGÈGE C., MARCHAL P., GICQUEL Q., DARMONI S., PEREIRA S., METZGER M-H. (2010). Linguistic and Temporal Processing for Discovering Hospital Acquired Infections from Patient Records. Proceedings of the ECAI 2010 Conference Workshop on Knowledge Representation for Health-Care (KRHC'10).

HAGEGE C. (2011). Linguistically Motivated Negation Processing. Proceedings CiCLing 2011, to appear.

HORAN T., ANDRUS M., DUDECK A. (2008). CDC/NHSN Surveillance Definition of Health Care-Associated Infection and Criteria for Specific Types of Infections in the Acute Care Setting. *AJIC: American Journal of Infection Control,* Volume 36, Issue 5, 309-332, AJIC: AMERICAN JOURNAL OF INFECTION CONTROL

PROUX D., MARCHAL P., SEGOND F., KERGOULAY I., DARMONI S., PEREIRA S., GICQUEL Q., METZGER M-H. (2009). Natural Language Processing to detect Risk Patterns related to Hospital Acquired Infections. Proceedings of *RANLP 2009,* Borovetz, Bulgaria, 865-881.